Advances in Marine Genomics

J. Mark Cock
Kristin Tessmar-Raible
Catherine Boyen
Frédérique Viard
Editors

Introduction to Marine Genomics

Springer

Advances in Marine Genomics

Volume 1

Series Editor

J. Mark Cock

CNRS, UMR 7139
Laboratoire International Associé Dispersal
and Adaptation in Marine Species
Station Biologique de Roscoff
Place Georges Teissier, BP74
29682 Roscoff Cedex
France

For further volumes:
http://www.springer.com/series/8684

J. Mark Cock · Kristin Tessmar-Raible · Catherine Boyen · Frédérique Viard
Editors

Introduction to Marine Genomics

Springer

Editors
J. Mark Cock
CNRS, UMR 7139
Laboratoire International Associé Dispersal and Adaptation in Marine Species
Station Biologique de Roscoff
Place Georges Teissier, BP74
29682 Roscoff Cedex
France
cock@sb-roscoff.fr

Catherine Boyen
Université Paris VI
CNRS UMR 7139
Labo. Végétaux Marins et Biomolécules
place Georges Teissier
29682 Roscoff CX
France
boyen@sb-roscoff.fr

Kristin Tessmar-Raible
Max F. Perutz Laboratories
University of Vienna Campus Vienna Biocenter
Dr. Bohr-Gasse 9/4
1030 Vienna
Austria
kristin.tessmar@mfpl.ac.at

Frédérique Viard
Université Paris VI
Equipe Diversité et Connectivité dans le Paysage Marin Côtier (DIV&CO)
place Georges Teissier
29682 Roscoff CX
France
viard@sb-roscoff.fr

ISBN 978-90-481-8616-7 e-ISBN 978-90-481-8639-6
DOI 10.1007/978-90-481-8639-6
Springer Dordrecht Heidelberg London New York

Library of Congress Control Number: 2009944167

Preface

Genomics can be defined as the study of the structure, function and diversity of genomes, genome being the collective term for all the genetic information contained in a particular organism. The difference between genomic approaches and classical biological approaches is essentially one of scale, the aim of genomics being to extend analyses that were originally limited to one or a small number of genes to large numbers of genes or even to the complete set that make up a genome. One consequence of this is that it is not always easy to define where classical approaches end and genomics begins, and this has even led to suggestions that there is no essential difference between the two types of approach. This is an extreme view however and, as the examples given in this book will illustrate, the need to develop high-throughput methodologies adapted to the analysis of multiple gene sets has led to considerable technical advances and to the development of new ways of thinking about biology.

Genomics as a discipline began with the first attempts to obtain large-scale sequence data for individual organisms, either by sequencing the genomic DNA or, where this was not practical, by sequencing large numbers of cDNAs.[1] This was initially a very costly process and consequently early efforts were focused on laboratory model organisms such as the bacteria *Escherichia coli* and the yeast *Saccharomyces cerevisiae* followed by the animal species *Caenorhabditis elegans* and *Drosophila melanogaster* and the plant *Arabidopsis thaliana*. The availability of extensive sequence data for these organisms then facilitated the development of additional genomic tools such as microarray systems for the analysis of gene expression and collections of sequence-tagged mutants.

Marine organisms were very poorly represented amongst these early genomic models and were to some extent left behind as the application of genomic approaches to several terrestrial models allowed an acceleration of our understanding of the biology, ecology and evolutionary history of these species. The situation has been rectified to a certain extent in recent years, essentially due to the reduced cost of DNA sequencing and associated enhanced capacity for analysis of very large datasets. The reduction in costs has not only allowed the application of genomic

[1] See the glossary for definitions of technical terms.

approaches to a much broader range of species but has also opened up new fields of genomics such as metagenomics and metatranscriptomics in which sequencing methodologies are used in new and innovative ways. Marine biology has been at the forefront of many of these new applications.

Genomic approaches are now being applied to a diverse catalogue of questions in marine biology. These include exploiting the enormous phylogenetic diversity of marine organisms to explore the evolution of developmental processes, characterising the marine ecosystems that play key roles in global geochemical cycles, searching for novel biomolecules and understanding ecological interactions within important marine ecosystems. Several of these domains have become increasingly important in the context of global climate change.

The objective of this book is to provide an overview of recent advances in the application of genomic approaches in the domain of marine biology. Each chapter covers a specific field, providing an introduction to that area and detailing how genomic approaches have been adapted to the specific requirements of that field of research. The final chapter discusses some practical questions of applying genomic techniques, particularly in terms of the bioinformatic challenges. Many of these fields of research addressed by individual chapters in this book will be covered in further detail by complete volumes within the Advances in Marine Genomics book series.

The first chapter looks at how genomic techniques are being used to explore and monitor biodiversity in the marine environment. Consideration of biodiversity is of particular interest in the light of accelerating climate change, which is increasingly affecting many key marine ecosystems. The chapter emphasises how biodiversity in the oceans is structured into ecosystems and the importance of interactions between different hierarchical levels within these ecosystems. Genomic approaches are described for the investigation of diversity at several different levels: taxonomic, genetic and functional.

Chapter 2 looks at the rapidly evolving field of metagenomics or environmental sequencing in which sequencing analysis is applied to communities of several (or many) species rather than single, isolated organisms. The emergence of metagenomics as a discipline is described, with particular reference to the important role played by marine ecosystems. The chapter emphasises the innovative nature of metagenomic approaches, particularly with respect to providing global views of ecosystems and as a means to describe species that cannot be cultivated in the laboratory. Recently developed approaches such as metatranscriptomics and metaproteomics are also discussed. Finally, this chapter provides some useful hints for researchers new to this field.

Chapter 3 discusses the relatively recent application of genomic methodologies to the study of marine organisms at the population level. The chapter looks both at studies at the DNA level, which address questions such as measuring biodiversity, mapping species boundaries and estimating drift in populations, and at the RNA level, which provides information about adaptation to particular habitats via selective processes or modifications in gene expression. Information about the capacity

of a species to adapt to perturbations of the environment are particularly important for predicting the effects of climate change or other anthropogenic factors. The chapter begins with an overview of the methodologies available for analyses at both the DNA and RNA level and then goes on to discuss the relative merits of the tools available with respect to the type of question that is being addressed. Finally, the chapter uses some examples to illustrate how genomic approaches have been and are being used in different contexts to answer questions about the structure and dynamics of populations and ecosystems in the marine environment.

The next two chapters concentrate on the impact of genomics on metazoan research. Chapter 4 describes how genomic approaches have been used to improve our understanding of the phylogeny of the metazoans. Early attempts to deduce the evolutionary history of the metazoans, based solely on morphological and developmental features, were susceptible to errors, due to factors such as morphological simplification of previously complex lineages and difficulties inherent in scoring and evaluating this type of feature. The introduction of gene-based phylogenetic methods led to a revolution in this field, leading to what is known as the "New View" of animal phylogeny, but certain relationships within the metazoans remained unresolved or controversial. Phylogenomic approaches, employing much more complete, "genome-scale" datasets, are now being employed to address these questions. This chapter describes how genome information is being used in phylogenomic studies and provides some examples of important insights produced by these approaches such as the recent re-evaluation of chordate relationships.

Chapter 5 looks specifically at the evolution of complexity in the animal lineage. Extension of work with model organisms to a much broader range of species, including many marine animals, has provide a more complete view of the evolution of complexity in the metazoans, particularly in terms of the underlying molecular genetic processes. These studies challenge the assumption that descendants are necessarily more complex than their ancestors and show that simplification has occurred in several lineages. By dissecting the molecular processes at work at different stages of evolution, these studies have shed light on several major evolutionary transitions, such as the transition to multicellularity or the origin of germ layers.

Chapter 6 focuses on the use of genomic methods to study marine algae. The term "algae" assembles a very diverse collection of taxonomic groups that all share one feature, the ability to carry out photosynthesis. These organisms play an important role in the planet's geochemical cycles and they are a rich source of novel biomolecules and bioprocesses. The important contribution of both genome sequencing and metagenomic analyses to the recent acceleration in our understanding of the biology of marine algae is discussed. The chapter also underlines the importance of the recent development of model organisms for several major algal groups.

Aquaculture is a rapidly expanding sector worldwide but the genomics resources available in this area remain poorly exploited. Chapter 7 discusses the application of genomic approaches to fish and shellfish and discusses how the data generated can be exploited to address aspects such as reproduction, growth and nutrition, product

quality and safety, and problems with pathogens. The chapter also looks at the evaluation of stock diversity and the potential for the use of genomic data in selection programs. Finally, the chapter discusses the use of genomic methods to study and monitor natural fish and shellfish populations and to understand interactions within their ecosystems.

The next chapter looks at the rapidly growing field of marine biotechnology. The chapter highlights the very broad spectrum of organisms that can be found in the seas in terms of phylogenetic diversity and discusses how both whole genome sequencing and large-scale metagenomics projects are rapidly expanding the genetic resources available for biotechnological exploitation of this part of the planet. As in several other domains, marine biotechnology has not developed to the same extent as its terrestrial counterpart, probably because of the difficulty in accessing the marine environment. The advent of genomic methodologies is changing this situation and exciting developments are expected in this domain in the coming years. The chapter describes the genomic resources that are available and provides examples of biotechnological products and processes that have been derived from a broad range of organisms including viruses, archaea, bacteria, algae, fungi and marine animals.

As mentioned above, the final chapter addresses some of the practical aspects of marine genomics. It looks at the different methods currently available for DNA sequencing and discusses the problem of data management. The latter has become an increasingly important consideration as the amount of data delivered by new sequencing technologies has expanded impressively over recent years. Bioinformatic processing of genomic data is then discussed including aspects such as EST clustering, genome assembly, gene prediction, assignment of gene functions and whole genome annotation. The chapter also provides an overview of transcriptome data analysis in particular those based on microarray hybridization technology.

Contents

Contributors

Michael J. Allen PML Applications, Plymouth Marine Laboratory, Plymouth PL1 3DH, UK, mija@pml.ac.uk

Luca Bargelloni Department of Public Health, Comparative Pathology and Veterinary Hygiene, Faculty of Veterinary Medicine, University of Padova, I-35020 Legnaro, Italy, luca.bargelloni@unipd.it

Thomas Bekel BRF/Computational Genomics group, CeBiTec, Universität Bielefeld, D33594 Bielefeld, Germany, thomas.bekel@cebitec.uni-bielefeld.de

Jochen Blom BRF/Computational Genomics group, CeBiTec, Universität Bielefeld, D33594 Bielefeld, Germany, jblom@cebitec.uni-bielefeld.de

François Bonhomme Université Montpellier II, Institut des Sciences de l'Evolution de Montpellier (ISEM), 34095 Montpellier Cedex, France, bonhomme@univ-montp2.fr

Pierre Boudry Ifremer, UMR M100 Physiologie and Ecophysiologie des Mollusques Marins, Technopôle Brest-Iroise, F-29280 Plouzané, France, pierre.boudry@ifremer.fr

Viviane Boulo Ifremer, CNRS Université de Montpellier 2, UMR 5119 ECOLAG CC 80, F-34095 Montpellier cedex 5, France, viviane.boulo@ifremer.fr

Jean-Paul Cadoret Physiology and Biotechnology Laboratory, Ifremer, 44311 Nantes, France, jean.paul.cadoret@ifremer.fr

M. Leonor Cancela Laboratory of Molecular Biology of Marine Organisms, Centre of Marine Sciences (CCMAR), University of Algarve, P-8005-139 Faro, Portugal; Faculty of Marine and Environmental Sciences, University of Algarve, P-8005-139 Faro, Portugal, lcancela@ualg.pt

G.R. Carvalho Molecular Ecology and Fisheries Genetics Laboratory, School of Biological Sciences, Bangor University, Bangor, Gwynedd LL57 2UW, UK, g.r.carvalho@bangor.ac.uk

Melody S. Clark Ecosystems, British Antarctic Survey, Natural Environment Research Council, Cambridge CB3 0ET, UK, mscl@bas.ac.uk

J. Mark Cock Laboratoire International Associé Dispersal and Adaptation in Marine Species, CNRS, UMR 7139, 29682 Roscoff Cedex, France, cock@sb-roscoff.fr

Susana M. Coelho UPMC Univ. Paris 06, The Marine Plants and Biomolecules Laboratory, UMR 7139, 29682 Roscoff Cedex, France, coelho@sb-roscoff.fr

Jonas Collén Station Biologique de Roscoff, UMR 7139, CNRS UPMC, 29682 Roscoff cedex, France, collen@sb-roscoff.fr

F.O. Costa Departamento de Biologia, Universidade do Minho, 4710-057 Braga, Portugal, fcosta@bio.uminho.pt

S. Creer Molecular Ecology and Fisheries Genetics Laboratory, School of Biological Sciences, Bangor University, Bangor, Gwynedd LL57 2UW, UK, s.creer@bangor.ac.uk

Jorge Dias Laboratory of Aquaculture, Centre of Marine Sciences (CCMAR), University of Algarve, P-8005-139 Faro, Portugal, jorgedias@ualg.pt

Michael Dondrup Computational Biology Unit, Unit BCCS, Thormøhlensgate 55, N-5008 Bergen, Norway, michael.dondrup@bccs.uib.no

Frank Oliver Glöckner Max Planck Institute for Marine Microbiology, 28359 Bremen, Germany, fog@mpi-bremen.de

Alexander Goesmann BRF/Computational Genomics group, CeBiTec, Universität Bielefeld, D33594 Bielefeld, Germany, agoesman@cebitec.uni-bielefeld.de

Nigel Grimsley Laboratoire ARAGO, UMR7628 Modèles en biologie cellulaire et évolutive, 66651 Banyuls-sur-mer, France, grimsley@obs-banyuls.fr

Bruno Guinand Université Montpellier II, Institut des Sciences de l'Evolution de Montpellier (ISEM), 34095 Montpellier Cedex, France, bruno.guinand@univ-montp2.fr

Svenja Heesch UPMC Univ. Paris 06, The Marine Plants and Biomolecules Laboratory, UMR 7139, 29682 Roscoff Cedex, France; Laboratoire International Associé Dispersal and Adaptation in Marine Species, CNRS, UMR 7139, 29682 Roscoff Cedex, France, svenja.heesch@sams.ac.uk

Kolja Henckel BRF/Computational Genomics group, CeBiTec, Universität Bielefeld, D33594 Bielefeld, Germany, khenckel@cebitec.uni-bielefeld.de

Arnaud Huvet Ifremer, UMR M100 Physiologie and Ecophysiologie des Mollusques Marins, Technopôle Brest-Iroise, F-29280 Plouzané, France, arnaud.huvet@ifremer.fr

Sebastian Jaenicke BRF/Computational Genomics group, CeBiTec, Universität Bielefeld, D33594 Bielefeld, Germany, sjaenick@cebitec.uni-bielefeld.de

Didier Jollivet Equipe Evolution et Génétique des populations Marines (EGPM), UMR 7144, CNRS-UPMC Station Biologique de Roscoff, 29682 Roscoff Cedex, France, jollivet@sb-roscoff.fr

Lutz Krause BioAnalytical Science Department, Nestlé Research Center, CH-1000 Lausanne 26, Switzerland, lutz.krause@rdls.nestle.com

Vincent Laizé Laboratory of Molecular Biology of Marine Organisms, Centre of Marine Sciences (CCMAR), University of Algarve, P-8005-139 Faro, Portugal, vlaize@ualg.pt

Sylvie Lapègue Laboratoire de Génétique et Pathologie, Ifremer, F-17390 La Tremblade, France, sylvie.lapegue@ifremer.fr

M. Le Goff-Vitry 32 rue Dupetit Thouars, 49 000 Angers, France, mclgv@yahoo.co.uk

Yannick Le Parco Station Marine d'Endoume, UMR 6540 DIMAR, CNRS & Université de la Méditerranée, Marseille, France, yannick.leparco@univmed.fr

Ricardo Leite Laboratory of Molecular Biology of Marine Organisms, Centre of Marine Sciences (CCMAR), University of Algarve, P-8005-139 Faro, Portugal; Faculty of Marine and Environmental Sciences, University of Algarve, P-8005-139 Faro, Portugal, rleite@ualg.pt

Burkhard Linke BRF/Computational Genomics group, CeBiTec, Universität Bielefeld, D33594 Bielefeld, Germany, burkhard.linke@cebitec.uni-bielefeld.de

Antonios Magoulas Institute of Marine Biology and Genetics (IMBG), Hellenic Centre for Marine Research (HCMR), 71003 Heraklion, Greece, magoulas@her.hcmr.gr

Ferdinand Marlétaz Station Marine d'Endoume, UMR 6540 DIMAR, CNRS & Université de la Méditerranée, Marseille, France, ferdinand.marletaz@me.com

L. Medlin Laboratoire Arago, Observatoire Océanologique, 66651 Banyuls sur Mer, France, medlin@obs-banyuls.fr

K. Metfies Alfred Wegener Institute, Am Handelshafen 12, D-27570 Bremerhaven, Germany, katja.metfies@awi.de

Anke Meyerdierks Max Planck Institute for Marine Microbiology, Celsiusstrasse 1, 28359 Bremen, Germany, ameyerdi@mpi-bremen.de

Sara Mira Laboratory of Molecular Biology of Marine Organisms, Centre of Marine Sciences (CCMAR), University of Algarve, P-8005-139 Faro, Portugal, smira@ualg.pt

Virginie Mittard-Runte BRF/Computational Genomics group, CeBiTec, Universität Bielefeld, D33594 Bielefeld, Germany, vrunte@cebitec.uni-bielefeld.de

Hervé Moreau Laboratoire ARAGO, UMR7628 Modèles en biologie cellulaire et évolutive, 66651 Banyuls-sur-mer, France, h.moreau@obs-banyuls.fr

Heiko Neuweger BRF/Computational Genomics group, CeBiTec, Universität Bielefeld, D33594 Bielefeld, Germany, hneuwege@cebitec.uni-bielefeld.de

Einar E. Nielsen Population Genetics Lab, Department of Inland Fisheries, Danish Institute for Fisheries Research, Technical University of Denmark, DK-8600 Silkeborg, Denmark, een@aqua.dtu.dk

Josep V. Planas Facultat de Biologia, Departament de Fisiologia, Universitat de Barcelona, E-08028 Barcelona, Spain, jplanas@ub.edu

Joel Querellou Ifremer, UMR 6197, Microbiology of Extreme Environments, Centre de Brest DEEP/LM2E, BP70, 29280 Plouzane, France, joel.querellou@ifremer.fr

Florian Raible Max F. Perutz Laboratories, Campus Vienna Biocenter, University of Vienna, A-1030 Vienna, Austria, florian.raible@mfpl.ac.at

Nerea Roher Facultat de Biologia, Departament de Fisiologia, Universitat de Barcelona, E-08028 Barcelona, Spain, nerea.roher@uab.cat

Elena Sarropoulou Institute of Marine Biology and Genetics, Hellenic Centre for Marine Research (HCMR), 71003 Heraklion, Greece, sarris@her.hcmr.gr

Susanne Schneiker-Bekel International NRW graduate school in Bioinformatics and Genome Research, CeBiTec, Universität Bielefeld, D33594 Bielefeld, Germany, Susanne.Schneiker@cebitec.uni-bielefeld.de

Patrick R.H. Steinmetz Department for Molecular Evolution and Development, University of Vienna, A-1090 Vienna, Austria, patrick.steinmetz@univie.ac.at

Arnaud Tanguy Equipe Evolution et Génétique des populations Marines (EGPM), UMR 7144, CNRS-UPMC Station Biologique de Roscoff, 29682 Roscoff Cedex, France, atanguy@sb-roscoff.fr

C.S. Tsigenopoulos Hellenic Centre for Marine Research, Institute of Marine Biology and Genetics, Heraklion, Crete, Greece, tsigeno@her.hcmr.gr

Frédérique Viard Equipe Evolution et Génétique des populations Marines (EGPM), UMR 7144, CNRS-UPMC Station Biologique de Roscoff, 29682 Roscoff Cedex, France, viard@sb-roscoff.fr

Filip A.M. Volckaert Laboratory of Animal Diversity and Systematics, Katholieke Universiteit Leuven, B-3000 Leuven, Belgium, Filip.Volckaert@bio.kuleuven.be

Chapter 1
Genomics in the Discovery and Monitoring of Marine Biodiversity

G.R. Carvalho, S. Creer, Michael J. Allen, F.O. Costa, C.S. Tsigenopoulos, M. Le Goff-Vitry, A. Magoulas, L. Medlin, and K. Metfies

Abstract Marine biodiversity encompasses a range of hierarchical levels, including genetic, species, ecosystem and functional diversity. Interactions among such levels determine ultimately the distribution and abundance, as well as evolutionary potential and resilience, of marine taxa. In the face of accelerating environmental change and ecosystem disruption, the detection and monitoring of structural and functional components becomes increasingly urgent. Classical ecological and conservation marine studies focused on species and communities- the emphasis has now shifted to enhancing our understanding of the relationships among the various components of biodiversity, especially their role in ecosystem services such as global nutrient recycling and climate. Here, we highlight the significance of genomics and genetic principles in elucidating the interactions among different biological levels of diversity, from genetic and cellular, to community and ecosystem-level processes. Genomic methods are especially powerful in disclosing previously undetected taxonomic (e.g. DNA barcoding), genetic (e.g. 454 sequencing) and functional (e.g. gene expression, analysis of metabolites) diversity, including the identification of new species and metabolic pathways.

1.1 Marine Biodiversity and Genomics – A Global Perspective

1.1.1 Marine Biodiversity: Structural and Functional Components

There are several indisputable facts about marine biodiversity: the exceptional levels that are found in our oceans; the accelerating range of threats to which marine taxa are exposed, and its importance in ecosystem functioning.

G.R. Carvalho (✉)
Molecular Ecology and Fisheries Genetics Laboratory, School of Biological Sciences, Bangor University, Bangor, Gwynedd LL57 2UW, UK
e-mail: g.r.carvalho@bangor.ac.uk

J.M. Cock et al. (eds.), *Introduction to Marine Genomics*,
Advances in Marine Genomics 1, DOI 10.1007/978-90-481-8639-6_1,

First, the levels of biodiversity in our oceans are exceptional. Oceans encompass approximately 70% of the planet's surface, and offer a diversity of habitats to support 28 phyla of animals, 13 of which are endemic to the marine realm, compared with 11 phyla from terrestrial habitats (only one endemic, Angel 1992). High species and phyletic diversity is commensurate with a corresponding plethora of life-styles from floaters and swimmers, to those withstanding partial aerial exposure in intertidal zones or inhabiting deep sea hydrothermal vents at >3,500 m. Moreover, because we know that life originated in the seas, marine taxa have been evolving for up to 2.7 billion years longer than terrestrial counterparts. Almost all extant phyla have marine representatives, compared to only approximately half having terrestrial representatives (Ray 1991). Importantly also, as advanced taxonomic methods become available (Savolainen et al. 2005), and as new technologies enable exploration of previously inaccessible habitats, many new marine species are being discovered (e.g. Santelli et al. 2008). These include both microscopic or microbial taxa (Venter et al. 2004, Goméz et al. 2007), and also more familiar larger organisms such as fish, crustaceans, corals and molluscs (Bouchet 2005). For example, the marine bryozoan, *Celleporella hyalina*, which was thought to be a single cosmopolitan species. DNA barcoding and mating tests revealed that geographic isolates comprised >20 numerous deep, mostly allopatric, genetic lineages (Gómez et al. 2007; Fig. 1.1). Moreover, such lineages were reproductively isolated, yet share very similar morphology, indicating rampant cryptic speciation. Such hidden diversity is exemplified by recent discoveries in waters off Australia, where over 270 new species of fish, ancient corals, molluscs crustaceans and sponges new to science were discovered among underwater mountains and canyons off Tasmania http://www.csiro.au/science/SeamountBiodiversity.html). This is also exemplified in marine transition zones between biogeographical provinces (e.g. between the Lusitanian and boreal provinces, Maggs et al. 2008). All but one of the cosmopolitan diatom species investigated to date are composed of multiple cryptic species (see review in Medlin 2007).

Second, despite the high levels of extant species diversity, marine systems are exposed to excessive and accelerating threats from environmental change and human activity. Threats such as pollution, over-exploitation, eutrophication, invasive species and climate change cause changes in distribution and abundance (Worm et al. 2006), as well as localized extinctions. It is important not only to understand the mechanisms and consequences of such change to generate predictions of response, but also importantly to enhance opportunities for recovery, resilience and reversibility of disturbance (Palumbi et al. 2008a). Third, marine biodiversity underpins the extent and dynamics of ecosystem functioning. Marine biota play a key role, for example, in global nutrient recycling and climate, and provide man with a multitude of resources and ecosystem services (products and processes provided by the natural environment), including carbon storage, atmospheric gas regulation, waste treatment, food provision and raw materials. Indeed, marine algae contribute up to 40% of global photosynthesis. Globally, such marine ecosystem services have been estimated to be valued in excess of $8.4 trillion per year for open oceanic systems and $12.6 trillion for coastal ecosystems (Costanza et al. 1997). For example,

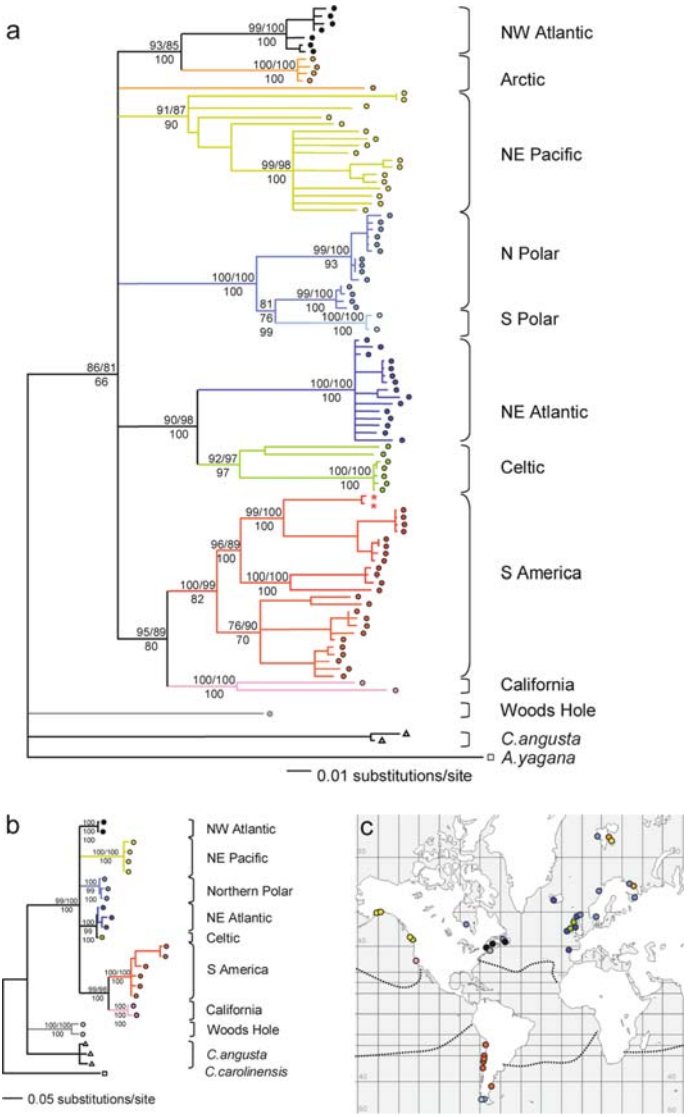

Fig. 1.1 Cryptic speciation of the marine bryozoan, *Celleporalla hyalina*. (After Gómez et al. 2007). (**a**) Maximum-likelihood tree of haplotype data from the barcoding gene, COI. (**b**) Maximum-likelihood tree of the nuclear gene, elongation factor, EF-1a haplotype data. Individuals traditionally described as *C. hyalina* are marked with circles coloured according to the geographical region listed to the right. (**c**) Map of the sample locations included in the genetic analysis. Coloured *circles* indicate the major lineage according to the phylogenetic analysis. *Dotted lines* indicate the limits of the temperate oceans (20C isotherm)

the deep sea plays a key role in ecological and biogeochemical processes globally. The conservation of such biodiversity is a priority to secure sustainable functioning of the world's oceans.

Fourth, there is increasingly compelling evidence that sustainable ecosystem services depend upon diverse biota (reviewed by Palumbi et al. 2008a). For example, using several independent indicators of ecosystem functioning and efficiency, a global-scale case study from 116 deep sea sites showed that ecosystem functioning was exponentially related to deep sea biodiversity (Danovaro et al. 2008, Fig. 1.2). Such a relationship, and similar such studies (Palumbi et al. 2008b), indicate that a higher biodiversity supports higher rates of ecosystem processes and an increased efficiency in which processes such as carbon production and cycling are carried out, with an associated increase in the biomass of key taxa. A loss of biodiversity, at least in this case, is likely therefore to be associated with a marked decline in ecosystem function.

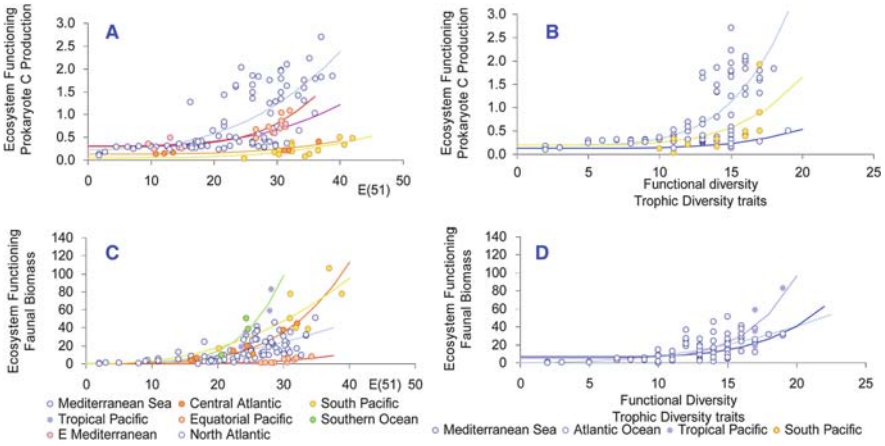

Fig. 1.2 Relationship between biodiversity and ecosystem function (after Danovaro et al. 2008). Data show that as the biodiversity of benthic meiofauna increases (estimated geographically, (**a**, **c**) and from diversity of trophic traits (**b**, **d**)), so do various proxies of ecosystem function (e.g. prokaryote C production (**a**, **b**), faunal biomass (**c**, **d**))

The combination of high and apparently dynamic species diversity in our seas with the global interdependence of biodiversity, energy flow and nutrient cycling provides a compelling case for increasing the rate at which we can identify novel taxa. In this chapter, we focus on how recent developments in genomic technologies and associated genetic theory provide an empirical and conceptual framework for the description and conservation of marine biodiversity. Crucially, we highlight the role that genomics might play in elucidating the linkages among different biological levels of diversity, from genetic and cellular, to community and ecosystem-level processes (Fig. 1.3). Genomic regions associated with particular phenotypes can be identified using quantitative trait loci (QTL, Witham et al. 2008). Candidate genes can then be identified from DNA sequence information, and anchored to genetic maps with sequence-tagged markers such as simple sequence repeats (SSR). Second, evidence for selection at candidate genes can be examined by screening natural populations for selective sweeps or local linkage disequilibrium, and/or rates

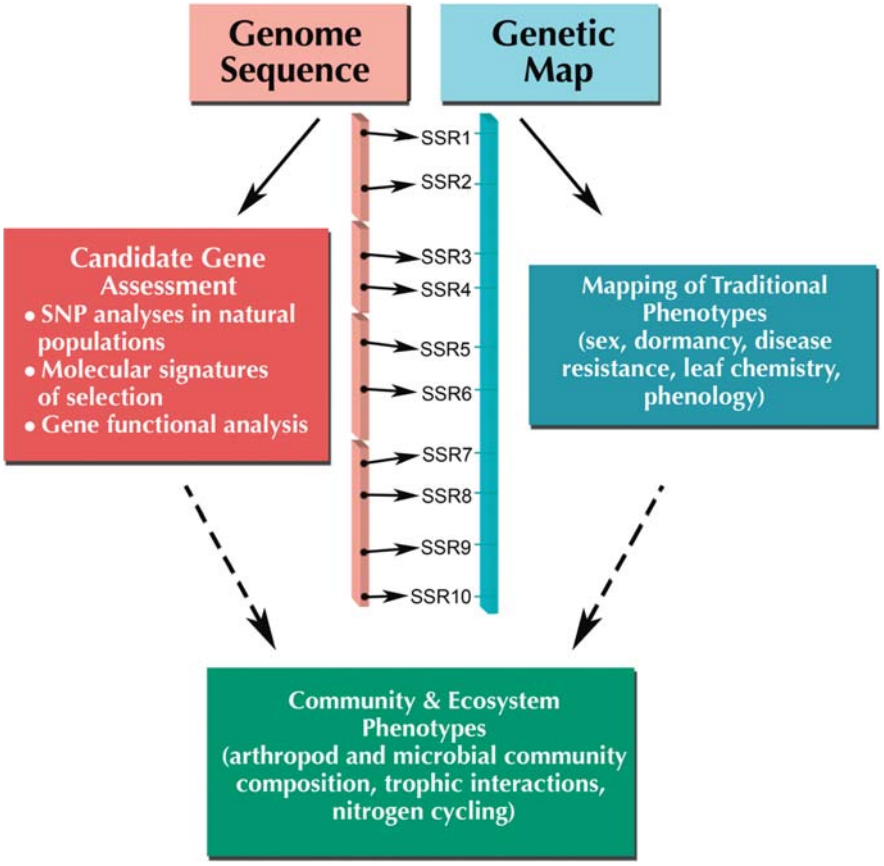

Fig. 1.3 Exploring linkages between genomic information and ecosystem function (after Witham et al. 2008). See text for further information. SSR, simple sequence repeats

of synonymous and non-synonymous polymorphisms. Third, genes with signature of selection can be subjected to functional analysis such as gene expression using microarrays or quantitative PCR. Finally, the link to community and ecosystems can be tackled using common garden experiments and wild populations.

Genomic applications to management of marine biota are at an early phase, but we identify opportunities in relation to relevant genetic and evolutionary principles. Exciting and novel opportunities now exist to detect and recognize life history stages and new species that hitherto remained inaccessible to classical taxonomic methods, such as high throughput DNA barcoding (Hebert et al. 2003) and metagenomics (Allen and Banfield 2005, DeLong 2005). Moreover, the ability to target specific gene structure and function has led to the discovery of new metabolic pathways (Venter et al. 2004) thereby increasing our ability to understand and explore functional links among marine biota, ecological processes and novel marine products. Such interrelationships coincide with the ecosystem approach to conservation

and management (Pikitch et al. 2004), whereby the focus of effort has shifted from individual components such as populations and species, to the interdependence of communities and trophic levels. Initially, we will describe briefly the nature of marine biodiversity, and then examine salient advances that have revolutionized our ability to study its dynamics and distribution. The synthesis is not intended to be comprehensive, but rather illustrates approaches and applications that are developed further in accompanying chapters.

1.1.2 The Nature of Marine Biodiversity

Biodiversity is an all inclusive term to describe the variety of living organisms and their environments, and it can be divided into four main components: (1) *Genetic diversity*, that refers to the within-species genetic variation, a crucial determinant of the ability of populations and species to withstand and recover from environmental perturbations; (2) *Species diversity*, which describes the variety of species or other taxonomic groups within an ecosystem, and represents the key identifiable units that determine the complexity and resilience of habitats; (3) *Ecosystem diversity*, that refers to the range of biological communities and the dynamics and nature of their interdependence and interactions with the environment. Diversity at this level is distinct from (1) and (2) in that it comprises both a living (biotic) and non-living (abiotic) component; (4) *Functional diversity*, which includes the array of biological processes, functions or characteristics of a specific ecosystem. Some argue that functional diversity may well be the most meaningful way of assessing biodiversity because it does not necessitate the cataloguing of all species within an ecosystem, and may thereby provide a tractable way for conserving marine natural systems. Although such a view would be well supported by genomic approaches, its application is constrained by the usual need to relate diversity to function at different spatial scales (Bulling et al. 2006, Naeem 2006), and the fact that many species and their function are, as yet, undefined.

1.1.3 Empirical and Conceptual Advances

Genomics within the current context is a scientific discipline that studies the structure, function and diversity of genes and gene products in the genome of a species with the aim of understanding the relationship between an organism and its biotic and abiotic environment. Various advances at different biological levels have revolutionized our ability to analyze genome structure and function (Wilson et al. 2005): (1) *At the level of DNA*: including the development of high throughput DNA sequencing, and recent enhancement using next generation sequencing technologies (Rothberg and Leamon 2008); (2) *At the level of gene expression* ("transcriptomics") using microarray or digital gene expression technology; (3) *At the level of protein products*, with improved analysis of proteins ("proteomics") using tandem

mass spectrometry and (4) *At the metabolic level*, with the detailed analysis of low-molecular-weight cellular constituents ("metabolomics") using a diversity of enhanced analytical techniques, such as pyrolysis gas chromatography and infrared spectrometry. Each has been associated with major improvements in throughput and volume of data produced, as described in Chapter 3. In addition to the developments in molecular biology, the genomics revolution is associated with three other technological advances that occurred in the 1990s in microtechnology, computing and communication (Van Straalen and Roelofs 2006).

Microtechnology: the ability to examine very small molecules on the scale of a few micrometers using new laser technology was crucial for the deployment of the gene chip.

Computing technology: The vast amount of DNA data generated by high throughput sequencing, and the analysis of expression matrices and protein databases requires considerable computing power for efficient bioinformatic analyses. The advent of high-speed computers and data-storage methods of immense capacity underpins genomic approaches.

Communication technology: The ability to access and integrate global databases using the Internet in real time is an essential component of both the design and interpretation of genomic data, especially as attention focuses increasingly on non-model organisms (Cock et al. 2005).

The essence then of the genomics approach for exploring biodiversity is to analyze an abundance of individuals and/or genes simultaneously, thereby enhancing the opportunities for matching genetic and phenotypic diversity at the component biological levels identified above.

Conceptually, there are three salient points to note in the application of genomics to the analyses of biodiversity. First, although the field of genomics was developed initially based on applications to model organisms such as baker's yeast (*Saccaromyces cerevisiae*), the fruit fly (*Drosophila melanogaster*), a nematode worm (*Caenorhabditis elegans*), mouse ear cress (*Arabidopsis thaliana*), and more recently the mouse (*Mus musculus*), techniques and interest now allow improved access to ecologically-well characterized species (Vera et al. 2008, Witham et al. 2008). Laboratory-based models have been useful for elucidating our understanding of basic processes governing growth and development, but generally are more restricted when predicting responses to environmental change and role in ecosystem processes. The recent increase in DNA sequence data, for example, of marine organisms including diatoms, sea urchins, *Hydra*, fish, shrimps, and brown algae (Wilson et al. 2005, Cock et al. 2005), together with the availability of rapidly increasing expression sequence tag (EST) libraries, has significantly enhanced the representation of marine taxa (see Chapter 3 and 7). Moreover, the targeting of so-called "key species" for genomic approaches across the evolutionary tree (Cock et al. 2005) as models for phylogenetically similar organisms will further broaden the range of life styles, adaptations and habitats that can be examined. Second, there has been a

surge of activity in the analysis of microbial communities by isolating and sequencing large fragments of DNA directly from the environment, an approach known as "metagenomics" (see Chapter 2). Such work has led to the discovery of novel biochemical pathways (e.g. Peers and Price 2006) and novel organisms (Massana et al. 2006). For example, the application of "whole-genome shotgun sequencing" to microbial populations collected from the Sargasso Sea identified at least 1,800 genomic species based on sequence relatedness, including 148 previously unknown bacterial phylotypes (Venter et al. 2004). The Sargasso Sea study additionally identified over 1.2 million previously unknown genes, including more than 782 new rhodopsin-like photoreceptors, revealing unprecedented levels of phyletic and functional oceanic microbial diversity. Such approaches increase further the ability to analyze directly samples from the wild, and promise to revolutionize our ability to characterize gene function in relation to organismal and ecosystem processes. Finally, the enhanced awareness that biodiversity underpins ecosystem resilience and recovery from perturbations (reviewed in Palumbi et al. 2008a), has shifted interest and focus of environmental management from individuals, populations and species, to communities and ecosystems, with particular attention on functional relationships between biotic diversity and ecosystem processes. Such a focus on relationships is ideally suited for genomic applications because it is the linkages between gene structure, function and phenotypic diversity at the cellular, metabolic and ecological processes that characterize variability in patterns of productivity, nutrient and energy flow.

Having provided a brief overview of key developments in marine genomics in relation to marine biodiversity, we now examine some molecular approaches for identifying marine species and classifying functional capacity of marine assemblages.

1.2 Molecular Identification of Marine Biodiversity

With the increasing focus on the relationship between biodiversity and ecosystem functioning on such processes as elemental cycling, production and trophic transfer (Chapin et al. 1997, Duffy and Stachowicz 2006), it is crucial to obtain realistic estimates of species diversity and quantification of functional capability. Particular challenges in the assessment of biodiversity in the marine environment arise from limited access to certain habitats, communities and the patchy distribution of species in pelagic environments. Accordingly, new genomic approaches have recently been developed that provide the means to access previously undetected biodiversity.

Species identification is the critical starting point of any research in marine biology. Conventional identification approaches based on phenotypic characters may be apparently straightforward. However there are various situations in which they may fail or have limited efficiency, such as cryptic species, inherently difficult taxonomic groups, or taxonomically ambiguous eggs and larvae (Kochzius et al. 2008). The discovery of new marine habitats and associated new species, increased threats to

marine species from on-going environmental change and habitat disturbance make it increasingly important to develop rapid and robust ways of describing and cataloguing marine biodiversity. Molecular tools of universal implementation, such as the recently proposed DNA barcodes ("a rigorously standardized sequence of a minimum length and quality from an agreed-upon gene, deposited in a major sequence database, and attached to a voucher specimen whose origins and current status are recorded") can counter conventional limitations, providing a simple, yet robust system to unambiguously identify not only whole individuals, but eggs, larvae and body fragments. Such approaches necessarily include a range of molecular identification strategies based on the analysis of homologous gene regions (e.g. COI, 16S, 18S and ITS) for delimiting species boundaries and in their discovery, as well as associated "genome screens" for the identification of functional capacity.

Hebert et al. (2003) introduced the concept of a DNA barcode, and proposed a new global approach to species identification, which offered great promise to counter many limitations of the classical taxonomic approach. The new approach was based on the premise that the sequence analysis of a short fragment of a single gene (eg, cytochrome c oxidase subunit 1, CO1), enables unequivocal identification of many animal species. Hence, the DNA barcode would provide a standardised tool for fast, simple, robust and precise species identification. The rationale and approach of DNA barcoding are essentially the same whichever region of the genome is selected. The basic premise is that for each currently known species an unequivocal match can be established with the DNA barcode obtained by comparing the same DNA regions. The resultant "matching hypothesis" underpins the rationale for implementing the new molecular bioidentification system, which importantly also extends and incorporates, rather than substitutes, the classical Linnaean system (Costa and Carvalho 2007). It is also assumed that a low between- versus within-species divergence represents reproductively isolated entities according to the biological species concept, thereby conforming to the Linnaean binomial system (Goméz et al. 2007). Reference barcoded specimens of each species that have been identified by experts are deposited in a museum and therefore available for double-checking and for long-term study. Once this reference database is complete, it can be used to assign an unknown sample to a known species.

Whilst COI-based molecular barcoding is gaining momentum, COI does not work effectively for the molecular identification of all eukaryotic taxa and different markers are used for bacteria, Archaea and viruses (see below). This is predominantly because in some eukaryotes, such as the nematoda, the COI gene is characterised by unusual molecular evolutionary rates and processes. In nematodes, COI has high mutational rates and is very A+T rich with biased substitution patterns (Blouin et al. 1998, Blouin 2000) making it rather unsuitable for DNA taxonomy studies (Hajibabaei et al. 2007). Such challenges are demanding the acknowledgement of alternative markers for certain taxa and attempts are being made to accommodate suites of markers into the DNA barcoding movement. Examples of alternative markers include ribosomal genes that have been used for decades to identify multiple suites of microbial eukaryotes. It was first demonstrated in the 1960s that ribosomal genes (rDNA) and their gene products (rRNA) could be used for the

taxonomic classification of microbial species (Doi and Igarashi 1965, Dubnau et al. 1965, Pace and Campbell 1971a, b). Over the past decades the comparative analysis of homologous gene sequences has become an indispensable approach to gain new insights into the phylogeny and diversity of microbial organisms (Díez et al. 2001, Evans et al. 2007). The genes coding for rRNA are particularly well suited for phylogenetic analysis, because they are universal, found in all cellular organisms; they are of relatively large size; and they contain both highly conserved and variable regions with no evidence for lateral gene transfer (Woese 1987). The number of ribosomal sequences is continuously growing. Currently it is in the range of ~550,000 sequences (Pruesse et al. 2007). Several publications describe the power of ribosomal sequences to identify both prokaryotic and eukaryotic micro-organisms (Amann et al. 1990; Simon et al. 2000; Groben et al. 2004). Direct cloning and sequencing of the small subunit ribosomal DNA (18S rDNA) from natural samples has, for example, permitted a broader view of the structure and composition of picoplankton communities (Giovannoni et al. 1990; López-Garcia et al. 2001, Medlin et al. 2006).

1.2.1 Diversity and Functional Analyses of Microbial Communities

With >99% of its organisms predicted to be unculturable, the marine environment represents the largest untapped reservoir of genomic diversity on the planet (Beja 2004). Indeed, with an average of more than a million viruses per millilitre of seawater and no means of propagation outside of their host organism, the marine viral fraction represents perhaps the most poorly sampled biomass on Earth. To date, PCR-based approaches (e.g. clone libraries, denaturing gradient gel electrophoresis (DGGE) (Muyzer 1999) and restriction fragment length polymorphism (RFLP) analyses) have offered us rare glimpses of the enormous diversity of marine microorganisms (Archaea, prokaryotes, eukaryotes) derived from the analysis of specific homologous genes (e.g. 16S or 18S ribosomal DNA) used to identify species or phylotypes (Kemp and Aller 2004). DGGE allows the rapid analysis and comparison of microbial communities. Compositional diversity can be visualized using DGGE where each band in principle represents a microbial phylotype (Tzeneva et al. 2008). Sequencing of a single band provides information on the related taxon and can be used, for example, in phylogenetic analysis. Since viruses are obligate intracellular parasites, they lack ribosomal encoding regions such as 16S and 18S, thereby necessitating the development of markers specific for each broad family (such as RNA polymerase and DNA polymerase) in order to perform similar biodiversity studies (Allen and Wilson 2008).

Whereas such studies highlight the sheer abundance of natural biodiversity, no real functional significance can be attributed to the communities in question, because the molecular markers used for biodiversity appraisals have conserved genomic functions. For years, the Holy Grail of biodiversity research has been not to merely study biodiversity but to assess functional biodiversity, that is, to determine how diversity contributes to ecosystem functioning. The study of functional

diversity has now been attempted with a variety of techniques including functional and environmental genomics, transcriptomics (using microarrays and quantitative PCR), proteomics and metabolic/metabolomic studies. Such studies have provided crucial clues to the functional significance of natural biodiversity through the study of well-known metabolic pathways.

Initially, perhaps the greatest success in marine functional genomics was achieved using bacterial artificial chromosome (BAC) and fosmid (f1 origin-based cosmid vector) libraries that allowed up to 300 kb of DNA inserts to be studied in *E. coli*. A multiplex PCR screen for rRNA fragments allows the identification of clones from specific phylogenetic groups. Full sequencing can then allow metabolic genes to be identified in order to infer the metabolic potential of the specific organism or phylogenetic group. Such an advance was first achieved by Stein et al. 1996 with a marine Archaea population from coastal waters, followed by (Suzuki et al. 2004)'s picoplankton study. BAC libraries derived from marine environments are now commonplace, yet still often require hundreds, if not thousands of litres of seawater to produce. Recent modifications include the use of terminal restriction fragment length polymorphism (TRFLP) and internal transcribed space, length heterogeneity PCR (ITS-LH-PCR) to screen for rRNA genes (Suzuki et al. 2004, Babcock et al. 2007). Perhaps the greatest success story to come from the BAC approach is the identification of a novel light driven photon pump (now known as a bacteriorhodopsin) in an uncultured SAR86 (γ proteobacterial) group BAC fragment (Beja et al. 2000). Since then the BAC approach has been used with both natural marine communities and specific marine strains such as the gammaproteobacterium *Congregibacter litoralis* (Fuchs et al. 2007), Eastern and Pacific oysters *Crassostrea virginica* and *C. gigas* (Cunningham et al. 2006) and protochordate *Botryllus schlosseri* (de Tomaso and Weissman 2003).

In addition to large-scale conventional sequencing approaches, the development of pyrosequencing technologies now allows direct shotgun sequencing of environmental samples (Blow 2008) on a hitherto unforeseen scale (Huse et al. 2007, Huber et al. 2007, Mou et al. 2008). Such metagenomic approaches (Handelsmann 2004) have had major implications for the marine environment with its massive pool (quite literally) of unstudied and unculturable organisms. To date, combinations of chain-termination and pyrosequencing approaches have been used to study marine bacterial, eukaryotic and viral diversity, sampled from diverse marine environments (coastal, open ocean, surface, deep sea, coral, subseafloor) (Breitbart et al. 2004, Venter et al. 2004, Angly et al. 2006, Culley et al. 2006, Sogin et al. 2006, Bench et al. 2007, Biddle et al. 2008, Dinsdale et al. 2008, Quaiser et al. 2008, Williamson et al. 2008). It is important to note though, that such a huge volume of information has its limitations: the vast majority of genes currently being identified are of unknown function and their organism of origin is often a complete mystery. Sometimes it is possible to find suitable phylogenetic markers associated with DNA fragments, but more often than not, the precise origin of the genomic sequence remains unknown. However, such challenges are now tractable, especially with the advent of ultrasequencing approaches.

Not all functional biodiversity studies require huge sequencing efforts. For example, a novel approach has seen a virus isolate-specific microarray used to probe for the presence of genes in other natural virus isolates from the same family (Allen et al. 2007). The application of this genome wide approach can actually provide important information on the functional potential of an organism by identifying regions or single genes that are diverse enough to avoid hybridization or conserved enough to allow hybridization to occur. Providing sufficient information is available about the genes under study, a brief overview can be gleaned on the potential of an organism to perform specific metabolic functions and pathways.

1.2.2 Between the Microbes and Metazoans: Eukaryotic Protists

The functioning of marine ecosystems is based on organisms that belong to the pico-, nano-, and micro- plankton. These three size fractions comprise bacteria and eukaryotic protists, such as phytoplankton or heterotrophic protozoa. In contrast to the important ecological role of microbes in the marine environment, in many respects knowledge about their biodiversity is limited. Biodiversity assessment in this group of species is challenged by various factors. First, the size of organisms, particularly in the picoplankton fraction, renders microscopic observation and identification very difficult or even impossible. Eukaryotic marine picoplankton are a group of species with limited information on biodiversity or ecology, but extremely high ecological relevance. In open oceanic oligotrophic waters, picoplankton contribute up to 80% of the biomass and constitute an important prey for heterotrophic nanoflagellates (Ishizaka et al. 1997, Caron et al. 1999). Diatoms are another example of ecologically important eukaryotic microbes with limited information on their biodiversity and ecology. They account for at least 20% of the annual global carbon fixation by photosynthesis (Mann 1999). However, in this case species size is not the limiting factor for biodiversity assessment, but rather an uncertainty of the significance of small morphological variation among closely related species. The identification of diatoms is challenged by limited morphological differentiation among closely related species and cryptic speciation is widespread among cosmopolitan species (Evans et al. 2007, Medlin 2007). The consequence of cryptic species is an underestimation of species numbers. However, species dimorphisms can in turn lead to an overestimation of biodiversity. Dimorphic species occur in different morphological appearances usually in different stages of a haplo-diploint life cycle and have been misleadingly identified as different species. This phenomenon has been observed, for example, in the microalgal classes Cryptophyceae (Hoef-Emden and Melkonian 2003) and Prymnesiophyceae (see review in Billard 1994).

1.2.2.1 Ribosomal Probes

The continually growing number of available algal 18S rDNA-sequences, as for example, in the Ribosomal Database Project (RDP, Maidak et al. 2001), and in

phylogenetic analyses, permit the design of hierarchically organized probe sets that specifically target the 18S-rDNA from higher taxonomic levels down to the species level (Groben et al. 2004). Molecular probes are powerful in the detection and monitoring of microbial diversity, especially in the pico-sized fraction. They can be used in combination with a wide variety of hybridisation-based methods, such as fluorescent in situ hybridization (FISH) (Eller et al. 2007), RNA-based nucleic acid biosensors (Metfies et al. 2005) and DNA-microarrays (or Phylochips) (Metfies et al. 2006). FISH is a widely accepted and practiced approach for quantitative surveillance of microbes, but processing and quantitative analysis of FISH samples with fluorescent microscopy can be tedious, slow, and time-demanding because only one probe can be processed at a time because of the limited choice in fluorochromes. Nevertheless, the application of molecular probes and FISH techniques in combination with flow cytometry has greatly increased the speed and accuracy of microbial identification, as for example, in the context of characterizing picoplanktonic communities (Biegala et al. 2003). Surveillance of microorganisms could be further refined by the application of new chip-based hybridization formats, such as nucleic acid biosensors and DNA-microarrays/phylochips, that allow parallel identification and quantification of multiple taxa in a single experiment. The identification is based on solid phase hybridization of molecular probes, immobilized to the surface of the sensor chips to the rRNA or rDNA of the target species (Metfies et al. 2006). Their feasibility for the assessment of picoplankton community composition with a Phylochip (Fig. 1.4) has been shown for the picoeukaryotic *prasinophytes* (Gescher et al. 2008).

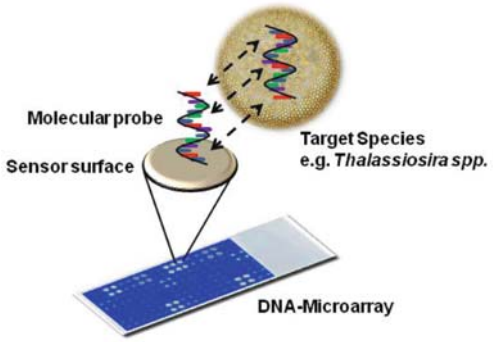

Fig. 1.4 Microbial species identification with DNA-Microarrays (PHYLOCHIPS). A PHYLOCHIP contains an ordered set of species specific molecular probes immobilized to the chip-surface. The identification of a microbial species is based on a specific interaction between the immobilized molecular probe and the complementary nucleic acid of the target species

1.2.2.2 Biodiversity Assessment at Sub-species Level

If information on species diversity within communities is to be supplemented by information on biodiversity below the species level, fingerprinting methods such as

restriction length polymorphism (RFLP), amplified fragment length polymorphism (AFLP) or microsatellites are appropriate methodologies. These well established methods are suited to resolve the degree of biodiversity within populations of certain species, and the utility of fingerprinting methods for microbial biodiversity assessment has been demonstrated in numerous publications (e.g. Adachi et al. 2003, John et al. 2004, Iglesias-Rodriguez et al. 2006, see review in Medlin 2007). In all studies to date, microbial populations in the oceans have been shown to have a distinct structure and gene flow can be restricted between geographically close areas. The oceans are far more fragmented in terms of population structure and gene flow than would have been believed decades ago.

1.2.3 Diversity and Ecological Analyses of Benthic Meiofaunal Communities

Soft-bottom benthic meiofauna are a ubiquitous, highly abundant community assemblage (ranging between 45 and 500 μm in size) that play a crucial role in marine ecosystem functioning and services. Comprised of between 50 (shallow water) and 90% (deep water) nematodes, meiofaunal assemblages contribute significantly to bentho-pelagic coupling in the form of nutrient cycling, water column processes, pollutant distribution, secondary production and stability of sediments (Snelgrove et al. 1997, Smith et al. 2000, Snelgrove et al. 2000). Despite their pivotal role in ecosystem functioning (Danovaro et al. 2008), our ability to construct mechanistic links between biodiversity and ecosystem services is significantly impeded by a poor understanding of global marine meiofaunal taxonomy and diversity. For example, a current estimate of global nematode diversity (c. 1 million species) remains a matter of conjecture (Lambshead and Boucher 2003). Only about 20,000 species have been described, around 4,000 of which are marine (Platt and Warwick 1983), and contemporary studies routinely recover between 30 and 40% of sampled taxa that are new to science (Lambshead and Boucher 2003). Such a knowledge gap is undoubtedly the result of the small size and apparent morphological conservatism of nematodes, rendering identification a considerable challenge to non-specialists. Indeed, even for experts, many male- or female-specific diagnostic morphological characters exhibit intraspecific variation and are restricted to mature individuals that can only be appraised using combinations of traditional light and electron microscopy combined with informed knowledge from specialist literature (Floyd et al. 2002, DeLey et al. 2005). From a logistical perspective, surveys have additionally shown that 120 times more scientific effort has to be expended in assigning only 10% of nematodes to known species, compared to parallel studies that successfully assigned all vertebrate morphospecies to known taxa (Lawton et al. 1998). The finding that more than 5,000 nematodes comprising up to 50 species can be recovered from a single 38 cm^3 temperate benthic fine sand sample, highlights the formidable challenge that meiobenthologists face when analyzing environmental samples.

In addition to the phylotype system employed to identify microbes, the past few years has witnessed the establishment of a DNA barcoding system of species identification for larger metazoans (Hebert et al. 2003). One drawback that arises when attempting to apply a barcoding framework to micro- and meiofaunal organisms is that in order to extract DNA from the organisms for PCR purposes, the whole animal is usually sacrificed, destroying the voucher specimen. An attractive solution to this problem lies in video capture and editing (VCE) microscopy (DeLey et al. 2005), which consists of recording a taxonomically relevant multifocal series of "3-D like" microscopy images as digital videoclips, that can act as "electronic voucher specimens". Images can then be deposited in appropriate publicly available repositories (e.g. NemAToL – http://nematol.unh.edu /) to be interrogated by the research community.

Commonly though, biodiversity questions revolve around communities of meiofaunal organisms, rather than single individuals. An alternative to the barcoding framework, analogous to methods employed in the study of prokaryotes, is to sequence a large number of individuals derived from environmental samples and recent research has highlighted the efficacy of an 18S rDNA molecular operational taxonomic unit (MOTU) scheme for environmental samples (Floyd et al. 2002, Blaxter and Floyd 2003, Bhadury et al. 2006). The MOTUs do not have any formal relationship with published species descriptions, but identifications can be achieved with existing databases, or future classifications in an approach that has been termed "reverse taxonomy" (Markmann and Tautz 2005). MOTU-based studies to date have investigated diversity via chain-termination sequencing of individual organisms, or cloning and sequencing a few hundred PCR products derived from environmental samples (Floyd et al. 2002, Blaxter et al. 2005). These data are highly informative, but molecular diversity accumulation curves typically do not reach an asymptote, suggesting that much larger sampling efforts are required to yield representative sets of organisms present in the meiobenthos (Markmann and Tautz 2005). As with many metagenetic (meta analyses of homologous gene regions) and metagenomic (meta analyses of multiple, fractionated genomes) endeavours, ultrasequencing and downstream microarray-based technologies are likely to offer significant opportunities to directly and simultaneously qualitatively assess molecular diversity (Creer 2008), akin to species richness measures. Moreover, a significant advantage of the use of VCE, is that functional diversity can be retrospectively assigned to environmental community representatives, by referring to the trophic mode, body size and life history strategy of the voucher specimens (Moens and Vincx 1997, Schratzberger et al. 2007).

1.2.4 DNA Barcoding and Fisheries

To illustrate certain principles of DNA barcoding and its role in the management of global marine resources, we consider briefly the application of the approach to marine fishes. In May 2004, an international consortium of organisations – the

Consortium for the Barcoding of Life (CBOL) (http://barcoding.si.edu/) – instigated the worldwide implementation of DNA barcoding, thus launching a unique large-scale horizontal genomics (one gene, many taxa) project. The first global DNA barcoding campaigns – the Fish Barcode of Life (FISH-BOL) (http://barcoding.si.edu/) and the All Birds Barcoding Initiative (ABBI) (http://www.barcodingbirds.org/) – were launched, with the intention of assembling a reference database of DNA barcodes for all fish and bird species respectively. FISH-BOL expects to complete most of the inventory of all known fish species of the world by 2010. To date, out of an estimated 29,112 fish species, approximately 5,600 have been barcoded using the CO1 target gene, representing 19% of global fish taxa. CBOL coordinates and promotes DNA barcoding on a worldwide scale, and endorses public access to DNA barcoding data. Both the Barcode of Life Database (BOLD) (Ratnasingham and Hebert 2007) and existing public genomic repositories (namely the GenBank of the National Center for Biotechnology Information (NCBI), the European Molecular Biology Laboratory (EMBL) and the DNA Data Bank of Japan (DDBJ)) provide free access to DNA barcoding data.

Fish provide a suitable model for testing the implementation of DNA barcoding at a worldwide scale, in terms of their phyletic diversity, economic value and environmental threats (Costa and Carvalho 2007, see also Chapter 7). Fish and fisheries resources comprise a key target group from which it is anticipated that DNA barcoding will bring larger and more immediate benefits (Lleonart et al. 2006). Such a system will offer a simple – and increasingly rapid and inexpensive – means of unambiguously identifying not only whole fish, but fish eggs and larvae, fish fragments, fish fillets and processed fish. Such a capability will generate more rigorous and extensive data on recruitment, ecology and geographic ranges of fisheries resources, improved knowledge of nursery areas and spawning grounds, as well as elucidation of taxonomy (Rock et al. 2008), with evident impacts at the fisheries management and conservation levels. For example, the possibility of rigorous identification of fish species from eggs and larvae could be particularly fruitful, since phenotypic identification of early life stages can be especially difficult (Pegg et al. 2006). A study testing the utility of molecular markers in species identification of fish eggs revealed that over 60% of the eggs were misidentified when phenotypic characters were used (Fox et al. 2005). Eggs from haddock and whiting may have been reported as cod eggs in previous surveys, possibly leading to an inflation of stock assessments of cod in the Irish Sea. Moreover, early stage haddock eggs were detected in the Irish Sea, indicating the presence of a spawning stock of this species previously unknown to that region (Fox et al. 2005). In the context of environmental change, induced, for instance, by global warming, the ability to rigorously identify fish species at all life history stages from egg to adult is particularly useful to assess escalating shifts in range distribution, spawning grounds and nursery areas (Fox et al. 2008).

Another valuable application of DNA barcoding is the identification of prey-remains from predators' stomach contents. Such studies could provide more detailed information about aquatic trophic chains, revealing which fish species are preyed upon by other fish species (Sigler et al. 2006) or seabirds (Phillips et al. 1999). This

ecological information could then be incorporated into models, as well as providing new data for use in management and conservation.

Potential forensic applications of fish DNA barcoding include the monitoring of fisheries quotas and by-catch, inspection of fisheries markets and products, the control of trade in endangered species, and improvements in the traceability of fish products (Ogden 2008). In Australian waters, for example, sharks are illegally captured, largely for their fins alone. Quality sharks' fins can sell for $6,000–$8,000/kg in Hong Kong, and it is estimated that globally more than 100 million sharks are killed every year. Sharks are particularly susceptible to overexploitation due to slow growth, their longevity and long gestation and low fecundity. Many species are morphologically very similar, and many are protected (R D Ward, pers comm.). A tool enabling precise identification of shark species from fins, from the fisheries boat to the soup in the restaurant, could be of great utility for law enforcement and conservation of endangered species (Chan et al. 2003). DNA barcoding could also be used for detection of fraudulent species substitutions in fish markets and fish food products, a practice that is generating concern among consumers (Wong and Hamer 2008). A striking example comes from the Red Snapper (*Lutjanus campechanus*), which is one of the most economically important fisheries in the Gulf of Mexico, and which has been subject to stringent fishing restrictions due to stock depletion. Marko et al. (2004) used sequences of the mtDNA gene cytochrome b, in an approach very similar to DNA barcoding, to show that up to 77% of the *L. campechanus* fillets were mislabelled in USA markets. Such a level of mislabelling may adversely affect estimates of stock size and contribute to the false impression among consumers and industry that the supply of fish is keeping up with demand.

Thus, DNA barcoding provides a standardised tool for describing and monitoring fish species diversity, not only in the wild, but also throughout the food supply chain in relation to legal enforcement and consumer protection (see Chapter 7). Moreover, a globally-accessible, standardised DNA barcoding data base means that non-experts may utilise the information to examine species identity, but importantly also allows a coordinated and extensive effort to document biodiversity from throughout species distributions.

1.2.5 Larvae in Marine Systems

The crucial influence of larval stages on population dynamics and on population connectivity, especially when adults are sessile, has long been acknowledged (Underwood and Fairweather 1989). Recent investigations of larval dispersal patterns have challenged traditional models, where larvae are dispersed as passive particles, by revealing the importance of larval behaviour and retention (Shanks and Brink 2005, Kinlan et al. 2005, Marta-Almeida et al. 2006). Accurately estimating larval distribution and abundance and assessing larval dispersal and recruitment patterns are all prerequisites for the sustainable management of marine resources (e.g. Taylor et al. 2002, Fox et al. 2005, Kochzius et al. 2008) and the restoration of

anthropogenically-degraded ecosystems, such as coral reefs (Bellwood et al. 2004, Shearer and Coffroth 2006). Including marine larvae in estimates of species richness is crucial, not only in relatively inaccessible or unexplored environments, such as the deep-sea (Grassle and Maciolek 1992), but also in more intensively documented coastal waters (Harding 1999, Lindley and Batten 2002).

Larvae, especially those of invertebrates, often represent an inscrutable component in marine fauna inventories (Mariani et al. 2003). Their ecological significance is still poorly understood, and the importance of larvae in marine food webs has only been recently highlighted (Bullard et al. 1999, Rosel and Kocher 2002, Metaxas and Burdett-Coutts 2006). Several factors account for this: small sizes, lack of diagnostic morphological characters allowing equivocal species identification, multiple larval stages, phenotypic plasticity, time and expertise required for accurate identification. When identification characters exist, they may often be restricted to certain geographical areas (reviewed in Rogers 2001). In some cases, the difficulty of rearing in captivity for reverse taxonomy is a further impediment to larval identification. Larvae, however, represent a fundamental component of marine ecosystems and a key step in the life cycle of many marine organisms.

Many molecular approaches have been developed to explore larval biology, and especially for species identification (Table 1.1). While most aim at high throughput analyses and are PCR-based, hence destructive, few preserve morphology. Despite the number of approaches developed, applications in ecological studies are still scarce.

Table 1.1 Recent molecular advances towards the identification of marine larvae

	Molecular approach	Organism targeted	Tested on larvae?	Used in ecological studies?
Hansen and Larsen (2005)	Single step nested multiplex PCR	*Mytilus edulis*/*Musculus marmoratus*, *Ensis* sp., *Myoida* spp., *Cariids*, *Spisula* spp., *Macoma*/*Abra* spp. (bivalves)	Yes	Yes
Patil et al. (2005)	COI nested PCR	*Crassostrea gigas* (Pacific oyster)	Yes	No
Noell et al. (2001)	Control region PCR	*Hyporhamphus melanochir*, *H. regularis* (Garfish)	Yes	No
Santaclara et al. (2007)	18S multiplex PCR and RFLP	*Xenostrobus secures*, *Mytillus galloprovincialis* (mussels)	Yes	No
Comtet et al. (2000)	ITS2 PCR and RFLP	*Bathymodiolus azoricus* (deep-sea vent bivalve)	Yes	No
Shearer and Coffroth (2006)	COI PCR and RFLP	*Agaricia agaricites*, *Porites astreoides* (scleractinian corals)	Yes	Yes
Karaiskou et al. (2007)	Cyt b PCR and RFLP	*Trachurus trachurus*, *T. mediterraneus*, *T. picturatus* (European horse mackerel species)	Yes	No

Table 1.1 (continued)

	Molecular approach	Organism targeted	Tested on larvae?	Used in ecological studies?
Hosoi et al. (2004)	D1/D2/D3 PCR and RFLP	*Barbatia* (*Abarbatia*) *virescens*, *Mytilus galloprovincialis*, *Pinctada martensii*, *Pinna bicolor*, *Chlamys* (*Azumapecten*) *farreri nipponensis*, *Anomia chinensis*, *Crassostrea gigas*, *Fuluvia mutica*, *Lasaea undulata*, *Moerella jedoensis*, *Theora fragilis*, *Ruditapes philippinarum*, *Paphia undulata* (bivalves)	Yes	No
von der Heyden et al. (2007)	Control region PCR and sequencing	*Merluccius capensis*, *M. paradoxus* (Cape hakes)	Yes	No
Richardson et al. (2007)	Cyt b PCR and sequencing	*Istiophorus platypterus*, *Makaira nigricans*, *Tetrapturus albidus*, *T. pfluegeri* (billfish), *Thunnus atlanticus*, *T. albacares*, *T. thynnus*, *T. obsesus*, *T. alalunga* (tuna)	Yes	Yes
Webb et al. (2006)	COI, 16S and 18S PCR and sequencing	Solasteridae, Echinidae, Ophiuridae, Echinidae, Syllida, Serpulidae, Nudibranchia, Gastropoda, Nemertea, Urochordata, Ophiuroidea, Terebellidea, *Adamussium colbecki*, (Antarctic scallop) *Parborlasia corrugatus*, (Antarctic nemertean)	Yes	No
Kirby and Lindley (2005)	16 S nested PCR and sequencing	*Echinocardium cordatum*, *Marthasterias glacialis*, *Spatangus purpureus*, *Amphiura filiformis* (echinoderms)	Yes	Yes
Barber and Boyce (2006)	COI PCR and sequencing	Stomatopod larvae (scleractinian corals)	Yes	Yes
Goffredi et al. (2006)	18S PCR and sandwich hybridisation	Thoracica, *Balanus glandula* (barnacles)	Yes	Yes
Jones et al. (2008)	18S PCR and sandwich hybridisation	*Carcinus maenas* (green crab), *Mytilus edulis* (native blue mussel), *Balanus* sp. (barnacle), *Osedax* sp. and *Ophelia* sp. (polychaetes)	Yes	Yes
Deagle et al. (2003)	PCR and DGGE	*Asterias amurensis* (northern Pacific seastar)	Yes	No

Table 1.1 (continued)

	Molecular approach	Organism targeted	Tested on larvae?	Used in ecological studies?
Fox et al. (2005)	TaqMan PCR	*Gadus morhua* (cod)	Yes	Yes
Vadopalas et al. (2006)	Real-time PCR	*Haliotis kamtschatkana* (pinto abalone)	Yes	No
Morgan and Rogers (2001)	Microsatellites	*Ostrea edulis* (flat oyster)	Yes	No
Zhan et al. (2008)	Microsatellites	*Chlamys farreri* (Zhikong scallop)	Yes	No
Barki et al. (2000)	AFLP	*Parerythropodium fluvum fluvum* (soft coral)	Yes	No
Arnold et al. (2005)	Blot hybridization with18S rRNA targeted probes	*Mercenaria* (bivalve)	Yes	Yes
Kochzius et al. (2008)	Microarrays	*Boops boops*, *Engraulis encrasicolus*, *Helicolenus dactylopterus*, *Lophius budegassa*, *Pagellus acarne*, *Scomber scombrus*, *Scophthalmus rhombus*, *Serranus cabrilla*, *Sparus aurata*, *Trachurus* sp., Triglidae (fish)	No	No
Pradillon et al. (2007)	Whole larvae in situ hybridization with 18S rRNA targeted probes	*Riftia pachyptila*, *Tevnia jerichonana* (hydrothermal vent polychaetes), *Crassostrea gigas* (Pacific oyster)	Yes	No
Le Goff-Vitry et al. (2007)	Whole larvae in situ hybridization with 18S rRNA targeted probes	*Ostrea edulis*, *Cerastoderma edule*, *Macoma balthica*, *Mytilus* sp., *Nucula* sp., *Glycymeris* sp., Pectinidae, Veneridae (bivalves)	Yes	No

The application of DNA barcoding for larval identification has faced technical obstacles, such as the applicability of a universal molecular marker (Shearer and Cofforth 2008), the restricted number of sequences available in databases (Barber and Boyce 2006, Webb et al. 2006, Richardson et al. 2007) and the lack of taxonomic resolution (Webb et al. 2006). Molecular approaches have addressed previously intractable larval ecological issues, yet, several major limitations still hamper their routine use in environmental studies, including quantification and

sensitivity problems associated with microarray platforms (Kochzius et al. 2008) and the influence of larval developmental on detection and optimisation of PCR with low quantities of target DNA (Patil et al. 2005). Morphological data therefore, remain the cornerstone of larval ecological studies (Minagawa et al. 2004, Richardson and Cowen 2004, Shanks and Brinks 2005). Future priorities should incorporate a more integrative approach, allying morphological, ecological and molecular data, in an attempt to understand the complexity and dynamics of marine life cycles (Will and Rubinoff 2004, Janzen et al. 2005, Dasmahapatra and Mallet 2006).

1.3 Marine Biodiversity and Ecosystem Function

1.3.1 Microbes in Novel Environments

Recent studies of microbial diversity have produced vast discoveries of previously unknown microorganisms, many of which have major impacts on oceanic processes. Deep-sea hydrothermal vents and cold seeps, for example, contain thriving chemosynthetic microbial communities; oceanic midwaters contain abundant Archaea. Large populations of picoplankton are the primary drivers for carbon fixation and for nitrogen recycling. Fundamental information about selection and evolution in the microbial world can be obtained if we can sample the global marine microbial genome. A full census of marine microorganisms and thus a complete list of metabolic processes is possible to achieve with intensive sampling (Scherer-Lorenzen 2005). Within all habitats, changing diversity has profound effects on biomass production, nutrient recycling, and ecosystem stability (Hughes et al. 2006). Higher biodiversity (genetic diversity) can afford a degree of ecological insurance against ecological uncertainty (Hughes et al. 2006).

1.3.2 Microbial Links in Ecosystem Processes

Knowing what "kinds" of organisms exist within populations and how the community structure changes in response to environmental changes is the only way to understand how biological systems force ocean function. Sophisticated measurements of microbial and metabolic diversity and how this diversity is linked to biogeochemical and physical processes is required to explore the dynamics of population biology, genome diversity and the metabolic basis of biogeochemical processes, especially at the microbial level where most of the unknown diversity lies. Scientists from all disciplines must interact to produce a predictive modelling framework to understand the interaction between members of complex microbial consortia and ocean biogeochemistry. Such predictions will challenge even the most advanced genetic technology and evolutionary theory because the evolution and diversity of the marine ecosystem is vastly different from terrestrial ones upon which

our current models are based. For example, the scale of disturbances in oceanic environments that can happen on a daily basis means that climax communities like those that can develop on land rarely develop in the ocean. DeLong and Karl (2005) are applying genomic technologies to search for the genomes of as-yet unidentified microbes in samples from the environment. Given the central role that microbes play in functioning ecosystems, ecogenomics has enormous implications for understanding of diversity in the oceanic microbial community (Eisen 2007), and can also shed significant insight into the health of ecosystems if the assemblage of microbes and their functions are eventually understood and mapped onto their geographic locations (Hoffman and Gaines 2008).

1.3.3 Environmental Change and Microbial Diversity

The world's oceans are facing many rapidly increasing threats from human disturbance. Ecosystems around the globe are currently undergoing dramatic changes in species composition because of the influence of human activity. The manner in which natural communities utilise resources affects the physical environment and interacts with other species suggests that biodiversity (that is, the composition, structure and function of an community) is essential for the functioning and/or sustainability of an ecosystem. These changes have usually led to a reduction in species diversity. Changes in species composition, species richness, and/or functional type affect the efficiency with which resources are utilised within an ecosystem, and suggest that biogeochemical functioning of an ecosystem will be impaired by biodiversity loss. Much remains unknown about how species richness or functional groups affects ecosytem level responses. Experiments to test the relationship between species richness and ecosystem function have been largely confined to terrestrial systems, but more recently marine systems have also been utilised, though with some conflicts between simulations and real observations. In recent years, many models have attempted to describe how changes in species richness may affect ecosystem function. A specific ecosystem function is thus seen as a function of (i) biodiversity and the functional traits of the organisms involved, (ii) associated biogeochemical processes, and (iii) the abiotic environment. The concept of resilience, as applied to an ecosystem, is loosely defined as the ability of the system to maintain its function when faced with novel disturbance. The concept is related to stability, but with its focus on maintenance of function and novel disturbance, resilience uniquely encompasses aspects of society's reliance on ecosystem services and increasing anthropogenic change (Webb 2007). Thus, ecologists can study resilience from a complex adaptive systems (CAS) approach or a social–ecological systems (SES) approach, which places equal weight on the human and ecological dimensions of ecosystem function and maintenance (Webb 2007).

Although documented global extinctions are rare in the marine environment, local extinctions and dramatic changes in abundance are widespread. The causes of this loss and its consequences for the functioning and stability of ecosystems

are the current focus of intense research activity, partly because of the threat to the goods and services that ecosystems provide to society. Much of the research to date has been controversial, with disagreement over the role of diversity as opposed to the roles of individual species or functional groups. Marine environments are more diverse at higher taxonomic levels than terrestrial systems and have higher levels of functional diversity. There are several hypotheses proposed, ranging from diversity having no effect on ecosystem function, through diversity driving ecosystem functioning and must be redundant to cope with the magnitude of changes, and the direction of change with diversity loss may or may not be predicatable. The question of how biodiversity affects ecosystem functioning is to search for communities differing in one aspect of biodiversity.

Few data are available to assess the potential ecosystem level importance of genetic diversity within species known to play a major functional role. However, Hughes and Stachowicz (2004) have shown that increasing genotypic diversity in a habitat-forming seagrass, *Zostera marina,* enhanced the community resistance to disturbance by grazing geese. In addition, the time required for the community to recovery to near predisturbance densities also decreased with increasing eelgrass genotypic diversity. A study on the population structure of the microalga *Phaeocystis antarctica* from all of the major continental gyres around Antarctica using microsatellites has shown that each continental gyre is highly genetically diverse with reduced gene flow between the gyres (Gaebler and Medlin, unpubl.). However, isolates from each gyre are physiologically distinct. For example, in those gryes with annual ice cover, isolates can survive a range of salinities from 18 to 70‰ with only a delay in maximum growth over time, whereas isolates from ice-free gyres can only survive at full strength sea water (33‰). Thus there appears to be an integrated phenotypic response to the environment such that should climate change occur, local extinctions would be expected and repopulation of an area would depend on gene flow and dispersal from other others. The high diversity within each gyre suggests that *P. antarctica* has exploited the niche available in each gyre to cover different possible combinations of environmental conditions that occur in that gyre and to make a more stable population.

1.4 Concluding Remarks

Documented shifts in the distribution and abundance of organisms in response to climate change (Umina et al. 2005, Bradshaw and Holzapfel, 2006) indicate that environments are changing rapidly. The ability of biota to persist is determined by their genetic constitution and ability to adapt physiologically. In extreme changes, an organism may respond in four ways: it migrates to a more favourable area, a physiological response increases the metabolic range for survival and reproduction, selective mortality results in local adaptation, or populations become locally extinct. Crucially, genetic markers can assist in monitoring the impact of environmental change at the level of DNA sequences, proteins or metabolites, as well as

assessing the nature of selection imposed by shifts in conditions, and estimating the potential of populations to respond to natural and anthropogenic change (Hoffmann and Willi 2008). In addition to the well established use of neutral molecular markers (Carvalho et al. 2002; Avise 2004), there has been an increasing shift in the study of adaptive genes directly involved in phenotypic response, that is, within a genomics framework. The excitement of many new species discoveries (Venter et al. 2004, Santelli et al. 2008) is, however, tainted by continuing population crashes of many exploited species, increase in number of endangered marine species, and continued degradation of habitats and ecosystem services (Palumbi et al. 2008b). The utilisation of genomic technologies, in combination with appropriate genetically and ecosystem-based conservation strategies offers one potent approach to halting such impacts (Fig. 1.5).

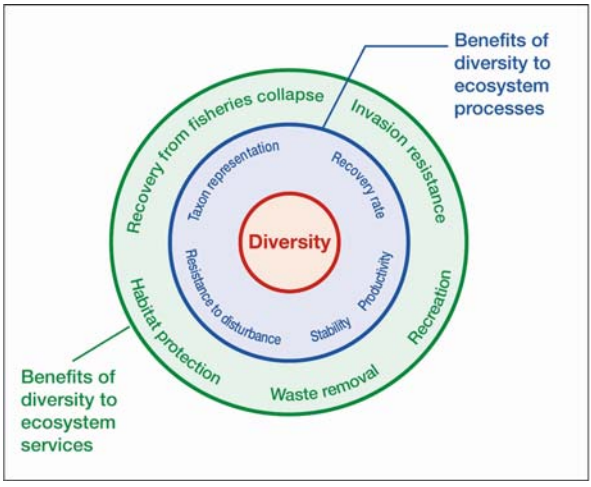

Fig. 1.5 Schematic representation of ecosystem benefits of marine biodiversity (after Palumbi et al. 2008b). Biodiversity (*red portion*) at the various biological levels (genetic, species, ecosystem and functional) enhances a variety of ecological processes (*blue portion*). Ecological processes enhance the benefits that ecosystems cab provide in terms of recovery, resistance, protection, recycling etc. (*green portion*)

The above inclusive consideration of marine biodiversity represents the range of hierarchical levels that description and monitoring can take place, as well as the close relationship among them. In many cases it is more practical to identify and monitor discrete species, such as when defining the management units for exploited species, whereas, a focus on genes and metabolic pathways may be more appropriate when tackling the effects of temperature on such processes as carbon assimilation in benthic biota. Ultimately, it is evident that the components of ecosystem stability – recovery, resistance and reversibility (Palumbi et al. 2008a), are indicators of overall resilience or robustness in the face of environmental change, and as such, necessitate the study of targets at different biological, spatial and temporal scales. As recently emphasized (Palumbi et al. 2008b), at least two key approaches can be identified

as priorities for managing marine biodiversity in the future: first, elucidating the link between diversity and ecosystem function, and second, undertaking empirical studies that demonstrate that increased biodiversity can enhance simultaneously and stabilize various measures of ecosystem functioning (Worm et al. 2006, Danovaro et al. 2008). Our belief is that genomics will play a pivotal role in both elements of such activity.

References

Adachi M, Kanno T, Okamoto R, Itakura S, Yamaguchi M, Nishijima T (2003) Population structure of Alexandrium (Dinophyceae) cyst formation-promoting bacteria in Hiroshima Bay, Japan. Appl Environ Microbiol 69:6560–6568

Allen EE, Banfield JF (2005) Community genomics and microbial ecology and evolution. Nat Rev Genet 3:489–498

Allen MJ, Martinez-Martinez J et al (2007) Use of microarrays to assess viral diversity: from genotype to phenotype. Environ Microbiol Apr 9(4):971–982

Allen MJ, Wilson WH (2008) Aquatic virus diversity accessed through omic techniques: a route map to function. Curr Opin Microbiol 11(3):226–232

Amann RI, Binder, BJ, Olson, RJ et al (1990) Combination of 16S rRNA-targeted oligonucleotide probes with flow cytometry for analysing mixed microbial populations. Appl Environ Microbiol 56:1919–1925

Angel M (1992) Managing biodiversity in the oceans. In: Peterson M (ed) Diversity of oceanic life. CSIS, Washington

Angly FE, Felts B et al (2006) The marine viromes of four oceanic regions. PLoS Biol 4(11):e368

Arnold WS, Hitchcock GL, Frischer ME, Wanninkhof R, Sheng P (2005) Dispersal of an induced larval cohort in a coastal lagoon. Limnol Oceanogr 50:587–597

Arise JC (2004) Molecular Markers, Natural History, and Evolution 2nd ed. Sinauer Assoc Inc, Sunderland, Massachusetts

Babcock DA, Wawrik B et al (2007) Rapid screening of a large insert BAC library for specific 16S rRNA genes using TRFLP. J Microbiol Methods 71(2):156–161

Barber P, Boyce SL (2006) Estimating diversity of Indo-Pacific coral reef stomatopods through DNA barcoding of stomatopod larvae. Proc R Soc B 273:2053–2061

Barki Y, Douek J, Graur D, Gateno D, Rinkevich B (2000) Polymorphism in soft coral larvae revealed by amplified fragment-length polymorphism (AFLP) markers. Mar Biol 136:37–41

Beja O (2004) To BAC or not to BAC: marine ecogenomics. Curr Opin Biotechnol 15(3):187–190

Beja O, Aravind L et al (2000) Bacterial rhodopsin: evidence for a new type of phototrophy in the sea. Science 289(5486):1902–1906

Bellwood DR, Hughes TP, Folke C, Nystro M (2004) Confronting the coral reef crisis. Nature 429:827–833

Bench SR, Hanson TE et al (2007) Metagenomic characterization of Chesapeake Bay virioplankton. Appl Environ Microbiol 73(23):7629–7641

Bhadury P, Austen MC et al (2006) Development and evaluation of a DNA-barcoding approach for the rapid identification of nematodes. Mar Ecol Prog Ser 320:1–9

Biddle JF, Fitz-Gibbon S et al (2008) Metagenomic signatures of the Peru Margin subseafloor biosphere show a genetically distinct environment. Proc Natl Acad Sci USA 105(30): 10583–10588

Biegala IC, Not F, Vaulot D, Simon N (2003) Quantitative assessment of picoeukaryotes in the natural environment by using taxon-specific oligonucleotide probes in association with tyramide signal amplification, fluorescence-in-situ-hybridization and flow cytometry. Appl Environ Microbiol 69:5519–5529

Billard C (1994) Life cycles. In: Green JC, Leadbeater BSC (eds). *The Haptophyte Algae*, The Systematics Association Special, Vol. 51. Clarendon Press, Oxford, pp 167–186

Blaxter M, Floyd R (2003) Molecular taxonomics for biodiversity surveys: already a reality. Trends Ecol Evol 18(6):268–269

Blaxter M, Mann J et al (2005) Defining operational taxonomic units using DNA barcode data. Phil Trans R Soc London B 360. doi:10.1098/rstb.2005.1725

Blouin MS (2000) Neutrality tests on mtDNA: unusual results from nematodes. J Hered 91(2): 156–158

Blouin MS, Yowell CA, Courtney CH, Dame JB (1998) Substitution bias, rapid saturation, and the use of mtDNA for nematode systematics. Mol Biol Evol 15(12):1719–1727

Blow N (2008) Exploring unseen communities. Nature 453:687–690

Bouchet P (2005) The mangitide of marine biodiversity. In: Duarte CM (ed) The exploration of marine biodiversity. Fundavión BBVA, Paris

Bradshaw, WE and Holzapfel, M (2006) Evolutionary response to rapid climate change. Science 312:1477–1478

Breitbart M, Felts B et al (2004) Diversity and population structure of a near-shore marine-sediment viral community. Proc Biol Sci 271(1539):565–574

Bullard SG, Lindquist N, Hay ME (1999) Susceptibility of invertebrate larvae to predators: how common are post-capture larval defenses?. Mar Ecol Prog Ser 191:153–161

Bulling MT, White PCL, Rafaelli D et al (2006) Using model systems to address the biodiversity-ecosystem functioning process. Mar Ecol Prog Ser 311:295–309

Caron DA, Peele ER, Lim EL, Dennett MR (1999) Picoplankton and nanoplankton and their trophic coupling in surface waters of the Sargasso Sea south of Bermuda. Limnol Oceanogr 44:259–272

Carvalho GR, van Oosterhout C, Hauser, L et al (2002) Measuring genetic variation in wild populations: from molecular markers to adaptive traits, In: Behringer, J Hails, RS, Godfrey C (eds) Genes in the Environment. Blackwell Science, pp. 91–111

Chan RWK, Dixon PI, Peperrell JG et al (2003) Application of DNA-based techniques for the identification of whaler sharks (*Carcharhinus* spp.) caught in protective beach meshing and by recreational fisheries off the coast of New South Wales. Fish Bull 101:910–914

Chapin FS, Walker BH et al (1997) Biotic control over the functioning of ecosystems. Science 277(5325):500–504

Cock JM, Scornet D, Coelho S et al (2005) Marine Genomics and the exploration of marine biodiversity. Biodiversity Conservation in the Coastal Zone. "Exploring Marine Biodiversity: Scientific and Technological Challenges", Madrid, 29th November 2005.

Comtet T, Jollivet D, Khripounoff A et al (2000) Molecular and morphological identification of settlement-stage vent mussel larvae, Bathymodiolus azoricus (Bivalvia: Mytilidae), preserved in situ at active vent fields on the Mid-Atlantic Ridge. Limnol Oceanogr 45:1655–1661

Costa FO, Carvalho GR (2007) The barcode of life initiative: synopsis and prospective societal impacts of DNA barcoding of Fish. Genom Soc Pol 3:29–40

Costanza R, d'Arge R, de Groot R et al (1997) The value of the world's ecosystem services and natural capital. Nature 387:253–260

Creer S (2008) New technologies. In: Lambshead PJD, Packer M (eds) Manual for the molecular barcoding of deep-sea nematodes. International Seabed Authority, United Nations, Kingston

Culley AI, Lang AS et al (2006) Metagenomic analysis of coastal RNA virus communities. Science 312(5781):1795–1798

Cunningham C, Hikima J et al (2006) New resources for marine genomics: bacterial artificial chromosome libraries for the Eastern and Pacific oysters (*Crassostrea virginica* and *C. gigas*). Mar Biotechnol (NY) 8(5):521–533

Danovaro R, Gambi C et al (2008) Exponential decline of deep-sea ecosystem functioning linked to benthic biodiversity loss. Curr Biol 18:1–8

Danovaro R, Gambi C, Dell'Anno A et al (2008) Exponential decline of deep-sea ecosystem functioning linked to benthic biodiversity loss. Curr Biol 18:1–8

Dasmahapatra KK, Mallet J (2006) DNA barcodes: recent successes and future prospects. Heredity 97:254–255

Deagle BE, Bax N, Hewitt CL, Patil JG (2003) Development and evaluation of a PCR-based test for detection of Asterias (Echinodermata : Asteroidea) larvae in Australian plankton samples from ballast water. Mar Freshwat Res 54:709–719

DeLey P, DeLey IT et al (2005) An integrated approach to fast and informative morphological vouchering of nematodes for applications in molecular barcoding. Phil Trans R Soc London B 360:1945–1958

DeLong EF (2005) Microbial community genomics in the ocean. Nat Rev Microbiol 3:459–469

DeLong EF and Karl DM (2005) Genomic perspectives in microbial oceanography. Nature 437:336–342

de Tomaso AW, Weissman IL (2003) Construction and characterization of large-insert genomic libraries (BAC and fosmid) from the Ascidian Botryllus schlosseri and initial physical mapping of a histocompatibility locus. Mar Biotechnol (NY) 5(2):103–115

Díez B, Pedrós-Alío C, Massana R (2001) Study of genetic diversity of eukaryotic picoplankton in different oceanic regions by small-subunit rRNA gene cloning and sequencing. Appl. Environ Microbial 67:2932–2941

Dinsdale EA, Pantos O et al (2008) Microbial ecology of four coral atolls in the Northern Line Islands. PLoS ONE 3(2):e1584

Doi RH, Igarashi RT (1965) Conservation of ribosomal and messenger ribonucleic acid cistrons in Bacillus species. J Bacteriol 90:384–390

Dubnau D, Smith I, Morell P, Marmur J (1965) Gene conservation in Bacillus species. I. Conserved genetic and nucleic acid base sequence homologies. Proc Natl Acad Sci USA 54(2): 491–498

Duffy JE, Stachowicz JJ (2006) Why biodiversity is important to oceanography: potential roles of genetic, species, and trophic diversity in pelagic ecosystem processes. Mar Ecol Prog Ser 311:179–189

Eisen JA (2007) Environmental shotgun sequencing: Its potential and challenges for studying the hidden world of microbes. PLoS Biol 5:e82

Eller G, Töbe K, Medlin LK (2007) A set of hierarchical FISH probes for the Haptophyta and a division level probe for the Heterokonta. J Plank Res 29:629–640

Evans KM, Wortley AH, Mann DG (2007) An assessment of potential diatom "Barcode" genes (cox1, rbcL, 18S and ITS rDNA) and their effectiveness in determining relationships in sellaphora (Bacillariophyta). Protist 158:349–364

Floyd R, Abebe E et al (2002) Molecular barcodes for soil nematode identification. Mol Ecol 11(4):839–850

Fox CJ, Taylor MI et al (2008) Mapping the spawning grounds of Noth Sea cod (*Gadus morhua*) by direct and indirect means. Proc R Soc Lond 275:1543–1548

Fox CJ, Taylor CJ, Pereyra R, Willasana MI, Rico C (2005) TaqMan DNA technology confirms likely overestimation of cod (*Gadus morhua* L.) egg abundance in the Irish Sea: Implications for the assessment of the cod stock and mapping of spawning areas using egg-based methods. Mol Ecol 14:879–884

Fuchs BM, Spring S et al (2007) Characterization of a marine gammaproteobacterium capable of aerobic anoxygenic photosynthesis. Proc Natl Acad Sci U S A 104(8):2891–2896

Gescher C, Metfies K, Frickenhaus S, Knefelkamp B, Wiltshire K, Medlin LK (2008) Feasibility of assessing the community composition of prasinophytes at the helgoland roads sampling site with a DNA microarray. Appl Environ Microbiol 74:5305–5316

Giovannoni SJ, Britschgi TB, Moyer CL et al (1990) Genetic diversity in sargasso sea bacterloplankton. Nature (London) 345:60–63

Goffredi SK, Jones WJ, Scholin CA, Marin R, Vrijenhoek RC (2006) Molecular detection of marine invertebrate larvae. Mar Biotech 8:149–160

Gómez A, Wright PJ, Lunt DH et al (2007) Mating trials validate the use of DNA barcoding to reveal cryptic speciation of a marine bryozoan taxon. Proc R Soc B 274:199–207

Grassle FJ, Maciolek NJ (1992) Deep-sea species richness: regional and local diversity estimates from quantitative bottom samples. Am Nat 139:313–341

Groben R, John U, Eller G, Lange M, Medlin LK (2004) Using fluorescently-labelled rRNA probes for hierarchical estimation of phytoplankton diversity. Nova Hedwigia 79:313–320

Hajibabaei M, Singer GA, Hebert PD, Hickey DA (2007) DNA barcoding: how it complements taxonomy, molecular phylogenetics and population genetics. Trends Genet 23(4):167–172

Handelsmann J (2004) Metagenomics: application of genomics to uncultured microorganisms. Microbiol Mol Biol Rev 68:669–685

Hansen BW, Larsen JB (2005) Spatial distribution of velichoncha larvae (Bivalvia) identified by SSNM-PCR. J Shell Res 24:561–565

Harding JM (1999) Selective feeding behavior of larval naked gobies (Gobiosoma bosc) and blennies (Chasmodes bosquianus and Hypsoblennius hentzi): preferences for bivalve veligers. Mar Ecol Prog Ser 179:145–153

Hebert PDN, Cywinska A, Ball SL, de Waard JR (2003) Biological identifications through DNA barcodes. Proc R Soc Lond B 270:313–321

Hoef-Emden K, Melkonian M (2003) Revision of the genus Cryptomonas (Cryptophyceae): a combination of molecular phylogeny and morphology provides insights in a long-hidden dimorphism. Protist 154:371–409

Hoffmann AA, Willi Y (2008) Detecting genetic responses to environment change. Nature Rev Genet 9:421–432

Hofmann GE, Gaines SE (2008) New tools to meet new challenges: Emerging technologies for managing marine ecosystems for resilience. BioScience 58(1):43–52

Hosoi M, Hosoi-Tanabe S, Sawada H et al (2004) Sequence and polymerase chain reaction–restriction fragment length polymorphism analysis of the large subunit rRNA gene of bivalve: Simple and widely applicable technique for multiple species identification of bivalve larva. Fish Sci 70:629–637

Huber JA, Welch DBM et al (2007) Microbial population structures in the deep marine biosphere. Science 318:97–100

Hughes TP, Bellwood DR, Folke C, Steneck RS, Wilson J (2006) New paradigms for supporting the resilience of marine ecosystems. Trends Ecol Evol 20:380–386

Hughes AR, Stachowicz JJ (2004) Genetic diversity enhances the resistance of a seagrass ecosystem to disturbance. Proc Natl Acad Sci 101:8898–9002

Huse SM, Huber JA, Morrison HG, Sogin ML, Welch DM (2007) Accuracy and quality of massively parallel DNA pyrosequencing. Genome Biol 8(7):R143

Igelsia-Rodriguez D, Schofield OM, Batley PJ, Medlin LK, Hayes PK (2006) Extensive intraspecific genetic diversity in the marine coccolithophorid Emiliania huxleyi: the use of microsatellite analysis in marine phytoplankton populations studies. J Phycol 42:526–536

Ishizaka J, Harada K, Ishikawa K, Kiyosawa H, Furusawa H, Watanabe Y, Ishida H, Suzuki K, Handa N, Takahash M (1997) Size and taxonomic plankton community structure and carbon flow at the equator, 175°E during 1990–1994. Deep Sea Res II 44:1927–1949

Janzen DH, Hajibabaei M, Burns JM et al (2005) Wedding biodiversity inventory of a large and complex Lepidoptera fauna with DNA barcoding. Phil Trans R Soc B 360:1835–1845

John U, Groben R, Beszteri B, Medlin L (2004) Utility of Amplified Fragment Length Polymorphisms (AFLP) to analyse genetic structures within the Alexandrium tamarense species complex. Protist Jun 155(2):169–179

Jones WJ, Preston CM, Marin R, Scholin CA, Vrijenhoek RC (2008) A robotic molecular method for in situ detection of marine invertebrate larvae. Mol Ecol Res 8:540–550

Jonston AW, Li Y, Ogilvie L (2005) Metagenomic marine nitrogen fixation- feast or famine? Trends Microbiol 13:416–420

Karaiskou N, Triantafyllidis A, Alvarez P et al (2007) Horse mackerel egg identification using DNA methodology. Mar Ecol 28:429–434

Kemp PF, Aller JY (2004) Bacterial diversity in aquatic and other environments: what 16S rDNA libraries can tell us. FEMS Microbiol Ecol 47:161–177

Kinlan BP, Gaines SD, Lester SE (2005) Propagule dispersal and the scales of marine community process. Div Distrib 11:139–148

Kirby RR, Lindley JA (2005) Molecular analysis of continuous plankton recorder samples, an examination of echinoderm larvae in the North Sea. J Mar Biol Ass UK 85:451–459

Kochzius M, Nölte M, Weber H et al (2008) Microarrays for identifying fishes. Mar Biotechnol 10:207–217

Lambshead PJD, Boucher G (2003) Marine nematode deep-sea biodiversity – hyperdiverse or hype?. J Biogeography 30(4):475–485

Lawton JHD, Bignell E et al (1998) Biodiversity inventories, indicator taxa and effects of habitat modification in tropical forest. Nature 391:72–76

Le Goff-Vitry MC, Chipman AD, Comtet T (2007) In situ hybridization on whole larvae: a novel method for monitoring bivalve larvae. Mar Ecol Prog Ser 343:161–172

Lindley JA, Batten SD (2002) Long-term variability in the diversity of North Sea zooplankton. J Mar Biol Ass UK 82:31–40

Lleonart J, Taconet M, Lamboeuf M (2006) Integrating information on marine species identification for fishery purposes. Mar Ecol Prog Ser 316:231–238

López-García P, Rodríguez-Valera F, Pedrós-Alió C et al (2001) Unexpected diversity of small eukaryotes in deep sea Antarctic plankton. Nature 409:603–607

Maggs CA, Castilho R, Foltz D, Henzler C, Jolly MT, Kelly J, Olsen J, Perez KE, Stam W, Väinölä R, Viard F, Wares J (2008) Evaluating signatures of glacial refugia for North Atlantic marine organisms?. Ecology 89:S108–S122

Maidak BL, Cole JR, Lilburn TG, Parker CTJ, Saxman PR, Farris RJ, Garrity GM, Olson GJ, Schmidt TM, Tiedje JM (2001) The RDP-II (Ribosomal Database Project). Nucleic Acids Res 29:173–174

Mann DG (1999) The species concept in diatoms. Phycologia 38:437–495

Mariani S, Uriz MJ, Turon X (2003) Methodological bias in the estimations of important meroplanktonic components from near-shore bottoms. Mar Ecol Prog Ser 253:67–75

Markmann M, Tautz D (2005) Reverse taxonomy: an approach towards determining the diversity of meiobenthic organisms based on ribosomal RNA signature sequences. Phil Trans R Soc London B 360:1917–1924

Marko B, Lee SC, Rice AM et al (2004) Mislabelling of a depleted reef fish. Nature 430:309–310

Marta-Almeida M, Dubert J, Peliz A, Queiroga H (2006) Influence of vertical migration pattern on retention of crab larvae in a seasonal upwelling system. Mar Ecol Prog Ser 307:1–19

Massana R, Terrado R, Forn I et al (2006) Distribution and abundance of unclutured heterotrophic flagellates in the world oceans. Env Microbiol 8:1515–1522

Medlin LK (2007) If everything is everywhere, do they share a common gene pool?. Gene 405:180–183

Medlin LK, Metfies K, Mehl H, Wiltshire K, Valentin K (2006) Picoplankton Diversity at the Helgoland Time Series Site as assessed by three molecular methods. Micro Ecol 167: 1432–1451

Metaxas A, Burdett-Coutts V (2006) Response of invertebrate larvae to the presence of the ctenophore Bolinopsis infundibulum, a potential predator. J Exp Mar Biol Ecol 334:187–195

Metfies K, Hujic S, Lange M, Medlin L (2005) Electrochemical detection of the toxic dinoflagellate *A. ostenfeldii* with a DNA Biosensor. Biosen Bioselec 20:1349–1357

Metfies K, Töbe K, Scholin C, Medlin LK (2006) Laboratory and field applications of ribosomal RNA probes to aid the detection and monitoring of harmful algae. In: Granéli E, Turner J (eds) Ecology of harmful algae. Springer, New York, pp 311–326

Minagawa G, Miller MJ, Aoyama J, Wouthuyzen S, Tsukamoto K (2004) Contrasting assemblages of leptocephali in the western Pacific. Mar Ecol Prog Ser 271:245–259

Moens T, Vincx M (1997) Observations on the feeding ecology of estuarine nematodes. J Mar Biol Assoc UK 77(1):211–227

Morgan TS, Rogers AD (2001) Specificity and sensitivity of microsatellite markers for the identification of larvae. Mar Biol 139:967–973

Mou X, Sun S et al (2008) Bacterial carbon processing by generalist species in the coastal ocean. Nature 451:708–711

Muyzer G (1999) DGGE/TGGE a method for identifying genes from natural ecosystems. Curr Opin Microbiol 2:317–322

Naeem S (2006) Expanding scales in biodiversity-based research: challenges and solutions for marine systems. Mar Ecol Prog Ser 311:273–283

Noell CJ, Donnellan S, Foster R, Haigh L (2001) Molecular discrimination of garfish Hyporhamphus (Beloniformes) larvae in southern australian waters. Mar Biotechnol 3: 509–514

Ogden R (2008) Fisheries Forensics: the use of DNA tools for improving compliance, traceability and enforcement in the fishing industry. Fish Fisheries 9:462–472

Pace B, Campbell LL (1971a) Homology of ribosomal ribonucleic acid of Desulfovibrio species with Desulfovibrio vulgaris. J Bacteriol 106(3):717–719

Pace B, Campbell LL (1971b) Homology of ribosomal ribonucleic acid diverse bacterial species with Escherichia coli and Bacillus stearothermophilus. J Bacteriol 107(2):543–547

Palumbi SR, McLeod KL, Grunbaum D (2008a) Ecosystems in action: lessons from marine ecology about recovery, resistance and reversibility. Bioscience 58:33–42

Palumbi SR, Sandifer PA, Allan JD (2008b) Managing for ocean biodiversity to sustain marine ecosystem services. Front Ecol Environ. doi: 10.1890/070135

Patil JG, Gunasekera RM, Deagle BE, Bax NJ (2005) Specific detection of pacific oyster (Crassostrea gigas) larvae in plankton samples using nested polymerase chain reaction. Mar Biotechnol 7:11–20

Peers G, Price NM (2006) Copper-containing plastocyanin used for electron transport by an oceanic diatom. Nature 44:341–344

Pegg GC, Snclair B, Briskey L et al (2006) MtDNA barcode identification of fish larvae in the southern Great Barrier Reef, Australia. Sci Mar 70:7–12

Phillips RA, Petersen MK, Lilliendahl K et al (1999) Diet of the northern fulmar *Fulmarus glacialis*: reliance on commercial fisheries?. Mar Biol 135:159–170

Pikitch EK, Santora C, Babcock EA, Bakun A, Bonfil R, Conover DO, Dayton P, Doukakis P, Fluharty D, Heneman B, Houde ED, Link J, Livingston PA, Mangel M, McAllister MK, Pope J, Sainsbury KJ (2004) Ecosystem-Based Fisheries Management. Science 305:346–347

Platt HM, Warwick RM (1983) Free-living marine nematodes. Part I British Enoplids. Cambridge University Press, Cambridge

Pradillon F, Schmidt A, Peplies J, Dubilier N (2007) Species identification of marine invertebrate early stages by whole-larvae in situ hybridisation of 18S ribosomal RNA. Mar Ecol Prog Ser 333:103–116

Pruesse E, Quast C, Knittel K, Fuchs BM, Ludwig W, Peplies J, Glöckner FO (2007) SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. Nucleic Acids Res 35(21):7188–7196

Quaiser A, Lopez-Garcia P et al (2008) Comparative analysis of genome fragments of Acidobacteria from deep Mediterranean plankton. Environ Microbiol 10:2704–2717

Ratnasingham S, Hebert PDN (2007) BOLD: The barcode of life data system (http://www.barcodinglife.org). Mol Ecol Notes 7:355–364

Ray GC (1991) Coastal-zone biodiversity patterns. BioScience 41:490–498

Richardson DE, Cowen RK (2004) Diversity of leptocephalus larvae around the island of Barbados (West Indies): relevance to regional distributions. Mar Ecol Prog Ser 282:271–284

Richardson DE, Vanwye JD, Exum AM, Cowen RK, Crawford DL (2007) High-throughput species identification: from DNA isolation to bioinformatics. Mol Ecol Notes 7:199–207

Rock J, Costa FO, Walker DI et al (2008) DNA barcodes for fish of the Antarctic Scotia Sea indicate priority groups for taxonomic and systematic focus. Antarctic Sci 20:253–262

Rogers AD (2001) Molecular ecology and identification of marine invertebrate larvae. In: Atkinson D, Thorndyke M (eds) Environment and animal development: genes, life histories and plasticity. BIOS Scientific Publishers Ltd., Oxford, pp 29–69

Rosel PE, Kocher TD (2002) DNA-based identification of larval cod in stomach contents of predatory fishes. J Exp Mar Biol Ecol 267:75–88

Rothberg JM, Leamon JH (2008) The development and impact of 454 sequencing. Nat Biotech 26:1117–1124

Santaclara FJ, Espineira M, Vieites JM (2007) Molecular detection of Xenostrobus securis and Mytillus galloprovincialis larvae in galician coast (Spain). Mar Biotechnol 9:722–732

Santelli CM, Orcutt BN, Banning E et al (2008) Abundance and diversity of microbial life in ocean crust. Nature 453:653–656

Savolainen R, Cowan RS, Vogler AP (2005) DNA barcoding of life. Phil Trans R Soc Lond B 360:1805–1980

Scherer-Lorenzen M (2005) Biodiversity and ecosystem functioning: basic principles. In: Wilhelm Barthlott, K Eduard L, Stefan P (eds) Biodiversity: structure and function. Encyclopedia of Life Support Systems (EOLSS), Developed under the Auspices of the UNESCO, Eolss Publishers, Oxford. [http://www.eolss.net]

Schratzberger M, Warr K et al (2007) Functional diversity of nematode communities in the southwestern North Sea. Mar Environ Res 63(4):368–389

Shanks AL, Brink L (2005) Upwelling, downwelling, and cross-shelf transport of bivalve larvae: test of a hypothesis. Mar Ecol Prog Ser 302:1–12

Shearer TL, Coffroth MA (2006) Genetic identification of Caribbean scleractinian coral recruits at the flower garden banks and the florida keys. Mar Ecol Prog Ser 306:133–142

Shearer TL, Coffroth MA (2008) Barcoding corals: limited by interspecific divergence, not intraspecific variation. Mol Ecol Res 8:247–255

Sigler MF, Hulbert LB, Lunsford CR et al (2006) Diet of Pacific sleeper shark, a potential Steller sea lion predator, in the north-east Pacific Ocean. J Fish Biol 69:392–405

Simon N, Campbell L, Ornolfsdottir E et al (2000) Oligonucleotide probes for the identification of three algal groups by dot biot and fluorescent whole-cell hybridization. J EuR Microbial 47:76–84

Smith CR, Austen MC et al (2000) Global change and biodiversity linkages across the sediment-water interface. Bioscience 50(12):1108–1120

Snelgrove PVR, Austen MC et al (2000) Linking biodiversity above and below the marine sediment-water interface. BioScience 50(12):1076–1088

Snelgrove P, Blackburn TH et al (1997) The importance of marine sediment biodiversity in ecosystem processes. Ambio 26:578–583

Sogin ML, Morrison HG et al (2006) Microbial diversity in the deep sea and the underexplored "rare biosphere". Proc Natl Acad Sci USA 103:12115–12120

Stein JL, Marsh TL et al (1996) Characterization of uncultivated prokaryotes: isolation and analysis of a 40-kilobase-pair genome fragment from a planktonic marine archaeon. J Bacteriol 178(3):591–599

Suzuki MT, Preston CM et al (2004) Phylogenetic screening of ribosomal RNA gene-containing clones in Bacterial Artificial Chromosome (BAC) libraries from different depths in Monterey Bay. Microb Ecol 48(4):473–488

Taylor MI, Fox C, Rico I, Rico C (2002) Species-specific TaqMan probes for simultaneous identification of (Gadus morhua L.), haddock (Melanogrammus aeglefinus L.) and whiting (Merlangius merlangus L. Mol Ecol Notes 2:599–601

Tzeneva VA, Heilig HG, van Vliet WA, Akkermans AD, de Vos WM, Smidt H (2008) 16S rRNA targeted DGGE fingerprinting of microbial communities. Methods Mol Biol 410:335–349

Umina PA, Weeks AR, Kearney MR et al (2005) A rapid shift in classic clinical pattern in *Drosophila* reflecting climate change. Science 308:691–693

Underwood AJ, Fairweather PG (1989) Supply-side ecology and marine benthic assemblages. Trends Ecol Evol 4:16–19

Vadopalas B, Bouma JV, Jackels CR, Friedman CS (2006) Application of real-time PCR for simultaneous identification and quantification of larval abalone. J Exp Mar Biol Ecol 334:219–228

Van Straalen NM, Roelofs D (2006) An introduction to ecological genomics. Oxford University Press, Oxford
Venter JC, Remington K et al (2004) Environmental genome sequencing of the Sargasso Sea. Science 304:66–74
Vera JC, Wheat CW, Fescemyer HW et al (2008) Rapid transcriptome characterisation for a non-model organism using 454 pyrosequencing. Mol Ecol 17:1636–1647
von der Heyden S, Lipnski MR, Matthee CA (2007) Species-specific genetic markers for identification of early life-history stages of Cape hakes, Merluccius capensis and Merluccius paradoxus in the southern Benguela. Curr J Fish Biol 70:262–268
Webb CT (2007) What is the role of ecology in understanding ecosystem resilience?. BioScience 5:470–471
Webb KE, Barnes DKA, Clark MS, Bowden DA (2006) DNA barcoding: A molecular tool to identify Antarctic marine larvae. Deep-Sea Res II 53:1053–1060
Will KW, Rubinoff D (2004) Myth of the molecule: DNA barcodes for species cannot replace morphology for identification and classification. Cladistics 20:47–55
Williamson SJ, Rusch DB et al (2008) The Sorcerer II global ocean sampling expedition: metagenomic characterization of viruses within aquatic microbial samples. PLoS ONE 3(1):e1456
Wilson K, Thorndyke M, Nilsen F et al (2005) Marine systems: moving into the genomics era. Mar Ecol 26:3–16
Witham TG, DiFazio SP, Schweitzer JA et al (2008) Extending genomics to natural communities and ecosystems. Science 320:492–495
Woese CR (1987) Bacterial evolution. Microbiol Rev 51:221–271
Wong EH-K, Hanner RH (2008) DNA barcoding detects market substitution in North American seafood. Food Res Int 41:828–837
Worm B, Barbier EB, Baumont N et al (2006) Impacts of biodiversity loss on ocean ecosystem services. Science 314:787–790
Zhan A, Bao Z, Hu X et al (2008) Accurate methods of DNA extraction and PCR-based genotyping for single scallop embryos/larvae long preserved in ethanol. Mol Ecol Res 8:790–795

Chapter 2
Metagenome Analysis

Anke Meyerdierks and Frank Oliver Glöckner

Abstract The term "metagenomics" represents a combination of molecular and bioinformatic tools used to assess the genetic information of a community without prior cultivation of the individual species. It is valuable for the study of microorganisms of which only a minor fraction is yet culturable. The collective genomes present in an environmental sample or in an enrichment of target cells are extracted and subject to sequence-based or functional analyses. The field of metagenomics is evolving very rapidly, especially due to newly developed high-throughput sequencing technologies and increased computational power. Metatranscriptome and – proteome analyses are increasingly combined with metagenomic studies in order to assess not only the genetic potential of a microbial community, but also the genes expressed in a particular environment. The present chapter gives a short historical overview of the early years of metagenome analyses, and of possible applications. Challenges regarding the molecular and bioinformatic part of metagenome analyses are discussed. The molecular section includes strategies to access a metagenome. Methods to enrich for cells or the DNA of certain subpopulations prior to metagenome analysis, as well as to extract, purify and amplify DNA are given. The construction and sequence-based screening of small and large insert metagenomic libraries as well as a library-independent metagenomic approach using a new high throughput sequencing technology are described. The bioinformatic section provides an overview of assembly and binning tools, gene prediction programs, and annotation systems. This section also addresses the problem of metagenomic studies on habitats with high microbial diversity. Moreover, approaches to analyse phylogenetic and functional diversity within a dataset are discussed. The aim of the chapter is to provide the reader with basic information on both the molecular and bioinformatic aspects of metagenome analysis, to give hints to further reading and, therefore, to enable the reader to use this valuable method in an appropriate way in his or her studies.

A. Meyerdierks (✉)
Max Planck Institute for Marine Microbiology, Celsiusstrasse 1, 28359 Bremen, Germany
e-mail: ameyerdi@mpi-bremen.de

J.M. Cock et al. (eds.), *Introduction to Marine Genomics*,
Advances in Marine Genomics 1, DOI 10.1007/978-90-481-8639-6_2,

2.1 Introduction

Microorganisms are the most abundant form of life on Earth, and catalyse key processes such as nitrogen fixation and the mineralization of organic matter. Considering that about 70% of our Earth's surface is covered by the oceans, a comprehensive understanding of microbial element cycling in marine environments is crucial if we are to understand these processes on a global scale. In the light of the current discussions about global warming, the formation, storage, emission, as well as degradation of greenhouse gases in marine environments is receiving a lot of attention. Although the importance of an investigation of biogeochemical cycles in the ocean has generally been acknowledged, the study of microbial populations in marine environments, their contribution to biogeochemical cycles, and the ecophysiology of individual species is still in its infancy. Metagenomics, defined as the cultivation independent approach to assess the genetic potential of organisms, has opened a new dimension in environmental research. This chapter is intended to provide an overview of the recent developments in metagenomics, ranging from lab technology to bioinformatics.

Antoni van Leeuwenhoek provided the first microscopic evidence for the existence of microorganisms in the late seventeenth century (for review: Hall 1989). It then took almost two centuries before microbiologists such as Louis Pasteur (for review: Schwartz 2001), Robert Koch (for review: Kaufmann and Schaible 2005), and Martinus Beijerinck (for review: Chung and Ferris 1996) started to describe microorganisms based on culturing and enrichment techniques. Many of these early studies in the nineteenth and early twentieth century targeted microorganisms of medical relevance, but pioneers such as Ferdinand Cohn already studied algae and photosynthetic bacteria. Cohn also described genera such as the large sulfur bacterium *Beggiatoa* (for review: Drews 2000). It was Sergei Winogradsky, who developed the concept of chemolithotrophy, revealing the essential role of microorganisms in biogeochemical processes. Moreover, he first isolated and described nitrogen-fixing as well as nitrifying bacteria (for review: Schlegel 1996). These first microbiological studies were all restricted to the isolation of microorganisms and their characterization in the laboratory. This changed when molecular techniques were introduced in the field of microbiology, and the 16S rRNA gene was discovered to be a phylogenetic marker that could be used to describe the diversity of uncultured microorganisms (Olsen et al. 1994). Early cultivation-independent investigations reported an immense array of completely unexpected microbial diversity in the environment (Torsvik et al. 1990). Moreover, the design and application of specific rRNA-targeted oligonucleotide probes allowed insights into the composition of microbial communities in situ (Stahl and Amann 1991, Amann and Fuchs 2008).

It is estimated that only about 1% of the microbial diversity in the biosphere has been revealed so far by means of standard cultivation techniques (Amann et al. 1995, Curtis et al. 2002). New cultivation strategies have already been introduced to gain access to this yet uncultured majority of microorganisms (Connon and Giovannoni 2002, Rappe et al. 2002, Zengler et al. 2002). However, currently cultivation is not

keeping pace with the flood of molecular studies of microbial communities which are identifying an increasing number of species and phyla. Examples of these molecular tools include powerful PCR-based methods that have been established for the direct amplification, cloning, and analysis of ribosomal RNA (rRNA) genes from the environment (Pace et al. 1985, Olsen et al. 1986, Giovannoni et al. 1990, Ward et al. 1990). Recently, with the development of a new generation of DNA amplification and sequencing techniques, a new dimension in 16S rRNA diversity analysis was opened, allowing massively parallel sequencing of a variable region of the 16S rRNA gene from environmental samples (Sogin et al. 2006).

It is often difficult to predict the ecophysiology of uncultivated microorganisms solely based on rRNA phylogeny. For that, also the "adaptive" pool of metabolic, resistance, and defence genes has to be investigated, because they ensure survival in the environment. The addition of metagenomic techniques to the toolbox available to molecular ecologists has opened a new window to study the metabolic equipment of uncultured microorganisms in detail, and has allowed bridging between diversity and function.

2.2 History and Application of Metagenomics

The term "metagenome analysis" was introduced by Jo Handelsman (1998). The term is derived from the statistical concept of meta-analyses, i.e. the process of statistically combining separate analyses, and genomics, the comprehensive analysis of an organism's genetic material. The method involves sequence-based or functional analysis of the collective genomes contained in an environmental sample based on genomic DNA fragments retrieved from a habitat, or an enrichment of target cells. Popular synonyms of "metagenome analysis" are "environmental genomics", "ecogenomics", and "community genomics".

Metagenome analyses were actually carried out before the term metagenomics was coined. Thomas M. Schmidt, Edward F. DeLong and Norman R. Pace were the first to screen a metagenomic library, with the objective of obtaining a PCR-independent, unbiased access to the microbial diversity in a marine ecosystem. They extracted DNA from about 8,000 l of an oligotrophic picoplankton sample from the north central Pacific Ocean and cloned it into a bacteriophage λ derived vector. Part (3.2×10^4 clones) of the resulting library (10^7 clones), with insert sizes of 10–20 kbp, was screened for 16S rRNA genes by hybridization with a mixed kingdom probe. This resulted in the identification of 16 unique clones. This group had already noted the power of metagenomics for the retrieval of sequence information associated with the 16S rRNA gene (Schmidt et al. 1991). The next step was made five years later by DeLong. With colleagues, he sequenced a metagenomic fosmid originating from an uncultured planktonic archaeon (Stein et al. 1996). With improved sequencing techniques and bioinformatic tools, metagenomics became increasingly popular in the late 1990s. Large metagenomic fragments were preferably cloned into high capacity vectors to create cosmids (Collins and Hohn 1978), fosmids (Kim

et al. 1992), and bacterial artificial chromosomes (BACs) (Shizuya et al. 1992) (see Section 2.3.5.2). The large inserts (> 30 kbp) in these clones offered the chance to gain access to whole operons, and to find a phylogenetic marker on the same contiguous region (contig) as a gene of interest, indicating the phylogenetic affiliation of the microorganism from which the genomic fragment originated. The discovery of proteorhodopsin based on large insert metagenomic libraries, by co-localisation of a rhodopsin-like gene and a 16S rRNA gene affiliating with the γ-proteobacterial SAR86 cluster on a 130 kb BAC clone, remains one of the most intriguing outcomes of these metagenomic studies (Beja et al. 2000, 2001). This and subsequent studies changed substantially our understanding of aerobic, anoxygenic phototrophy in the oceans and its relevance to global carbon and energy budgets (Bryant and Frigaard 2006). More recently, it has even been demonstrated that complete bacterial genomes can be reconstructed, based on metagenomic fosmid clones. This approach was used for a rice cluster I bacterium. For this project, about 3,700 fosmid clones were constructed from an enrichment culture and used to assemble a complete genome (Erkel et al. 2006). Moreover, massive fosmid insert-end sequencing has been applied to analyse depth-variable community trends in carbon and energy metabolism of planktonic communities (DeLong et al. 2006).

With decreasing costs and improved high-throughput sequencing techniques, metagenome analysis based on small insert libraries (about 1.5–3 kbp) became "en vogue". These approaches were particularly successful in low diversity habitats, for example in a study of the microbial community in an acid-mine drainage (Tyson et al. 2004). In a second study, of the symbionts of the marine worm *Olavius algarvensis*, small insert shotgun library sequencing was complemented with fosmid sequencing (Woyke et al. 2006). Both of these studies provided fascinating insights into the genetic capabilities of the respective microbial communities and the potential interactions of microbes with each other, and with their host. The largest massive small insert metagenomic library sequencing experiment carried out to date was recently completed by the Global Ocean Sampling expedition (Rusch et al. 2007, Yooseph et al. 2007). This study was a follow up to the Sargasso Sea shotgun sequencing project (Venter et al. 2004), which had already led to the identification of 148 new phylotypes, and 69,901 novel genes, including 782 new proteorhodopsin genes.

Recently, a new DNA sequencing method, called pyrosequencing, has been developed. The technique does not involve cloning and is therefore free of cloning biases (Margulies et al. 2005). Third generation pyrosequencing technology generates sequences of just over 400 bp. This technique in its infancy with read length of about 100 bp already proved to be valuable for the study of microbial diversity (Leininger et al. 2006, Sogin et al. 2006) and metagenome (Edwards et al. 2006) analyses.

In addition to inventions and improvements in the field of DNA sequencing, isothermal multiple displacement amplification (MDA) using bacteriophage ϕ29-polymerase has generated considerable interest from molecular ecologists. The technique allows the amplification of genomic DNA from minute amounts of environmental samples to obtain insights into microbial communities such as those

found in soil (Abulencia et al. 2006). MDA is also used to amplify DNA from single cells. Although this technique still has its drawbacks (Hutchison and Venter 2006) it is valuable as a method for obtaining an initial overview of the genome of a single uncultured bacterium (Marcy et al. 2007). Such single cell approaches aim at the isolation of single uncultured microorganisms from the environment, and at the subsequent description of the genetic potential of each individual organism. This approach, therefore, involves analysing the metagenome cell by cell. Single cell genomics and metagenomics are therefore complementary approaches to assess the genomes of uncultured organisms. Some of the techniques discussed below are applied in both, metagenomics and single cell genomics.

2.3 Technical Challenges in Metagenome Analysis

2.3.1 Strategies to Assess the Metagenome

The construction and analysis of metagenomic libraries follows a general scheme (Fig. 2.1). Genomic DNA is either isolated directly from an environmental sample, or following enrichment. The DNA generally requires further purification in order to remove contaminants such as polyphenolic substances or metal ions, which could interfere with the subsequent enzymatic manipulation of the DNA. If insufficient DNA is retrieved for downstream processing, multiple displacement amplification (MDA) can be applied to increase DNA quantity (Binga et al. 2008). In the next steps, the genomic DNA is processed for cloning into large (BAC, fosmid, Cosmid) or small (plasmid) insert vectors or for direct DNA sequencing without prior library construction (e.g. pyrosequencing). If libraries are constructed, arraying of individual clones is often recommended, and an initial characterisation of the library with respect to parameters such as the number of recombinant clones, and the average insert-size should be carried out. The libraries can be used directly for insert-end sequencing or they can be screened for selected DNA sequences or expressed proteins. If large insert clones carry a gene of interest, they can be completely sequenced. This sequencing step is then followed by bioinformatic analysis of the dataset. This includes, depending on the approach, a binning of sequences belonging to the same species or group, the assembly of contigs, the functional assignment of predicted genes, and the interpretation of specific metabolic capabilities.

Ecologists have various microbiological and molecular tools available to them. The choice of tool depends on the question to be answered by the metagenomic study. Issues that need to be considered are the diversity of the microbial community in the habitat of interest. One also needs to decide whether the study is targeted towards the genetic capability of the entire community or of one population within the community. Depending on this, a decision for or against an enrichment of cells or DNA has to be made. In addition, the quality, length and amount of sequences that will be required to address the aim of the study, has to be defined.

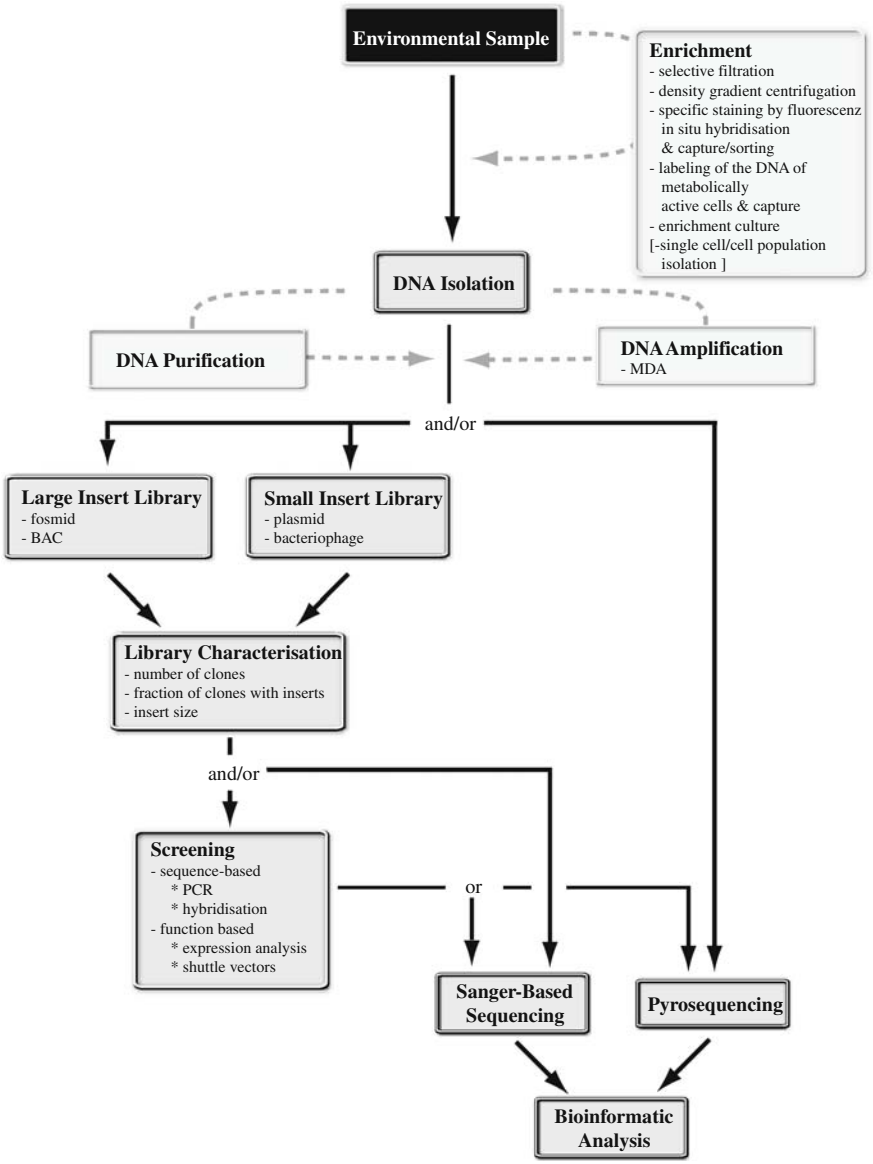

Fig. 2.1 Metagenomic workflow. Steps indicated by dashed lines are optional

The genetic composition of a highly diverse microbial community can be assessed using several different approaches. A simple inventory of genetic capabilities without gene order and phylogenetic affiliation of single genes is best obtained by shotgun sequencing of small insert libraries (Venter et al. 2004, Tringe et al.

2005, Rusch et al. 2007) or library-independent pyrosequencing (Angly et al. 2006, Edwards et al. 2006, Biddle et al. 2008, Dinsdale et al. 2008). If, however, gene order and phylogenetic affiliation of gene fragments are of interest, end-sequencing or screening of large insert libraries is the method of choice (DeLong et al. 2006). This approach allows selection and further sequencing of clones that carry genes of interest.

Studies aimed at studying the genetic capability of certain species within a highly diverse microbial community should use an enrichment of target cells to reduce screening and sequencing effort. If this is successful, a shotgun sequencing approach or a pyrosequencing approach followed by sequence assembly can be applied. If it is only possible to obtain a small number of target cells at a high level of purity, a DNA amplification step has to be included (Mussmann et al. 2007, Podar et al. 2007). If pre-enrichment of target populations is not possible, the generation of large insert libraries can be attempted as this allows a considerably reliable phylogenetic affiliation of sequences (Beja et al. 2000, Krüger et al. 2003, Teeling et al. 2004, Bryant et al. 2007).

An intelligent combination of the available tools can result in significant savings in both time and money.

2.3.2 Enrichment Strategies

There are various methods available to enrich subpopulations of microbial communities prior to a metagenomic study.

Plankton samples have been fractionated by serial filtration through filters of different pore sizes. Eukaryotic phytoplankton and particles were removed using meshes and filters with pore sizes ranging from 20 to 0.8 μm (Schmidt et al. 1991, Stein et al. 1996, Beja et al. 2000, de la Torre et al. 2003, Lopez-Garcia et al. 2004, Venter et al. 2004, Angly et al. 2006, Culley et al. 2006, DeLong et al. 2006, Martín-Cuadrado et al. 2007, Rusch et al. 2007). The filtrate was either further subdivided (Rusch et al. 2007) or prokaryotic cells were concentrated using filters of generally 0.22 μm pore size, or by centrifugation (Schmidt et al. 1991, Beja et al. 2000). For metagenomic studies of viruses, prokaryotic cells were removed using filters of 0.22 μm pore size (Angly et al. 2006, Culley et al. 2006, Rusch et al. 2007). Prokaryotes inhabiting marine invertebrates have been successfully separated from host tissue cells by density gradient centrifugation (Hughes et al. 1997, Schirmer et al. 2005, Woyke et al. 2006, Robidart et al. 2008). Benthic microbial consortia were detached from sediment particles by sonication, separated from those particles by density gradient centrifugation, and concentrated and separated from single cells by filtration (Schleper et al. 1998, Hallam et al. 2003). In other cases, it may be even possible to manually isolate cell filaments. This was applied in a metagenome analysis of multicellular *Beggiatoa* filaments (Mussmann et al. 2007). Enrichment of genomic DNA based on G+C content is an alternative to the enrichment of intact cells, if the G+C content of the target cells is significantly different from that of the

bulk DNA. Selective lysis of cells, based for example on a difference in cell wall composition, or biochemical properties may also be used, when an enrichment of a subpopulation of a community is necessary.

Cell populations may also be tagged by in situ hybridisation using 16S rRNA-targeted nucleic acid probes. These probes can be labelled, for example using biotin, and target cells captured with streptavidin coated paramagnetic beads and a magnet (Stoffels et al. 1999). Another enrichment method using microplates has also been described (Zwirglmaier et al. 2004). An improved protocol combining catalysed-reporter deposition fluorescence in situ hybridisation (CARD-FISH) with magnetic bead capture, called Magneto-FISH, has recently been used to enrich methanotrophic microbial consortia from sediments (Pernthaler et al. 2008). Cell populations stained by fluorescence in situ hybridisation (FISH), have also been sorted by flow cytometry (Podar et al. 2007).

In addition to these cell enrichment methods based on physical or molecular characteristics, cell populations can also be separated based on their metabolic capabilities. The genomic DNA of the active fraction of a microbial community can be labelled by a short incubation with 5-bromo-2′-deoxyuridine (BrdU). BrdU is a thymidine analogue, which is incorporated into newly synthesised DNA of replicating cells. Such labelled cells can be visualised by immunostaining and sorted by flow cytometry. Alternatively, the labelled DNA can be captured with BrdU-specific monoclonal antibodies coupled to paramagnetic beads (Urbach et al. 1999, Mou et al. 2008). This method can also be used to label cells growing in response to a specific stimulus (Borneman 1999). Alternatively, microbial communities can be incubated with substrates labelled with stable isotopes (Dumont and Murrell 2005, Dumont et al. 2006, Neufeld et al. 2008). In one study, marine surface water was incubated with ^{13}C-labelled methanol. Microorganisms capable of metabolizing methanol incorporated ^{13}C. The resulting heavy DNA could be separated from light DNA by caesium chloride density gradient centrifugation, and was used for metagenome analysis of the genetic capability of the methanol utilizing fraction of the microbial community (Neufeld et al. 2008).

Another way of increasing the fraction of target cells is to enrich using classical microbiological techniques. The circular genome of the methanogenic, rice cluster I bacterium (Erkel et al. 2006) and the almost complete genome of the uncultured anammox bacterium *Kuenenia stuttgartiensis* (Strous et al. 2006) were obtained from metagenomic approaches based on such enrichments.

A general trend in the analysis of uncultured prokaryotes is single cell genomics. In one study, for example, single cells of marine bacterioplankton were sorted on 96-well plates by high speed fluorescence activated cell sorting (FACS) (Stepanauskas and Sieracki 2007). Moreover, microfluidic devices have been developed to allow single cell isolation based on dilution (Ottesen et al. 2006), or targeted selection of individual cells (Marcy et al. 2007). Micromanipulation has been applied to isolate FISH-stained cells (Ishoy et al. 2006, Kvist et al. 2007). Micromanipulation using optical tweezers (Huber et al. 1995), and laser capture microdissection (Gloess et al. 2008, Thornhill et al. 2008) are alternative tools to select single cells from environmental samples.

2.3.3 Isolation and Purification of Genomic DNA

The method applied for the isolation of metagenomic DNA depends largely on the downstream metagenomic analysis. Metagenomic DNA for library independent analysis as well as small insert and large insert shotgun libraries can be retrieved by "liquid phase-based" methods. This means, that the environmental sample is directly mixed with an appropriate lysis buffer. The maximum fragment size generally obtained is about 150–200 bp. The method is hardly suitable for the construction of BAC libraries with average insert sizes exceeding 50 kbp (Rondon et al. 2000, MacNeil et al. 2001). For the generation of BAC libraries, the environmental samples are generally embedded in a matrix, such as agarose, prior to cell lysis. This prevents excessive shearing of the DNA.

The "liquid phase-based" method is the most straightforward DNA isolation technique. It is applicable to nearly every environmental sample. One of the commercially available DNA isolation kits can be used for this purpose. However, if silica-based purification kits are used, this is often at the expense of a smaller average fragment size and a lower yield of DNA. If larger fragment sizes for large insert library construction are needed or the sample is limited, other protocols are preferable. One such protocol was published by Zhou et al. (1996). Cells are lysed in a high salt buffer containing Proteinase K, cetyl trimethyl ammonium bromide (CTAB), and sodium dodecyl sulfate (SDS). The lysis of gram-positive bacteria is enhanced by several freeze and thaw cycles. DNA extracted from plankton samples may be used directly for downstream sequencing or library construction. DNA from other samples, such as sediment samples, may need further purification in order to be suitable for downstream enzymatic manipulation. This purification can be accomplished by gel electrophoresis (Rondon et al. 2000, Quaiser et al. 2003), anion exchange chromatography (Krüger et al. 2003), density gradient centrifugation (MacNeil et al. 2001, Courtois et al. 2003), or by using commercially available kits.

Isolation of high quality genomic DNA fragments larger than 200 kbp is more difficult. DNA fragments of more than 100 kbp are easily sheared by pipetting or sample mixing. Therefore, the sample needs to be embedded in agarose, and dialysed against different buffers supplemented with enzymes and detergents to remove proteins and lipids from the embedded cells, leaving naked DNA behind (Green et al. 1997, Sambrook and Russel 2001). After these treatments, plankton samples can be used directly for BAC library construction (Beja et al. 2000). The use of this method with other, more contaminated environmental samples, such as sediment samples, resulted in agarose plugs with partly degraded genomes and enzyme inhibitors, which could not be removed by the lysis procedure. Further purification of the DNA, for example by electrophoresis through conventional agarose gels or two-phase agarose gels containing polyvinylpyrrolidone (Quaiser et al. 2002), or dialysis against a high salt and formamide containing buffer (Liles et al. 2008), was necessary in these cases. However, this procedure often leads to a reduction in quantity and shearing of the DNA, which may be afterwards inadequate for the construction of large insert BAC libraries.

2.3.4 Amplification of Genomic DNA

Comprehensive metagenome analyses generally require microgram amounts of genomic DNA in order to be representative (see below). Theoretically, less DNA would be sufficient for a representative metagenome analysis, if the microbial diversity in the sample is low or strategies have been applied to reduce the diversity and to enrich for genomic DNA of target cell populations. However, standard metagenomic library construction and sequencing techniques are not compatible with DNA amounts much below the microgram level. In some studies, large amounts of genomic DNA cannot be obtained, e.g., from the deep biosphere (Webster et al. 2003, Biddle et al. 2008). These small quantities of starting DNA need to be amplified before metagenome analysis. To overcome this problem, different genomic DNA amplification methods have been developed (Telenius et al. 1992, Zhang et al. 1992, Breitbart et al. 2002, Breitbart and Rohwer 2005, Pinard et al. 2006) of which one, the multiple displacement amplification (MDA) with bacteriophage ϕ29 polymerase, is now widely used in metagenome analyses.

The DNA polymerase of bacteriophage ϕ29 from *Bacillus subtilis* is required for the replication of the 19,285 bp phage genome (Blanco and Salas 1985b). In addition to its DNA polymerase activity, this single polypeptide of 66,520 dalton (Blanco and Salas 1984, Watabe et al. 1984) has a $3' \rightarrow 5'$ exonuclease activity (Blanco and Salas 1985a). The error rate of the ϕ29 polymerase ($\sim 10^{-5}$ to 10^{-7}) (Esteban et al. 1993, Nelson et al. 2002, Zhang et al. 2006) is about 10–100-fold lower than the mutant frequency determined for Taq DNA polymerase (Eckert and Kunkel 1990) and the processivity is high. The replication of the bacteriophage ϕ29 genome in a protein primed reaction was determined to last 8 min (Blanco et al. 1989). The ϕ29 polymerase can also elongate DNA-primed reactions. The circular single stranded M13 genome (~7,250 bases) was isothermally replicated in 5 min, producing a strand of more than 70 kb after 40 min, due to the strand displacement capability of the ϕ29 polymerase (Blanco et al. 1989). This strand displacement capability leads to a successive detachment of the 5′-end of the synthesised DNA molecule from the template strand by the ϕ29 polymerase during DNA synthesis, resulting in a rolling circle amplification (RCA) (Fire and Xu 1995, Liu et al. 1996).

In (meta)genome analyses, the ϕ29 DNA polymerase is used in a multiply-primed amplification with 3′-terminally modified, exonuclease resistant, random hexamers as primers. This results in an exponential, hyperbranched amplification of DNA of 10,000-fold (Dean et al. 2001), or even more (Dean et al. 2002). The hexanucleotides bind at several sites to the template DNA, and are elongated by the ϕ29 polymerase. If the polymerase reaction reaches the 5′-end of the next hexamer binding to the DNA, the preceding DNA strand is displaced and the displaced single strand is again subject to the binding of hexamers and elongation. The result is a network of amplified DNA (Fig. 2.2). Disadvantages of this method include the formation of chimeras, amplification biases, and non-specific DNA amplification. Chimera formation during MDA occurs with a frequency of one rearrangement per 10–22 kb of MDA product (Zhang et al. 2006, Lasken and Stockwell 2007). In 85%

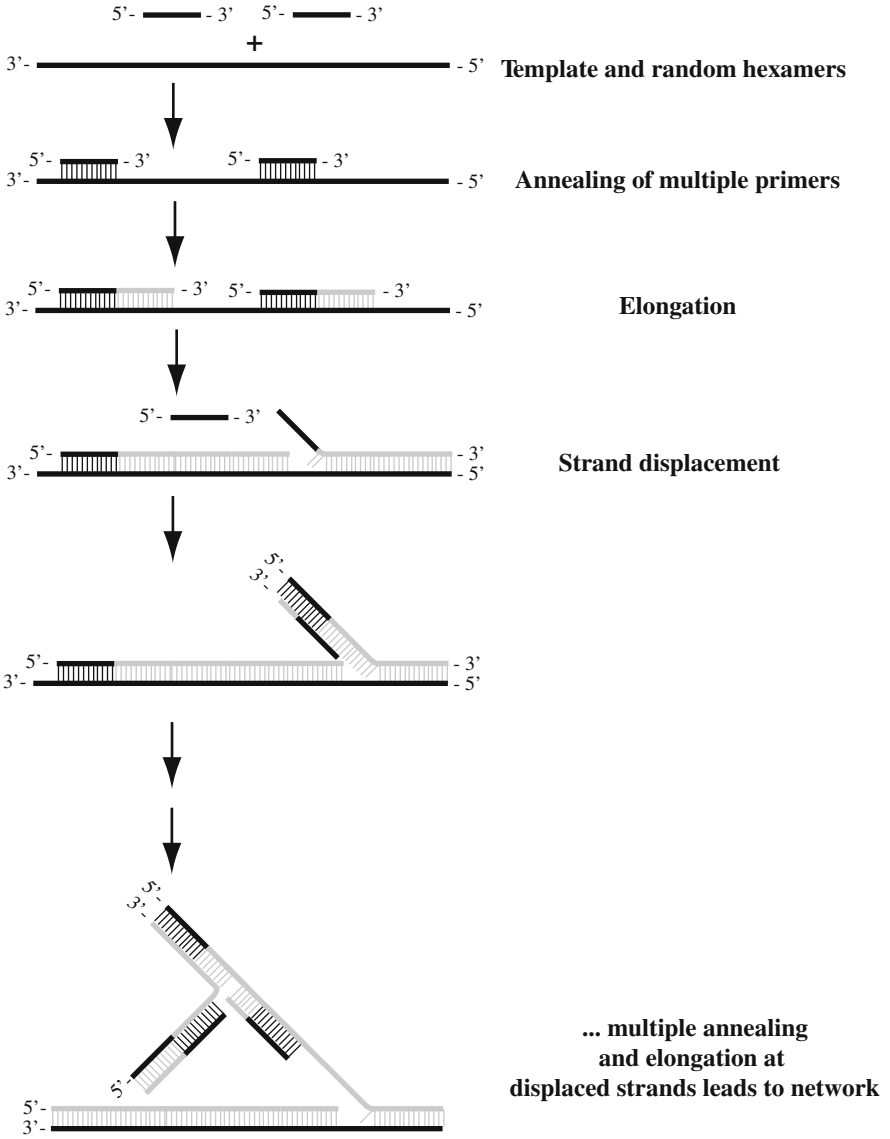

Fig. 2.2 Multiple displacement amplification of DNA. Template and random hexamers are indicated in black. Newly synthesised DNA is coloured in grey (after Binga et al. 2008)

of all cases inverted sequences were observed (Lasken and Stockwell 2007). The frequency of chimeric sequences in MDA-based libraries could be reduced by a combination of ϕ29 polymerase debranching, S1 nuclease treatment to hydrolyse single stranded regions, and nick translation using DNA polymerase I. However, this still resulted in about 6–8% chimeric inserts in a small insert library of about

3 kb insert size (Zhang et al. 2006). Increased sequencing effort and iterative assembly strategies have also been applied to overcome this problem (Zhang et al. 2006). The second issue, the amplification bias observed for MDA reactions (Dean et al. 2002, Hosono et al. 2003, Abulencia et al. 2006, Yokouchi et al. 2006) is of importance when the input is reduced to several or only one microbial cell (Raghunathan et al. 2005, Zhang et al. 2006, Kvist et al. 2007, Podar et al. 2007). As this bias seems to be sequence independent, this limitation of the method may be overcome by deep sequencing, and the combination of multiple MDA reactions from the same DNA sample (Raghunathan et al. 2005, Abulencia et al. 2006, Zhang et al. 2006). Non-specific DNA amplification is especially observed when minute amounts of DNA are amplified, and might originate from primer dimers or trace contaminants of DNA (Raghunathan et al. 2005, Zhang et al. 2006). MDA has been used in several studies for whole genome amplification (WGA) of single uncultured microorganisms, isolated from environmental samples by single cell isolation methods (Zhang et al. 2006, Kvist et al. 2007, Marcy et al. 2007, Stepanauskas and Sieracki 2007).

When a mixed sample of eight bacterial genomes of different sizes and G+C contents was amplified by MDA a significant bias towards the amplification of certain strains was observed (Abulencia et al. 2006). A case study with low biomass samples from contaminated soils revealed only small changes in the taxonomic groups detected, and a slightly higher number of species were inferred based on 16S rRNA libraries constructed from MDA treated community DNA (Abulencia et al. 2006). An investigation of microbial communities in the deep biosphere revealed a similar phylogenetic and functional category distribution in MDA amplified and non-amplified samples. However, the phylogenetic distribution of 16S rRNA genes and ribosomal proteins differed significantly in this study, probably due to the low number of such genes in the dataset (Biddle et al. 2008). An evaluation of the MDA-associated bias using 16S rRNA denaturing gradient gel electrophoresis (DGGE) fingerprints revealed a slight amplification bias when the DNA input in the MDA reaction was below one nanogram (Neufeld et al. 2008). MDA-amplified DNA has already been successfully used to construct a large insert fosmid library from heavy DNA retrieved from a stable isotope probing experiment. Chimeric artefacts were observed at the ends and within a fosmid insert, making a verification of the gene order by PCR necessary. Nevertheless, the approach represents a significant step forward in the assessment of the genetic potential, including complete operon structures, of uncultured, active microorganisms in marine surface waters (Neufeld et al. 2008).

Other studies in which the amplification of metagenomic DNA was a prerequisite for the study of uncultured marine microbes include a study of marine viral assemblages (Angly et al. 2006), an analysis of the genetic potential of two multicellular filaments of uncultured, marine, large sulfur bacteria (Mussmann et al. 2007), and an analysis of marine methanotrophic microbial consortia, isolated by a combination of fluorescence in situ hybridisation and magnetic bead capture (Pernthaler et al. 2008).

2.3.5 Construction and Analysis of Metagenomic Libraries

2.3.5.1 Small Insert Metagenomic Libraries

Small insert metagenomic libraries (generally 1.5–3 kbp insert size) in plasmid or bacteriophage λ derived vectors are a repository of the metagenome, which can be screened by sequence-based and, more often, function-based assays. Early studies of such small insert libraries (Cottrell et al. 1999, Henne et al. 1999) focused on the analysis of single metabolic genes of uncultured microorganisms. Small insert libraries in reporter gene constructs proved to be valuable to identify novel catabolic operons in a substrate-induced gene expression screening (Uchiyama et al. 2005). In a variety of studies small insert libraries have been constructed as a prerequisite for Sanger-based metagenome sequencing. High throughput sequencing of small insert libraries has proved to be a powerful method for the assessment of genome drafts (e.g. Tyson et al. 2004, Venter et al. 2004, Woyke et al. 2006), and for the study of the phylogenetic and metabolic diversity in complex habitats (Venter et al. 2004, Rusch et al. 2007). The construction of small insert libraries is straightforward. Nevertheless, there are several major drawbacks when small insert libraries are used in sequence-based metagenomic studies. First of all, a phylogenetic marker, which allows the reliable assignment of a fragment to a certain phylogenetic group, is generally missing on small inserts. If insufficient sequence data is available or if microbial diversity is high within the sample, as is the case for soil samples (Handelsman et al. 2002, Tringe et al. 2005), the reconstruction of draft genomes or even the assignment of genetic capabilities to certain species is almost impossible. Moreover, the *in silico* assembly of larger fragments involves a risk of creating chimeras, especially when the overlaps of the reassembled fragments are too short or repetitive elements are present.

2.3.5.2 Large Insert Metagenomic Libraries

Metagenomic libraries with large DNA inserts of > 30 kbp (large insert libraries) have the advantage that genome fragments can often be assigned to a specific species. This assignment can be made based on the presence of a phylogenetic marker gene, such as the 16S rRNA gene (Schleper et al. 1997, Beja et al. 2002), on the same DNA fragment. Moreover, an *in silico* assignment of large fragments to a reference sequence carrying a phylogenetic marker can be attempted, e.g., by comparing nucleotide frequencies (Krüger et al. 2003, Teeling et al. 2004). Large insert libraries can give access to complete operons (Krüger et al. 2003) or genomic islands (Schübbe et al. 2003, Mußmann et al. 2005), along with a lower risk of chimera formation and a reasonable sequencing effort. Therefore, large insert libraries were favoured in early metagenomic studies. Improved single-cell and sequencing technologies along with a decrease of sequencing costs per base pair have changed the focus within the past few years. It needs to be seen whether large insert repositories of metagenomes will be dispensable in future metagenome analyses.

Large DNA fragments can be cloned into a variety of vectors (Green et al. 1997, Tao and Zhang 1998). Among these, three are commonly used for the construction of metagenomic libraries: cosmid, fosmid and bacterial artificial chromosome (BAC) vectors. Cosmid vectors are conventional plasmids in which one or two bacteriophage λ *cos* sites have been integrated (Collins and Hohn 1978). This allows highly efficient in vitro packaging of the DNA with bacteriophage λ packaging extracts. The average size of DNA fragments that can be cloned in cosmid vectors is between 30 and 45 kbp. This is dependent on the size of the vector. The combined size of vector plus insert needs to be 78–105% of an average bacteriophage λ genome in order to be suitable for DNA packaging into bacteriophage λ phage heads (Sambrook and Russel 2001). Once in the host cell, cosmids are present at a high copy number. This facilitates further screening of the libraries (Collins and Hohn 1978, Sambrook and Russel 2001). Cosmids have been used in the construction of several metagenomic libraries (Entcheva et al. 2001, Piel 2002, Courtois et al. 2003, Schmeisser et al. 2003, Sebat et al. 2003, Lopez-Garcia et al. 2004). However, the system has two major drawbacks. Firstly, chimeras, rearrangements, and deletions have been observed (Monaco and Larin 1994), secondly the insert sizes are more or less uniform, but relatively small compared to a complete genome.

Fosmid vectors (Kim et al. 1992) are currently the most popular vectors for metagenome analyses (e.g., Stein et al. 1996, Schleper et al. 1997, 1998, Beja et al. 2002, Quaiser et al. 2002, 2003, Meyerdierks et al. 2005, DeLong et al. 2006, Neufeld et al. 2008). Fosmid cloning also takes advantage of the in vitro packaging system for bacteriophage λ, resulting in insert sizes similar to those obtained with cosmid vectors. In contrast to cosmid vectors, however, fosmid vectors are derived from the *E. coli* F(ertility)-factor and carry the replication and partition sequences of the F-factor plasmid. As a result, there are only 1–2 fosmid copies per cell and two different fosmids cannot be maintained in a single cell. The expression of toxic gene products that could be lethal for the host is reduced due to the lower copy number, and chimeras are less frequent. To facilitate screening, improved fosmid vectors have been constructed, which carry a second inducible replicon (Wild et al. 2002). This allows maintenance of the library in the low copy state, and the selective induction of 10–50 copies per cell for screening and further analysis. With these improved vectors fosmid cloning and screening is straightforward. This may be the reason for the success of these vectors in metagenome analyses.

BAC vectors (Shizuya et al. 1992) have been developed to overcome the insert size limitation of fosmids, intrinsic to the bacteriophage λ packaging system. BAC vectors, like fosmid vectors, carry the F-factor replication and partition sequences from *E. coli*. and are also available with a second inducible replicon (Handelsman et al. 2002, Wild et al. 2002). The main difference between these two cloning systems is the method used to introduce the recombinant DNA into the host. In contrast to fosmid cloning, electroporation is used for BAC library construction, to circumvent the size limitation of the bacteriophage λ packaging system. Specific *E. coli* host strains are used which have been shown to transform well with large insert constructs (Sheng et al. 1995). It has been shown that DNA fragments larger than 300 kbp can be cloned into BAC vectors and stably maintained in *E. coli* (Shizuya

et al. 1992, Kim et al. 1996, Zimmer and Verrinder 1997). For example, Bryant et al. recently analysed a 271 kbp metagenomic BAC clone from a phototrophic mat community for phylogenetic marker and phototrophy genes (Bryant et al. 2007). Nevertheless, BAC cloning is up to 100–1,000 times less efficient than fosmid cloning due to the low efficiency of electroporation as compared to in vitro packaging and transition. Moreover, the average insert size of BAC libraries is negatively correlated with the number of BAC clones generated (Leonardo and Sedivy 1990, Woo et al. 1994, Sheng et al. 1995, Zimmer and Verrinder 1997). For groups working on metagenomics, especially for those focusing on soil or marine sediments, BAC cloning is still a challenge, and often the average insert size of the constructed metagenomic BAC libraries is not much higher than those of fosmid libraries (e.g., Hughes et al. 1997, Beja et al. 2000, Rondon et al. 2000, MacNeil et al. 2001, de la Torre et al. 2003, Dumont et al. 2006). In a recent publication describing the construction of BAC libraries from soil samples, it was noted that "... large insert library construction is to some degree an idiosyncratic process..." (Liles et al. 2008). This clearly evokes the difficulties that may be encountered, especially during BAC cloning. Nevertheless, the larger an insert is the less laborious is genome walking, especially in a library from a highly diverse sample.

The technical aspects of the construction of large insert libraries have been described in detail, for example by Sambrook and Russel (2001) and by Green et al. (1997), as well as in the manuals of cloning kits. An overview is given in Fig. 2.3. Most of the refinements published over the years will not be discussed here due to space restrictions. However, one important step should be mentioned. This is the size selection of DNA fragments prior to cloning, as this greatly influences the size and quality of large insert libraries. When constructing fosmid libraries, the inclusion of an effective size selection step prevents chimera formation resulting from the ligation of two, or more, genome fragments with one vector molecule. Moreover, cloning efficiency may be influenced as ligation products which are too small or too large will not be packaged into phage heads. When constructing BAC libraries, smaller constructs will be preferentially introduced into the host cell during electroporation, and this can lead to a smaller average insert size of the resulting library (Osoegawa et al. 1998). For BAC libraries, size selection should be performed by pulsed field gel electrophoresis (PFGE), instead of conventional agarose gel electrophoresis. In a regular agarose gel electrophoresis with a constant electric field, DNA molecules larger than 15–25 kbp migrate with nearly identical mobilities. In contrast, PFGE allows size-dependent separation of DNA molecules of up to 5 Mbp in agarose gels (Sambrook and Russel 2001). Two rounds of size selection are preferable if large amounts of DNA are separated on the gel (Osoegawa et al. 1998, Rondon et al. 2000).

2.3.5.3 Metagenomic Library Size

For microorganisms in pure culture the library size (N) needed to cover the genome with a given probability is defined as $N = [\ln (1–P)]/[(\ln (1–f))]$, with P being the desired probability, and f being the fractional proportion of the genome in a

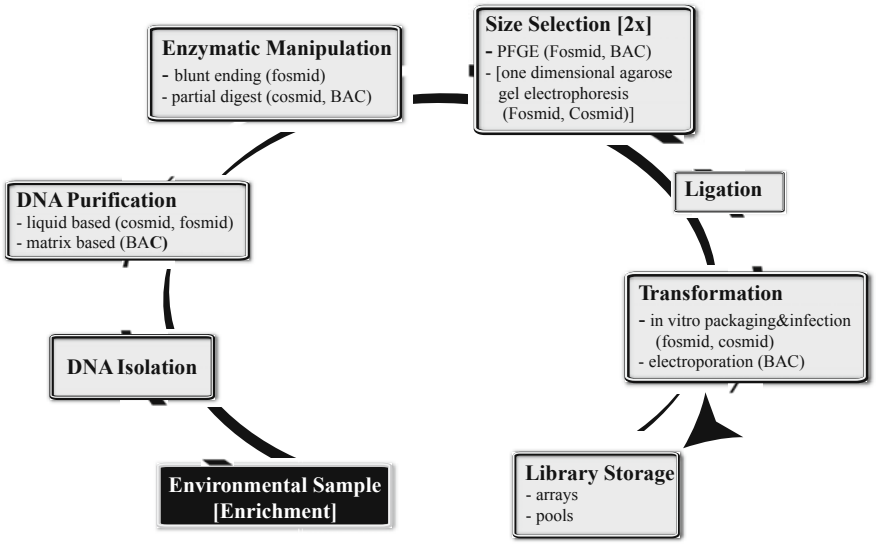

Fig. 2.3 Overview of the construction of large insert metagenomic libraries

single recombinant (e.g. average insert size/genome size) (Sambrook and Russel 2001). In metagenome analysis, several additional factors have to be considered. The size of a metagenomic library that is needed to be representative for an environmental sample, or that contains statistically one clone carrying a marker gene of interest is difficult to calculate. Among the additional factors that have to be taken into account are the diversity and composition of the microbial community, and the genome sizes of the different microorganisms in the sample. The average genome size, calculated for the 815 bacterial and archaeal genomes listed on the homepage of the National Center for Biotechnology Information (September 2009; http://www.ncbi.nlm.nih.gov/) is about 3.2 Mbp. Nevertheless, genome size can vary considerably. Currently the two extremes are the genomes of *Candidatus* Carsonella ruddii (159,662 bp), an endosymbiotic gammaproteobacterium, and *Sorangium cellulosum* "So ce 56" (13 Mbp), a soil bacterium belonging to the Myxobacteria. In addition, incomplete lysis of cells from complex environmental samples may result in a biased library. Repetitive elements and an extreme G+C content may also introduce a bias (Abulencia et al. 2006) as well as the presence of toxic gene products such as phage genes (Edwards and Rohwer 2005).

2.3.5.4 Storage of Metagenomic Libraries

Metagenomic libraries may be stored at several different stages. If the library cannot be plated at once, the ligation or the packaged phage heads can be stored at –80°C, after shock freezing. Prior to deep freezing, the phage solution has to be supplemented with DMSO (Sambrook and Russel 2001, Promega Corporation 2007)

or glycerol (Epicentre Biotechnologies 2007). Also, the recombinant cells may be stored in glycerol at –80°C immediately after transformation.

The method used for long-term storage of plated libraries depends on many factors: the uniqueness of the sample, the library size, and the method chosen for the subsequent screening. The storage of several copies of a library is generally recommended. The libraries can be stored at –80°C in pools of recombinant clones that have been washed off their agar plates using culture medium containing cryoprotectant (e.g. 5–7% glycerol). In cases where approaches such as insert-end sequencing are planned, picking and arraying of individual clones is generally conducted.

2.3.5.5 Screening of Metagenomic Libraries

Metagenomic libraries can either be screened by sequence-based or function-based approaches.

Sequence-based approaches include (i) PCR-screening, (ii) hybridisation, and (iii) insert-end sequencing. PCR-screening is the fastest method to screen a library, especially, when DNA pools or pools of recombinant clones have been prepared. Examples for complex pooling schemes have already been published (Kim et al. 1996, Asakawa et al. 1997). Several strategies have been developed to prevent amplification of the host's chromosomal DNA by cross-reacting primers. In one study, the chromosomal DNA of the host was selectively hydrolysed with an ATP-dependent DNase prior to PCR-screening (Beja et al. 2000, Liles et al. 2003). Host specific, terminally modified oligonucleotides have also been included in the PCR reaction (Goodman and Liles 2001, Liles et al. 2003). Another study added restriction fragment length polymorphism (RFLP) analysis to the PCR protocol in order to identify positive clones (Liles et al. 2003). Screening by hybridisation has been performed using colony blots (Asakawa et al. 1997, Osoegawa et al. 2000), high density DNA arrays (Rondon et al. 1999), and microarrays (Park et al. 2008). Finally, insert-end sequences can be determined to find certain genes (Kube et al. 2005) or overlapping clones for genome walking (Meyerdierks et al. 2005), and also to characterise and compare microbial assemblages (DeLong et al. 2006).

Function-based screening is based on heterologous expression of the cloned genes by the host cell. This approach is predominantly applied to identify enzymes for biotechnological purposes. If the transcription and translation machinery of the host and donor strain are incompatible, shuttle vectors may be used to transfer the library into a different host strain (Handelsman et al. 2002, Riesenfeld et al. 2004).

2.3.6 Library Independent Metagenome Analysis

The dideoxynucleotide-based DNA sequencing method was for decades the standard method for DNA sequencing. Improvements to this method, which was first described by Sanger et al. (1977), have included the use of fluorescent- instead of radioactive-labelled dideoxynucleotides as chain terminators (Prober et al. 1987). The use of this method in (meta)genome sequencing requires the construction of

shotgun libraries prior to sequencing. The most impressive examples of bulk Sanger-based sequencing of small insert metagenomic shotgun libraries to date are the sequencing of the metagenome of marine surface waters in the Sargasso Sea (Venter et al. 2004) and its follow up, the Global Ocean Survey (Rusch et al. 2007, Yooseph et al. 2007).

Since the beginning of the twenty-first century, several new sequencing methods have become available. These are allowing highly parallel, fast and cheap DNA sequencing, and they do not require the construction of shotgun libraries (Mardis 2008). Of these, pyrosequencing technology, developed by 454 Life Sciences (Margulies et al. 2005), is currently the most widely used in metagenome analysis. Pyrosequencing is based on a sequencing by synthesis technology. DNA is fragmented (300–800 bp), blunt ended, and two types of adapters, one with a biotin tag, are ligated to the fragments. Fragments carrying the same adaptor on each side are removed by streptavidin-biotin interactions. The resulting denatured, single stranded DNA fragments are diluted so that, statistically, only one fragment binds to one DNA capture bead. Subsequently, the DNA fragments are amplified in an emulsion PCR reaction in a highly parallel manner, using primers specific for the adaptor regions. About 400,000 (GS FLX) to >1 million (GS FLX Titanium) of those beads, carrying the amplified fragments are loaded onto a PicoTiterPlate so that each well contains no more than one bead. Sequencing enzymes are added and the fluidics subsystem of the sequencer flushes individual nucleotides in a fixed order across the wells. The reaction mixture is composed of the sequencing primer, the DNA template, the enzymes DNA polymerase, ATP-sulfurylase, luciferase and apyrase, as well as the substrates, adenosine 5′phosphosulfate (APS) and luciferin. The primer is elongated by the incorporation of a deoxynucleotide triphosphate (dNTP), which is catalysed by the DNA polymerase. Pyrophosphate (PPi) is released in a quantity equimolar to the amount of incorporated nucleotide. This is important because more than one nucleotide can be incorporated in a single step if a stretch of identical nucleotides is present on the target DNA. The ATP-sulfurylase converts pyrophosphate to ATP in the presence of APS. This ATP fuels the luciferase-mediated conversion of luciferin to oxyluciferin, generating visible light in amounts that are proportional to the amount of synthesised ATP. The light is detected by a charge coupled device (CCD) camera, resulting in a peak in a so called pyrogram™. Each light signal is proportional to the number of nucleotides incorporated. Each cycle is completed by an apyrase step that degrades unincorporated dNTPs and excess ATP before the next dNTP is added.

Pyrosequencing has many advantages. Massive parallel sequencing generates a vast amount of sequence information. The method is about 100 times faster than Sanger-based sequencing (Rogers and Venter 2005), and the costs per base pair are lower (Wheeler et al. 2008). There is no cloning bias and the method is less sensitive to sequencing hard stops, often found in genomes with high G+C content (Goldberg et al. 2006, Wheeler et al. 2008). A major disadvantage is the short length of the sequence reads compared to Sanger-based sequencing (∼700 bp). However, the first generation 454 GS20 sequencer with its 100 base pair reads was already

sufficient to back up genome sequencing projects (Goldberg et al. 2006), and for re-sequencing projects (Margulies et al. 2005). It was also used to obtain insights into the metabolic equipment of single uncultured TM7 cells (Marcy et al. 2007) and of about 600 cells of a single filament of uncultured marine *Beggiatoa* (Mussmann et al. 2007). Pyrosequencing technology was also a great step forward in the analyses of the metabolic capabilities of microbial communities and the differences between microbial assemblages present in habitats with distinctly different biogeochemical parameters (Angly et al. 2006, Edwards et al. 2006, Biddle et al. 2008, Dinsdale et al. 2008). However, the phylogenetic affiliation of identified genes, a binning of sequences belonging to distinct microbial species or clades and therefore a reliable assignment of functions to distinct microbial groups in a mixed microbial community was difficult due to short read lengths (Frias-Lopez et al. 2008). With the 454 GS FLX system read length improved to 200–300 bp and more than 400 bp reads can be achieved with the new 454 GS FLX Titanium system. Another issue related to this technique is the higher error rate of pyrosequencing due to the poor resolution of homopolymer stretches in template sequences (Margulies et al. 2005, Huse et al. 2007, Wheeler et al. 2008). However, this can be markedly improved by rigorous exclusion of low quality reads from the dataset (Margulies et al. 2005, Huse et al. 2007, Wheeler et al. 2008).

2.4 Bioinformatic Challenges in Metagenome Analysis

The typical workflow for metagenomic data analysis follows the well established analysis schema for full or draft genome sequencing projects (Glöckner and Meyerdierks 2006, Stothard and Wishart 2006). After the raw reads have been obtained, either from large or short insert size clone libraries, assembly is usually the first step in data processing. If longer contigs (continuous sequences) or scaffolds (contigs that still contain sequencing gaps) can be successfully established, gene calling and subsequent annotation is performed to gain insights into the phylogenetic and functional diversity of the sample. To get an overview of the functional and metabolic capacities represented in the sample, the protein coding genes are often mapped against Subsystems (Overbeek et al. 2005), the Kyoto Encyclopaedia of Gene and Genomes (KEGG) (Kanehisa et al. 2004) and the Clusters of Orthologous Groups of proteins (COGs) (Tatusov et al. 1997). Comparative metagenomics (Tringe et al. 2005, DeLong et al. 2006) and functional metagenomics approaches such as metatranscriptomics (Poretsky et al. 2005, Frias-Lopez et al. 2008) and metaproteomics (Ram et al. 2005, Wilmes and Bond 2006) are currently emerging techniques that aim at obtaining a more dynamic understanding of the differences and adaptations of the organisms to their environment.

Although processing metagenomic sequence data may seem to be straightforward a priori, this process is far from trivial in practise. In fact it is the same as trying to reconstruct a puzzle with millions of pieces where most of them show a similar colour and texture. When highly diverse environments, such as marine ecosystems, are being studied and especially when the data produced corresponds to shallow

sequencing and/or short read sequencing reads, analysis is particularly difficult. Metagenomic data analysis is still in its infancy when it comes to meeting this sort of challenge. To address these problems, this description of the bioinformatic analysis of metagenomic data will be split into the analysis of assembled metagenomes and the analysis of short read metagenomics resulting from high throughput/high diversity approaches.

2.4.1 Fragment Assembly and Binning

To obtain fragments that can be assembled into scaffolds and contigs of several kilobases in length it is best to sequence large insert constructs (BACs or fosmids). These constructs are generally sequenced using a shotgun approach as described, for example, by Sambrook and Russel 2001. About 400 sequencing reactions are required to sequence a fosmid sized fragment ($\sim$ 40 kb) to eight-fold coverage. Because all the resulting reads belong to the same large insert construct, assembly is then easily performed with standard programs such as Phrap (www.phrap.com), JAZZ (Aparicio et al. 2002), Arachne 2 (Jaffe et al. 2003) or the Celera Assembler (http://sourceforge.net/projects/wgs-assembler). If sequencing capacities are limited the crucial step in this approach is the selection of the appropriate BACs or fosmids for sequencing. Standard approaches include screening for phylogenetic or metabolic marker genes or random insert-end sequencing.

If a whole metagenome shotgun approach is chosen, the situation becomes immediately more complicated. The success of such a strategy strongly depends on the phylogenetic complexity of the sample. In low diversity environments, containing only a small number of dominant species (5–10), it is possible to obtain a good set of assembled contigs and scaffolds of up to several megabases in length with reasonable sequencing efforts (Tyson et al. 2004, Martin et al. 2006, Woyke et al. 2006). Nevertheless, problems can arise from biology due to high genome complexity even within closely related species (Johnson and Slatkin 2006) and the presence of ubiquitous genomic elements like transposons, viruses and inserted phages (Salzberg and Yorke 2005). Technical problems are even more pronounced. In a recent study Mavromatis et al. (Mavromatis et al. 2007) showed an increased chance for chimeric contigs particularly for sequence fragments below 8 kb. In general this is in line with the fact that none of the currently used assembly programs has been specifically built for metagenomes. In particular, more progressive aligners such as Phrap or JAZZ try to incorporate as many of the reads as possible, resulting in a large number of contigs. This is the primary goal in single genome assembly programs. For metagenomes a more conservative approach such as that implemented by the Arachne assembler is preferable, as this significantly decreases the chance of misassemblies.

To maintain an optimal balance between the number of contigs and the probability of obtaining chimeras, iterative cycles of “binning” and assembly can be performed.

The term binning describes the process in which sequence fragments of variable sizes are clustered into "bins", where each of the bins most probably resembles a single organism. Several binning approaches have been proposed over the last years, the simplest just mapping occasionally occurring phylogenetic marker genes, or all genes on the contigs, to taxonomic groups based on best BLAST-hits (Treusch et al. 2004, Huson et al. 2007). More sophisticated approaches use intrinsic genomic signatures based on oligonucleotide frequencies for fragment correlation or even phylogenetic classification. Numerous studies have shown, that oligonucleotide frequencies within DNA sequences exhibit species-specific patterns (Karlin et al. 1998); and for tetranucleotides it has even been demonstrated that their frequencies carry an innate but weak phylogenetic signal (Pride et al. 2003). The currently available techniques can be classified into supervised (classification) and unsupervised (clustering) methods (McHardy and Rigoutsos 2007). Here supervised means that sequence fragments are classified based on their intrinsic DNA signatures to one or many classes that have been modelled based on prior knowledge (e.g. all bacterial genomes). Recent examples for this approach are naïve Bayesian classifiers (Sandberg et al. 2001) and Phylophytia (McHardy et al. 2007). The main advantage of these systems is that they provide a direct feedback about the phylogenetic composition in the metagenome after presenting the metagenomic fragments to the trained classifiers.

The main drawback of supervised classifiers is the availability of accurately classified sequence information for training. Optimal results are only obtained in case sample specific classes e.g. trained on closely related sequenced genomes are available (Mavromatis et al. 2007, McHardy et al. 2007, Warnecke et al. 2007). This is often problematic when working with environmental samples.

Unsupervised methods like TETRA (Teeling et al. 2004) or Self Organizing Maps (SOM) (Abe et al. 2005, Abe et al. 2006) do not need a training set and produce sequence clusters independently of phylogenetic assignments. In a subsequent mapping step phylogenetic marker genes, often present on at least on one fragment in the clusters, are used to assign the sequence clusters to an organism. The clear advantage of unsupervised methods is their ability to take a broad set of sequence features, like GC content, read depth, into account, in addition to oligonucleotide frequencies and distributions. These methods are able to assign metagenomic fragments to coherent organism bins, even if no prior information is available about the prevailing organisms in a sample. The power of unsupervised correlation of sequence fragments has been recently demonstrated for a consortium of microbes involved in the anaerobic oxidation of methane (Krüger et al. 2003, Meyerdierks et al. 2005), in single cell genomic analysis of *Beggiotoa* sp. (Mussmann et al. 2007) and in the reconstruction of the genomes of the four symbionts of the marine oligochaete *Olavius* sp. (Woyke et al. 2006).

Nevertheless, both supervised and unsupervised methods are based on the statistical analysis of sequence data and therefore perform poorly if sequence fragments are shorter than 5–8 kb. For assembled data this is not a major problem because currently the standard assembly methods fall short for contigs of less than 8 kb in length (Mavromatis et al. 2007). A solution for the huge amount of single reads, which are

consequently accumulating in complex environmental samples, could be an initial round of conservative assembly combined with an unsupervised pre-binning of the resulting contigs and scaffolds. These "seed" bins can then be used as templates for further sequence correlation and clustering of the single reads or merged mate pairs, or to train systems like Phylophytia for classification. However, this will only work if sequencing depth is sufficient to produce the initial contigs, or if sequences from large insert constructs are available.

2.4.2 Gene Prediction

Classical gene predictors like ZCurve (Guo et al. 2003), Glimmer (Delcher et al. 2007) or a combination of several gene finders (Glöckner et al. 2003) work reasonably well for high quality contigs from assembled shotgun sequences or large insert constructs. If only a small number of contigs needs to be analysed a slight overpredition of 10–20%, typical of intrinsic gene predictors like ZCurve or Glimmer, can be accepted and these can later be eliminated by manual curation. Gene predictors using extrinsic strategies such as similarity based searches against existing genomic and metagenomic databases to delineate coding regions (Badger and Olsen 1999, Krause et al. 2006) require more time and tend to underpredict genes due to a lack of homologous information in the databases. In the case of large metagenomic datasets, processing time and potential loss of information can be a severe limitation. Furthermore, metagenomic gene predictors have to cope with (1) fragmented genes, (2) low sequence quality leading to frameshifts and (3) high phylogenetic diversity, which limits the initial training step of intrinsic gene finders. MetaGene, a recently developed genefinder for the analysis of metagenomic fragments, is able to cope with most of these limitations. MetaGene utilises di-codon frequencies estimated by the GC content of a given sequence. This estimation, in combination with measures of the length distribution of ORFs, the distance from the leftmost start codons and the orientation and distance of neighbouring ORFs, extracted by statistical analysis of around 130 bacterial and archaeal genomes, are used for gene prediction (Noguchi et al. 2006). The system is quite fast and, based on our experiences, currently provides the best results when tested on a broad range of metagenomic assemblies from prokaryotes.

2.4.3 Functional Annotation

Functional annotation can be regarded as the most important step in the process of analysing genomic fragments obtained from metagenomic studies. It is at this stage of the process that the investigator obtains a substantial insight into the abundance of the genetic potential available in the environmental sample being studied. Annotation should be carried out with care since poor annotations will – like the proverbial first ice crystal – start a snowball effect by continuous error propagation

in the public databases. In general, errors can be introduced by inconsistencies in functional assignments between and even within a single genome and by a simplistic procedure for assigning potential functions to the genes found. Unfortunately, there is currently no "gold-standard" for consistent (meta)genome annotation available and there are no binding rules which have to be followed by all annotators (Raes et al. 2007). To exploit the currently available data sources for functional predictions, comprehensive software systems are needed to store, analyse and visualise data, and to support the decision process by providing information from various sequence-based analysis tools. A "state-of-the-art" analysis pipeline for genomic data should include gene finding and standard bioinformatic tools for similarity-, pattern- and profile-based searches as well as prediction of signal peptides, transmembrane helices, transfer RNAs and other stable RNAs. In addition, the analysis of global and local G+C content and skews as well as codon usage and other statistical parameters can be helpful in distinguishing coding from non-coding regions. Adequate annotation systems should include automatic annotation of the protein coding regions as well as user friendly annotation facilities for manual refinement via annotation jamborees (for more details see Stothard and Wishart 2006). Finally, the reconstruction of metabolic pathways and networks helps to transform the wealth of sequence information into biological knowledge.

2.4.4 Web Based Annotation Pipelines

Current, initiatives such as the Community Sequencing Program at the Joint Genome Institute (JGI), the Microbial Genome Sequencing Project funded by the Gordon and Betty Moore Foundation, or collaborations with Genoscope, have enabled researchers worldwide to get their genome or metagenome of interest sequenced. Moreover, these centres can fund a number of projects internally, solving another important problem of cost. Unfortunately, when the sequences are recovered by the research laboratories, the processing of this data can pose a serious problem due to a lack in expertise with installing, maintaining and/or implementing the software needed for genome annotation. Initially, bioinformatic support is often provided by the sequencing centers through web-based systems, such as the Integrated Microbial Genomes (IMG and IMG/M) system (Markowitz et al. 2006, 2008), Magnifying Genomes (Vallenet et al. 2006) or CAMERA (Seshadri et al. 2007). Further online systems that accept raw (meta)genomic sequence data for processing and provide web-based visualisation of results such as BASys and PUMA2 (Van Domselaar et al. 2005, Maltsev et al. 2006) have been set up to support functional assignments and metabolic reconstruction. Recently, the RAST Server for rapid annotation using subsystem technology (Aziz et al. 2008) and comparative genomics (Overbeek et al. 2005, Ye et al. 2005) has been released. Besides gene prediction and annotation this system uses annotations to reconstruct the metabolic networks that are functioning in the studied environment and then makes all this data available for download. An experimental system specifically

designed for metagenome analysis is currently being developed and will be available at http://metagenomics.ics.nmpdr.org

2.4.5 Annotation Systems for Local Installation

The arguments for installing an annotation system locally are mainly an improved flexibility in handling and visualizing the data (Gans and Wolinsky 2007). After the first round of data mining using web-based annotation systems, users normally ask for custom tailored views on their data. To deal with such demands, full access to the databases and tools is required. Since the advent of genomics at the beginning of the 1990s several annotation systems have been made available for local installation. The most prominent are MAGPIE (Gaasterland and Sensen 1996), PEDANT Pro (Frishman et al. 2001), WIT/ERGO (Overbeek et al. 1999, 2000) and ARTEMIS (Rutherford et al. 2000). Currently, one of the most advanced is the recently developed GenDB system (Meyer et al. 2003) which has been adopted by the Networks of Excellence "Marine Genomics Europe" as their standard platform for data processing and annotation.

Without going into details about the pros and cons of the different annotation systems (an overview can be found in Stothard and Wishart 2006), we would like to suggest that the data model and versatility of GenDB is the most appropriate for the emerging demands of metagenomics. Starting with gene prediction, several options can currently be chosen in GenDB: 1. to run single gene finders such as Glimmer (Delcher et al. 2007), Critica (Badger and Olsen 1999) or 2. to use a combination of two gene finders united in the tool Reganor (McHardy et al. 2004). Exchanging the preconfigured gene predictors by MetaGene or combined systems (Bauer et al. 2006) can be done easily using a local installation. The standard tools and databases for providing functional observations for the predicted genes can be found in Table 2.1. Information about the location of a protein can be procured by

Table 2.1 Standard tools and databases providing functional observations

Tool	Databases	References
BLASTn	GenBank	http://www.ncbi.nlm.nih.gov
	EMBL	http://www.ebi.ac.uk/
BLASTp or BLASTx	GenBank	http://www.ncbi.nlm.nih.gov
	UniProt	Apweiler et al. (2004) and
	Swiss-Prot	Boeckmann et al. (2003)
HMMER	Pfam	Bateman et al. (2004)
InterProscan[a]	InterPro	Mulder et al. (2003)

[a]InterPro itself is a metadatabase that provides access to commonly used signature database such as Prosite, Prints, Pfam, ProDom, SMART, TIGR-fams, SCOP, Cath and MSD see http://www.ebi.ac.uk/interpro/.

the prediction of signal peptides with signalP (Bendtsen et al. 2004) and transmembrane helices prediction with TMHMM (Krogh et al. 2001). tRNAscan-SE (Lowe and Eddy 1997) can be used to find and assign transfer RNAs within the sequence.

Once the calculations are finished, automatic annotation systems like Metanor, provided by GenDB, or MicHanThi (Quast 2006) attempt to automatically generate annotations for all the predicted genes based on the observations returned by the individual tools. This supports the manual annotation process by providing additional information for decision making. The MicHanThi system is currently able to delineate and annotate hypothetical and conserved hypothetical genes in a nearly quantitative manner. For genes with significant hits in primary or secondary databases, such as UniProt or Pfam, the system is consistently able to assign the correct functional category. In the subsequent manual annotation process, every predicted gene should be investigated for significant hits to entries in the databases. Starting with hits to Swiss-Prot and taking into account Pfam and InterPro results the annotators have to integrate the information, read additional literature, and finally assign a function to the genes. The history system implemented in GenDB tracks all annotation changes and thus parallel annotations by different experts, or automatic annotations systems, can be handled for every gene. After finishing the annotation process, metabolic reconstructions can be performed to the extent that this is possible with the dataset in hand. The easiest way to do this is to automatically map the EC-numbers to the corresponding KEGG pathway maps (Kanehisa et al. 2004) provided by the GenDB system.

The rather slow web based visualisation system of GenDB has recently been complemented by JCoast a software tool for data mining and comparison of prokaryotic (meta)genomes (Richter et al. 2008). JCoast offers a flexible graphical user interface (GUI), as well as an application programming interface (API) that facilitates direct back-end data access. The system offers individual genome, cross-genome and metagenome analysis, and assists the biologist to explore large and complex (meta)genomics datasets. The system can also work independently of an existing GenDB installation as long as an appropriate database backend exists.

2.4.6 High Diversity Environments, Shallow Sequencing and Short Read Technologies

Mastering the bioinformatic analysis of metagenomes becomes increasingly difficult as the environments studied increase in their diversity, especially if this factor is combined with shallow sequencing. This has been nicely demonstrated by the recently published Global Ocean Sampling campaigns (Venter et al. 2004, Seshadri et al. 2007, Yooseph et al. 2007). Although billions of bases were obtained, only two genomes could be reconstructed and reasonably large scaffolds could only be assembled for the most dominant community members. Nevertheless, the incredible repertoire of genes now available in our public databases has already stimulated a broad range of follow up research aimed at investigating the diversity and function

of microorganisms in marine ecosystems (DeLong 2005, DeLong and Karl 2005, Harrington et al. 2007, Sabehi et al. 2007, Yutin et al. 2007, Kagan et al. 2008).

When assembly fails, gene prediction has to cope with the additional problem of fragmentation. When this is the case, BLASTx searches of the "environmental gene tags" (EGT) (Tringe and Rubin 2005) against UniProt or Swiss-Prot can at least provide some information about the available functional space. Two strategies can be used to compensate for these limitations: (1) pooling of sequence data from several sampling sites to increase coverage, (2) establishment of a set of reference genomes as templates for assembly and comparison (see the Gordon and Betty Moore Marine Microbiology Project). Pooling has been used for the GOS datasets (Rusch et al. 2007), although this has incidentally caused a nightmare for ecologists. Building up of the "oceans community genome" has led to the lost of data about specific adaptations to oceanic provinces and the correlation of these adaptations with habitat parameters. This problem is compounded when short read technologies like pyrosequencing are used. A recent study on the influence of read length on functional predictions showed that 100 bp fragments missed on average 72% of the BLASTx hits found by long reads (~750 bp) in the Sargasso, AMD (Tyson et al. 2004) and Chesapeake bay datasets (Bench et al. 2007, Wommack et al. 2008). Nevertheless, it has to be mentioned that a recent study by Mou et al. (Mou et al. 2008), which analysed populations involved in the metabolism of organic carbon compounds, showed that valuable information can be obtained even with short read sequencing, provided that targeted marine microcosm approaches are used.

2.4.7 Metagenome Descriptors for Comparative Metagenomics

Descriptors of phylogenetic and functional diversity can be used to obtain a better understanding of the metagenome under investigation and to compare metagenomes especially from environments with high biological diversity.

2.4.7.1 Phylogenetic Diversity

Several approaches can be used to describe phylogenetic diversity: (1) binning and classification of the fragments as described above (2) phylogenetic analysis of the ribosomal RNA genes (3) best BLAST hit mapping (4) analysis of single-copy or equal-copy marker genes. Phylogenetic assignment of the ribosomal RNA genes is rather straightforward. They can be mapped against the up to date ribosomal RNA (rRNA) sequence databases provided by SILVA (Pruesse et al. 2007) or the Ribosomal Database Project II (Cole et al. 2007). The advantage of the SILVA databases are that they provide quality checked and aligned sequences for *Eukarya* as well. Furthermore, in addition to 16S/18S databases, 23S/28S rRNA databases are provided for download at www.arb-silva.de. The SILVA compatible ARB software suite can be used for detailed phylogenetic tree reconstruction, providing advanced alignment, tree reconstruction and visualization tools (Ludwig et al. 2004).

A commonly used alternative to show an overview of the taxonomic composition of the metagenome, is the analysis of the taxonomic affiliation of the best hit by searching for similarities of the contigs or genes to the UniProt or nr databases (Treusch et al. 2004, DeLong et al. 2006, Turnbaugh et al. 2006). Due to misclassifications of the sequences in the databases this is often not very accurate and curated genome databases like the genomesDB provided by GenBank and implemented in the JCoast system (Richter et al. 2008) are preferable for taxonomic breakdowns. A recently introduced alternative is based on the mapping of single-copy or equal-copy marker genes onto a reference species phylogeny (Ciccarelli et al. 2006, von Mering et al. 2007). The advantage of the system is that 31 phylogenetic marker genes can be used, compared to only two available markers when rRNA sequences are applied. It is important to note that the coverage of the phylogenetic protein marker databases is still rather small (only around 600 per gene) when compared to the more than 800,000 publicly available rRNA genes. Nevertheless, the two methods provide different views of the data and should therefore be regarded as complementary (Raes et al. 2007).

2.4.7.2 Functional Diversity

As described above, functional diversity should be primarily accessed and described by classical annotation and metabolic reconstruction approaches involving manual curation. Unfortunately, the lack of common standards means that there is a significant difference in the quality of data handling and interpretation and this can hamper comparative metagenomic analyses (Raes et al. 2007). Furthermore, the incredible amount of data produced by metagenomic sequencing campaigns often renders time consuming approaches unrealistic. In consequence, the analysis is often limited to the determination of general statistical descriptors such as over- and underrepresentation of genes (Tringe et al. 2005) or the richness, membership and structure of microbial communities (Schloss and Handelsman 2008). These approaches are often the only available option for obtaining comparative insights into the functional adaptations of the populations, especially when only shallow or short read length sequencing data is available for several sampling sites. Successful examples of this sort of analysis include the investigation of stratified microbial assemblages in the North Pacific Subtropical Gyre (DeLong et al. 2006), the comparison of mesocosms amended with DMSP (Mou et al. 2008) and the investigation of several viral communities (Edwards and Rohwer 2005, Angly et al. 2006, Culley et al. 2006).

In summary, for high diversity environments a “gene-centric” approach invoking the descriptors mentioned above, plus additional ones like differences in functional categories, is useful to obtain a better view of the ecology and function of the microbial communities. With low diversity communities, where only a few dominating species are present, the classical “genome-centric” approach is often preferable because this provides more detailed information. The assignment of the assembled contigs to organism bins combined with subsequent annotation usually enables the reconstruction of individual metabolic properties and leads to

sound hypotheses about how the organisms share their resources and energy (Tyson et al. 2004, Meyerdierks et al. 2005, Martin et al. 2006, Woyke et al. 2006). The challenge is now to expand the bioinformatic toolbox so that it can deal with complex environments and provide an integrated understanding of marine ecosystem functioning.

2.5 Outlook

In the current context of global change, it is crucial to generate a broad understanding of the key players and processes that regulate life on earth. Marine ecosystems, which cover more than 70% of the earth surface, represent the majority of the planetary biomass and contribute significantly to the global cycles of matter and energy. Microorganisms are known to be the "gatekeepers" of these processes and insights into their life-style and fitness will enhance our ability to monitor, model and predict future changes. The capability of sequencing DNA samples from natural environments without prior cultivation of the organisms present in these samples provide an unprecedented opportunity to investigate the microbial diversity and functions on the molecular level. In the long run this will allow us to address questions which are central for marine ecology: (1) Which microbes are in the environment? (2) How abundant are they? (3) What is their functional potential? and (4) How are their activities and adaptations linked to environmental conditions?

The power of metagenomics to provide descent answers to some of the questions has been recently shown. Dinsdale et al. (2008) demonstrated that functional differences of nine qualitatively categorized, discrete biomes can be used for discrimination. Gianoulis et al. (2009) took this a step further by calculating metabolic footprints based on an ensemble of weighted pathways that maximally covaries with a combination of environmental variables. They even suggest that such footprints can be used as environmental indicators when no measurable environmental factors are available.

Although metagenomics has been shown to be a formidable tool, there are still several limitations that need to be addressed and solved. A major obstacle is that it is often impossible to assign specific abilities to individual organisms, especially if highly diverse environments are sampled. The general descriptors indicating the prevailing phylogenetic and functional diversity cannot provide answers to questions such as: "Who does what?" and "How do they work together and exchange energy and nutrients?" Single cell genomics based on physical separation of cells and subsequent whole-genome multiple displacement amplification (MDA) (Lasken 2007) is an emerging technology. This approach has been shown to provide valuable insights into the genomic potential of large sulfide oxidisers such as *Beggiatoa* sp. (Mussmann et al. 2007) and a set of marine organisms isolated from the Gulf of Maine bacterioplankton (Stepanauskas and Sieracki 2007).

Another common criticism of sequence dominated metagenomics is the discrepancy between the rate of data accumulation and the rate of knowledge generation.

To progress from piling up a rather static repertoire of gene inventories more information about the expression of the genes, with respect to different sampling sites and environmental conditions, is needed. Although, still in an early phase metatranscriptomics (Poretsky et al. 2005, Bailly et al. 2007) and metaproteomics (Ram et al. 2005, Wilmes and Bond 2006) are becoming increasingly visible and promise insights into the dynamics and regulation of genes in naturally occurring microbial communities. In a recent example Frias-Lopez et al. (Frias-Lopez et al. 2008) investigated gene expression in ocean surface waters by cDNA pyrosequencing. In addition to identifying expressed genes from key metabolic pathways, the study detected a significant amount of highly expressed hypothetical genes. These genes were most probably involved in the specific adaptation of the organisms and should be primary targets for further functional assessment.

To transfer the current flood of metagenomic data into biological knowledge, data generation and storage need to be organised and standardised to handle the sequences, genomes, genes and predicted metabolic functions and relate them to the environment that is being analysed (DeLong and Karl 2005, Lombardot et al. 2006, Markowitz 2007). The Genomic Standards Consortium (GSC) has been recently established to work on a richer description of our complete collection of genomes and metagenomes (Field et al. 2007, 2008). The Minimum Information about a Metagenome Sequence (MIMS) initiative intends to standardise contextual data acquisition so that investigators are required to anchor their metagenomes in space and time by providing a minimum of information including GPS coordinates plus depth/altitude and sampling time (see http://gensc.org). More information about the physical-chemical characteristics of the samples is needed so that genomic features can be correlated with habitat properties. This would also allow organism-specific adaptations to be identified and the role and impact of organisms on the environment to be assessed. To complement missing environmental information and to obtain a dynamic picture of the stability of the ecosystem under investigation, in situ measurements can be complemented by global data layers. An initial approach aimed at integrating (meta)genomic information with interpolated habitat parameters from global ocean data layers is already available (see www.megx.net; Lombardot et al. 2006). If this approach is developed further, it might be possible to overlay functional diversity, metabolic pathways and phylogenetic diversity upon physical, chemical and biological data to map microbial features onto marine provinces.

To allow sound comparisons between the different metagenomic investigations standardization of the processing pipeline is another aspect that is crucial. Cyberinfrastructures like the CAMERA project in the marine field will be helpful in this respect by processing a significant portion of the data. Raes et al. (Raes et al. 2007) have proposed MINIMESS, the MINImal Metagenome Sequence analysis Standard, where basic parameters for assembly, species and functional composition and coverage as well as biological and technical factors need to be documented for each metagenome. Such efforts needs to be merged and made transparent to the data providers and biologists. It will be a major community effort to work on the integrity and quality of our ever growing metagenomic inventories to improve our

understanding of how marine ecosystems function. In the long run these data could represent the starting point for modelling the complex interplay and networks found in the environment, leading to a new science of ecosystems biology.

Acknowledgments We cordially thank R. Amann, L. Raggi, I. Pizzetti for reading the manuscript and valuable comments. The work was funded by the Max Planck Society.

References

Abe T, Sugawara H, Kanaya S et al (2006) A novel bioinformatics tool for phylogenetic classification of genomic sequence fragments derived from mixed genomes of uncultured environmental microbes. Polar Biosci 20:103–112

Abe T, Sugawara H, Kinouchi M et al (2005) Novel phylogenetic studies of genomic sequence fragments derived from uncultured microbe mixtures in environmental and clinical samples. DNA Res 12:281–290

Abulencia CB, Wyborski DL, Garcia JA et al (2006) Environmental whole-genome amplification to access microbial populations in contaminated sediments. Appl Environ Microbiol 72: 3291–3301

Amann R, Fuchs BM (2008) Single-cell identification in microbial communities by improved fluorescence in situ hybridization techniques. Nat Rev Microbiol 6:339–348

Amann RI, Ludwig W, Schleifer KH (1995) Phylogenetic identification and in situ detection of individual microbial cells without cultivation. Microbiol Rev 59:143–169

Angly FE, Felts B, Breitbart M et al (2006) The marine viromes of four oceanic regions. PLoS Biol 4:2121–2131

Aparicio S, Chapman J, Stupka E et al (2002) Whole-genome shotgun assembly and analysis of the genome of *Fugu rubripes*. Science 297:1301–1310

Apweiler R, Bairoch A, Wu CH et al (2004) UniProt: the Universal Protein knowledgebase. Nucleic Acids Res 32:D115–D119

Asakawa S, Abe I, Kudoh Y et al (1997) Human BAC library: construction and rapid screening. Gene 191:69–79

Aziz RK, Bartels D, Best AA et al (2008) The RAST server: rapid annotations using subsystems technology. BMC Genomics 9:75

Badger JH, Olsen GJ (1999) CRITICA: coding region identification tool invoking comparative analysis. Mol Biol Evol 16:512–524

Bailly J, Fraissinet-Tachet L, Verner MC et al (2007) Soil eukaryotic functional diversity, a metatranscriptomic approach. ISME J 1:632–642

Bateman A, Coin L, Durbin R et al (2004) The PFAM protein families database. Nucleic Acids Res 32:D138–D141

Bauer M, Kube M, Teeling H et al (2006) Whole genome analysis of the marine Bacteroidetes 'Gramella forsetii' reveals adaptations to degradation of polymeric organic matter. Environ Microbiol 8:2201–2213

Beja O, Aravind L, Koonin EV et al (2000) Bacterial rhodopsin: evidence for a new type of phototrophy in the sea. Science 289:1902–1906

Beja O, Koonin EV, Aravind L et al (2002) Comparative genomic analysis of archaeal genotypic variants in a single population and in two different oceanic provinces. Appl Environ Microbiol 68:335–345

Beja O, Spudich EN, Spudich JL et al (2001) Proteorhodopsin phototrophy in the ocean. Nature 411:786–789

Bench SR, Hanson TE, Williamson KE et al (2007) Metagenomic characterization of Chesapeake bay virioplankton. Appl Environ Microbiol 73:7629–7641

Bendtsen JD, Nielsen H, von Heijne G et al (2004) Improved prediction of signal peptides: SignalP 3.0. J Mol Biol 340:783–795

Biddle JF, Fitz-Gibbon S, Schuster SC et al (2008) Metagenomic signatures of the Peru Margin subseafloor biosphere show a genetically distinct environment. Proc Natl Acad Sci U S A 105:10583–10588

Binga EK, Lasken RS, Neufeld JD (2008) Something from (almost) nothing: the impact of multiple displacement amplification on microbial ecology. ISME J 2:233–241

Blanco L, Bernad A, Lazaro JM et al (1989) Highly efficient DNA-synthesis by the phage phi-29 DNA-Polymerase – symmetrical mode of DNA-replication. J Biol Chem 264: 8935–8940

Blanco L, Salas M (1984) Characterization and purification of a phage phi-29–encoded DNA-polymerase required for the initiation of replication. Proc Natl Acad Sci USA-Biol Sci 81:5325–5329

Blanco L, Salas M (1985a) Characterization of a 3′-5′ exonuclease activity in the phage phi-29-encoded DNA-polymerase. Nucleic Acids Res 13:1239–1249

Blanco L, Salas M (1985b) Replication of phage phi-29 DNA with purified terminal protein and DNA-polymerase – synthesis of full-length phi-29 DNA. Proc Natl Acad Sci U S A 82: 6404–6408

Boeckmann B, Bairoch A, Apweiler R et al (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res 31:365–370

Borneman J (1999) Culture-independent identification of microorganisms that respond to specified stimuli. Appl Environ Microbiol 65:3398–3400

Breitbart M, Rohwer F (2005) Method for discovering novel DNA viruses in blood using viral particle selection and shotgun sequencing. Biotechniques 39:729–736

Breitbart M, Salamon P, Andresen B et al (2002) Genomic analysis of uncultured marine viral communities. Proc Natl Acad Sci U S A 99:14250–14255

Bryant DA, Costas AMG, Maresca JA et al (2007) *Candidatus* Chloracidobacterium thermophilum: an aerobic phototrophic acidobacterium. Science 317:523–526

Bryant DA, Frigaard NU (2006) Prokaryotic photosynthesis and phototrophy illuminated. Trends Microbiol 14:488–496

Chung KT, Ferris DH (1996) Martinus Willem Beijerinck (1851–1931) – Pioneer of general microbiology. ASM News 62:539–543

Ciccarelli FD, Doerks T, von Mering C et al (2006) Toward automatic reconstruction of a highly resolved tree of life. Science 311:1283–1287

Cole JR, Chai B, Farris RJ et al (2007) The ribosomal database project (RDP-II): introducing myRDP space and quality controlled public data. Nucleic Acids Res 35:D169–D172

Collins J, Hohn B (1978) Cosmids – type of plasmid gene-cloning vector that is packageable in vitro in bacteriophage lambda-heads. Proc Natl Acad Sci U S A 75:4242–4246

Connon SA, Giovannoni SJ (2002) High-throughput methods for culturing microorganisms in very-low-nutrient media yield diverse new marine isolates. Appl Environ Microbiol 68: 3878–3885

Cottrell MT, Moore JA, Kirchman DL (1999) Chitinases from uncultured marine microorganisms. Appl Environ Microbiol 65:2553–2557

Courtois S, Cappellano CM, Ball M et al (2003) Recombinant environmental libraries provide access to microbial diversity for drug discovery from natural products. Appl Environ Microbiol 69:49–55

Culley AI, Lang AS, Suttle CA (2006) Metagenomic analysis of coastal RNA virus communities. Science 312:1795–1798

Curtis TP, Sloan WT, Scannell JW (2002) Estimating prokaryotic diversity and its limits. Proc Natl Acad Sci U S A 99:10494–10499

de la Torre JR, Christianson LM, Beja O et al (2003) Proteorhodopsin genes are distributed among divergent marine bacterial taxa. Proc Natl Acad Sci U S A 100:12830–12835

DeLong EE (2005) Microbial community genomics in the ocean. Nat Rev Microbiol 3:459–469

DeLong EF, Karl DM (2005) Genomic perspectives in microbial oceanography. Nature 437: 336–342

DeLong EF, Preston CM, Mincer T et al (2006) Community genomics among stratified microbial assemblages in the ocean's interior. Science 311:496–503

Dean FB, Hosono S, Fang LH et al (2002) Comprehensive human genome amplification using multiple displacement amplification. Proc Natl Acad Sci U S A 99:5261–5266

Dean FB, Nelson JR, Giesler TL et al (2001) Rapid amplification of plasmid and phage DNA using phi29 DNA polymerase and multiply-primed rolling circle amplification. Genome Res 11:1095–1099

Delcher AL, Bratke KA, Powers EC et al (2007) Identifying bacterial genes and endosymbiont DNA with Glimmer. Bioinformatics 23:673–679

Dinsdale EA, Edwards RA, Hall D et al (2008) Functional metagenomic profiling of nine biomes. Nature 452:629–U632

Drews G (2000) The roots of microbiology and the influence of Ferdinand Cohn on microbiology of the 19th century. FEMS Microbiol Rev 24:225–249

Dumont MG, Murrell JC (2005) Stable isotope probing – linking microbial identity to function. Nat Rev Microbiol 3:499–504

Dumont MG, Radajewski SM, Miguez CB et al (2006) Identification of a complete methane monooxygenase operon from soil by combining stable isotope probing and metagenomic analysis. Environ Microbiol 8:1240–1250

Eckert KA, Kunkel TA (1990) High fidelity DNA-synthesis by the *Thermus aquaticus* DNA-polymerase. Nucleic Acids Res 18:3739–3744

Edwards RA, Rodriguez-Brito B, Wegley L et al (2006) Using pyrosequencing to shed light on deep mine microbial ecology. BMC Genomics 7:57

Edwards RA, Rohwer F (2005) Viral metagenomics. Nat Rev Microbiol 3:504–510

Entcheva P, Liebl W, Johann A et al (2001) Direct cloning from enrichment cultures, a reliable strategy for isolation of complete operons and genes from microbial consortia. Appl Environ Microbiol 67:89–99

Epicentre Biotechnologies (2007) CopyControl^{TM} Fosmid Library Production Kit

Erkel C, Kube M, Reinhardt R et al (2006) Genome of Rice Cluster I archaea – the key methane producers in the rice rhizosphere. Science 313:370–372

Esteban JA, Salas M, Blanco L (1993) Fidelity of phi-29 DNA-polymerase – comparison between protein-primed initiation and DNA polymerization. J Biol Chem 268:2719–2726

Field D, Garrity G, Gray T et al (2007) eGenomics: cataloguing our complete genome collection III. Comp Funct Genom 2007:1–7

Field D, Garrity G, Selengut J et al (2008) Towards a richer description of our complete collection of genomes and metagenomes: the "Minimum Information about a Genome Sequence" (MIGS) specification. Nat Biotechnol 26:541–547

Fire A, Xu SQ (1995) Rolling replication of short DNA circles. Proc Natl Acad Sci U S A 92: 4641–4645

Frias-Lopez J, Shi Y, Tyson GW et al (2008) Microbial community gene expression in ocean surface waters. Proc Natl Acad Sci U S A 105:3805–3810

Frishman D, Albermann K, Hani J et al (2001) Functional and structural genomics using PEDANT. Bioinformatics 17:44–57

Gaasterland T, Sensen CW (1996) MAGPIE: automated genome interpretation. Trends Genet 12:76–78

Gans JD, Wolinsky M (2007) Genomorama: genome visualization and analysis. BMC Bioinformatics 8:204

Gianoulis TA et al (2009) Quantifying environmental adaptation of metabolic pathways in metagenomics. Proc Natl Acad Sci U S A 106:1374–1379

Giovannoni SJ, Britschgi TB, Moyer CL et al (1990) Genetic diversity in Sargasso Sea bacterioplankton. Nature 345:60–63

Glöckner FO, Kube M, Bauer M et al (2003) Complete genome sequence of the marine planctomycete *Pirellula* sp. strain 1. Proc Natl Acad Sci U S A 100:8298–8303

Glöckner FO, Meyerdierks A (2006) Metagenome analysis. In: Stackebrandt E (ed) Molecular identification, systematics, and population structure of prokaryotes. Springer-Verlag, Heidelberg

Gloess S, Grossart HP, Allgaier M et al (2008) Use of laser microdissection for phylogenetic characterization of polyphosphate-accumulating bacteria. Appl Environ Microbiol 74:4231–4235

Goldberg SMD, Johnson J, Busam D et al (2006) A Sanger/pyrosequencing hybrid approach for the generation of high-quality draft assemblies of marine microbial genomes. Proc Natl Acad Sci U S A 103:11240–11245

Goodman RM, Liles M (2001) Template specific termination in a polymerase chain reaction. US Patent 6,248,567

Green ED, Birren B, Klapholz S et al (1997) Genome analysis: a laboratory manual, 1st edn. Cold Spring Harbor Laboratory Press, New York

Guo FB, Ou HY, Zhang CT (2003) ZCURVE: a new system for recognizing protein-coding genes in bacterial and archaeal genomes. Nucleic Acids Res 31:1780–1789

Hall AR (1989) The Leeuwenhoek Lecture, 1988 – Antoni van Leeuwenhoek 1632–1723. Notes Rec R Soc Lond 43:249–273

Hallam SJ, Girguis PR, Preston CM et al (2003) Identification of methyl coenzyme M reductase A (*mcr*A) genes associated with methane-oxidizing archaea. Appl Environ Microbiol 69: 5483–5491

Handelsman JRM (1998) Molecular biological access to the chemistry of unknown soil microbes – a new frontier for natural products. Chem Biol 5:R245–R249

Handelsman J, Liles M, Mann D et al (2002) Cloning the metagenome: culture-independent access to the diversity and functions of the uncultivated microbial world. In: Functional microbial genomics. Academic Press Inc., San Diego, pp 241–255

Harrington ED, Singh AH, Doerks T et al (2007) Quantitative assessment of protein function prediction from metagenomics shotgun sequences. Proc Natl Acad Sci U S A 104: 13913–13918

Henne A, Daniel R, Schmitz RA et al (1999) Construction of environmental DNA libraries in *Escherichia coli* and screening for the presence of genes conferring utilization of 4-hydroxybutyrate. Appl Environ Microbiol 65:3901–3907

Hosono S, Faruqi AF, Dean FB et al (2003) Unbiased whole-genome amplification directly from clinical samples. Genome Res 13:954–964

Huber R, Burggraf S, Mayer T et al (1995) Isolation of a hyperthermophilic archaeum predicted by in situ RNA Analysis. Nature 376:57–58

Hughes DS, Felbeck H, Stein JL (1997) A histidine protein kinase homolog from the endosymbiont of the hydrothermal vent tubeworm *Riftia pachyptila*. Appl Environ Microbiol 63:3494–3498

Huse SM, Huber JA, Morrison HG et al (2007) Accuracy and quality of massively parallel DNA pyrosequencing. Genome Biol 8:9

Huson DH, Auch AF, Qi J et al (2007) MEGAN analysis of metagenomic data. Genome Res 17:377–386

Hutchison CA, Venter JC (2006) Single-cell genomics. Nat Biotechnol 24:657–658

Ishoy T, Kvist T, Westermann P et al (2006) An improved method for single cell isolation of prokaryotes from meso-, thermo- and hyperthermophilic environments using micromanipulation. Appl Microbiol Biotechnol 69:510–514

Jaffe DB, Butler J, Gnerre S et al (2003) Whole-genome sequence assembly for mammalian genomes: Arachne 2. Genome Res 13:91–96

Johnson PLF, Slatkin M (2006) Inference of population genetic parameters in metagenomics: a clean look at messy data. Genome Res 16:1320–1327

Kagan J, Sharon I, Beja O et al (2008) The tryptophan pathway genes of the Sargasso Sea metagenome: new operon structures and the prevalence of non-operon organization. Genome Biol 9:R20

Kanehisa M, Goto S, Kawashima S et al (2004) The KEGG resource for deciphering the genome. Nucleic Acids Res 32:D277–D280

Karlin S, Campbell AM, Mrazek J (1998) Comparative DNA analysis across diverse genomes. Ann Rev Genet 32:185–225

Kaufmann SHE, Schaible UE (2005) 100th anniversary of Robert Koch's Nobel Prize for the discovery of the tubercle bacillus. Trends Microbiol 13:469–475

Kim UJ, Birren BW, Slepak T et al (1996) Construction and characterization of a human bacterial artificial chromosome library. Genomics 34:213–218

Kim UJ, Shizuya H, Dejong PJ et al (1992) Stable propagation of cosmid sized human DNA inserts in an F-factor based vector. Nucleic Acids Res 20:1083–1085

Krause L, Diaz NN, Bartels D et al (2006) Finding novel genes in bacterial communities isolated from the environment. Bioinformatics 22:E281–E289

Krogh A, Larsson B, von Heijne G et al (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J Mol Biol 305:567–580

Krüger M, Meyerdierks A, Glockner FO et al (2003) A conspicuous nickel protein in microbial mats that oxidize methane anaerobically. Nature 426:878–881

Kube M, Beck A, Meyerdierks A et al (2005) A catabolic gene cluster for anaerobic benzoate degradation in methanotrophic microbial Black Sea mats. Syst Appl Microbiol 28:287–294

Kvist T, Ahring BK, Lasken RS et al (2007) Specific single-cell isolation and genomic amplification of uncultured microorganisms. Appl Microbiol Biotechnol 74:926–935

Lasken RS (2007) Single-cell genomic sequencing using Multiple Displacement Amplification. Curr Opin Microbiol 10:510–516

Lasken RS, Stockwell TB (2007) Mechanism of chimera formation during the Multiple Displacement Amplification reaction. BMC Biotechnol 7:19

Leininger S, Urich T, Schloter M et al (2006) Archaea predominate among ammonia-oxidizing prokaryotes in soils. Nature 442:806–809

Leonardo ED, Sedivy JM (1990) A new vector for cloning large eukaryotic DNA segments in *Escherichia coli*. Bio-Technology 8:841–844

Liles MR, Manske BF, Bintrim SB et al (2003) A census of rRNA genes and linked genomic sequences within a soil metagenomic library. Appl Environ Microbiol 69:2684–2691

Liles MR, Williamson LL, Rodbumrer J et al (2008) Recovery, purification, and cloning of high-molecular-weight DNA from soil microorganisms. Appl Environ Microbiol 74:3302–3305

Liu DY, Daubendiek SL, Zillman MA et al (1996) Rolling circle DNA synthesis: small circular oligonucleotides as efficient templates for DNA polymerases. J Am Chem Soc 118:1587–1594

Lombardot T, Kottmann R, Pfeffer H et al (2006) Megx.net – database resource for marine ecological genomics. Nucleic Acids Res 34:D390–D393

Lopez-Garcia P, Brochier C, Moreira D et al (2004) Comparative analysis of a genome fragment of an uncultivated mesopelagic crenarchaeote reveals multiple horizontal gene transfers. Environ Microbiol 6:19–34

Lowe TM, Eddy SR (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res 25:955–964

Ludwig W, Strunk O, Westram R et al (2004) ARB: a software environment for sequence data. Nucleic Acids Res 32:1363–1371

MacNeil IA, Tiong CL, Minor C et al (2001) Expression and isolation of antimicrobial small molecules from soil DNA libraries. J Mol Microbiol Biotechnol 3:301–308

Maltsev N, Glass E, Sulakhe D et al (2006) PUMA2 – grid-based high-throughput analysis of genomes and metabolic pathways. Nucleic Acids Res 34:D369–D372

Marcy Y, Ouverney C, Bik EM et al (2007) Dissecting biological "dark matter" with single-cell genetic analysis of rare and uncultivated TM7 microbes from the human mouth. Proc Natl Acad Sci U S A 104:11889–11894

Mardis ER (2008) Next-generation DNA sequencing methods. Annu Rev Genomics Hum Genet 9:387–402

Margulies M, Egholm M, Altman WE et al (2005) Genome sequencing in microfabricated high-density picolitre reactors. Nature 437:376–380

Markowitz VM (2007) Microbial genome data resources. Curr Opin Biotechnol 18:267–272

Markowitz VM, Ivanova NN, Szeto E et al (2008) IMG/M: a data management and analysis system for metagenomes. Nucleic Acids Res 36:D534–D538
Markowitz VM, Korzeniewski F, Palaniappan K et al (2006) The integrated microbial genomes (IMG) system. Nucleic Acids Res 34:D344–D348
Martin HG, Ivanova N, Kunin V et al (2006) Metagenomic analysis of two enhanced biological phosphorus removal (EBPR) sludge communities. Nat Biotechnol 24:1263–1269
Martín-Cuadrado AB, López-García P, Alba JC et al (2007) Metagenomics of the deep Mediterranean, a warm bathypelagic habitat. PLoS One 2:e914
Mavromatis K, Ivanova N, Barry K et al (2007) Use of simulated data sets to evaluate the fidelity of metagenomic processing methods. Nat Methods 4:495–500
McHardy AC, Goesmann A, Pühler A et al (2004) Development of joint application strategies for two microbial gene finders. Bioinformatics 20:1622–1631
McHardy AC, Martin HG, Tsirigos A et al (2007) Accurate phylogenetic classification of variable-length DNA fragments. Nat Methods 4:63–72
McHardy AC, Rigoutsos I (2007) What's in the mix: phylogenetic classification of metagenome sequence samples. Curr Opin Microbiol 10:499–503
Meyer F, Goesmann A, McHardy AC et al (2003) GenDB – an open source genome annotation system for prokaryote genomes. Nucleic Acids Res 31:2187–2195
Meyerdierks A, Kube M, Lombardot T et al (2005) Insights into the genomes of archaea mediating the anaerobic oxidation of methane. Environ Microbiol 7:1937–1951
Monaco AP, Larin Z (1994) YACs, BACs, PACs and MACs – artificial chromosomes as research tools. Trends Biotechnol 12:280–286
Mou XZ, Sun SL, Edwards RA et al (2008) Bacterial carbon processing by generalist species in the coastal ocean. Nature 451:708–U711
Mulder NJ, Apweiler R, Attwood TK et al (2003) The InterPro Database, 2003 brings increased coverage and new features. Nucleic Acids Res 31:315–318
Mussmann M, Hu FZ, Richter M et al (2007) Insights into the genome of large sulfur bacteria revealed by analysis of single filaments. PLoS Biol 5:1923–1937
Mußmann M, Richter M, Lombardot T et al (2005) Clustered genes related to sulfate respiration in uncultured prokaryotes support the theory of their concomitant horizontal transfer. J Bacteriol 187:7126–7137
Nelson JR, Cai YC, Giesler TL et al (2002) TempliPhi, phi 29 DNA polymerase based rolling circle amplification of templates for DNA sequencing. Biotechniques, 44–47
Neufeld JD, Chen Y, Dumont MG et al (2008) Marine methylotrophs revealed by stable-isotope probing, multiple displacement amplification and metagenomics. Environ Microbiol 10: 1526–1535
Noguchi H, Park J, Takagi T (2006) MetaGene: prokaryotic gene finding from environmental genome shotgun sequences. Nucleic Acids Res 34:5623–5630
Olsen GJ, Lane DJ, Giovannoni SJ et al (1986) Microbial ecology and evolution – a ribosomal-RNA Approach. Ann Review of Microbiol 40:337–365
Olsen GJ, Woese CR, Overbeek R (1994) The winds of (evolutionary) change – breathing new life into microbiology. J Bacteriol 176:1–6
Osoegawa K, Tateno M, Woon PY et al (2000) Bacterial artificial chromosome libraries for mouse sequencing and functional analysis. Genome Res 10:116–128
Osoegawa K, Woon PY, Zhao B et al (1998) An improved approach for construction of bacterial artificial chromosome libraries. Genomics 52:1–8
Ottesen EA, Hong JW, Quake SR et al (2006) Microfluidic digital PCR enables multigene analysis of individual environmental bacteria. Science 314:1464–1467
Overbeek R, Begley T, Butler RM et al (2005) The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. Nucleic Acids Res 33:5691–5702
Overbeek R, Fonstein M, D'Souza M et al (1999) The use of gene clusters to infer functional coupling. Proc Natl Acad Sci U S A 96:2896–2901

Overbeek R, Larsen N, Pusch GD et al (2000) WIT: integrated system for high-throughput genome sequence analysis and metabolic reconstruction. Nucleic Acids Res 28:123–125

Pace NR, Stahl DA, Olsen GJ et al (1985) Analyzing natural microbial populations by rRNA sequences. ASM News 51:4–12

Park SJ, Kang CH, Chae JC et al (2008) Metagenome microarray for screening of fosmid clones containing specific genes. FEMS Microbiol Lett 284:28–34

Pernthaler A, Dekas AE, Brown CT et al (2008) Diverse syntrophic partnerships from deep-sea methane vents revealed by direct cell capture and metagenomics. Proc Natl Acad Sci U S A 105:7052–7057

Piel J (2002) A polyketide synthase-peptide synthetase gene cluster from an uncultured bacterial symbiont of *Paederus* beetles. Proc Natl Acad Sci U S A 99:14002–14007

Pinard R, de Winter A, Sarkis GJ et al (2006) Assessment of whole genome amplification-induced bias through high-throughput, massively parallel whole genome sequencing. BMC Genomics 7:21

Podar M, Abulencia CB, Walcher M et al (2007) Targeted access to the genomes of low-abundance organisms in complex microbial communities. Appl Environ Microbiol 73:3205–3214

Poretsky RS, Bano N, Buchan A et al (2005) Analysis of microbial gene transcripts in environmental samples. Appl Environ Microbiol 71:4121–4126

Pride DT, Meinersmann RJ, Wassenaar TM et al (2003) Evolutionary implications of microbial genome tetranucleotide frequency biases. Genome Res 13:145–158

Prober JM, Trainor GL, Dam RJ et al (1987) A system for rapid DNA sequencing with fluorescent chain-terminating dideoxynucleotides. Science 238:336–341

Promega Corporation (2007) Technical Bulletin: Packagene® Lambda DNA Packaging System.

Pruesse E, Quast C, Knittel K et al (2007) SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. Nucleic Acids Res 35:7188–7196

Quaiser A, Ochsenreiter T, Klenk HP et al (2002) First insight into the genome of an uncultivated crenarchaeote from soil. Environ Microbiol 4:603–611

Quaiser A, Ochsenreiter T, Lanz C et al (2003) Acidobacteria form a coherent but highly diverse group within the bacterial domain: evidence from environmental genomics. Mol Microbiol 50:563–575

Quast C (2006) MicHanThi – design and implementation of a system for the prediction of gene functions in genome annotation projects. Diploma thesis. Department of Computer Science and Microbial Genomics Group. University Bremen and Max Planck Institute for Marine Microbiology, Bremen

Raes J, Foerstner KU, Bork P (2007) Get the most out of your metagenome: computational analysis of environmental sequence data. Curr Opin Microbiol 10:490–498

Raghunathan A, Ferguson HR, Bornarth CJ et al (2005) Genomic DNA amplification from a single bacterium. Appl Environ Microbiol 71:3342–3347

Ram RJ, VerBerkmoes NC, Thelen MP et al (2005) Community proteomics of a natural microbial biofilm. Science 308:1915–1920

Rappe MS, Connon SA, Vergin KL et al (2002) Cultivation of the ubiquitous SAR11 marine bacterioplankton clade. Nature 418:630–633

Richter M, Lombardot T, Kostadinov I et al (2008) JCoast – A biologist-centric software tool for data mining and comparison of prokaryotic (meta)genomes. BMC Bioinformatics 9:177

Riesenfeld CS, Schloss PD, Handelsman J (2004) Metagenomics: genomic analysis of microbial communities. Ann Rev Genet 38:525–552

Robidart JC, Bench SR, Feldman RA et al (2008) Metabolic versatility of the *Riftia pachyptila* endosymbiont revealed through metagenomics. Environ Microbiol 10:727–737

Rogers YH, Venter JC (2005) Genomics – Massively parallel sequencing. Nature 437:326–327

Rondon MR, August PR, Bettermann AD et al (2000) Cloning the soil metagenome: a strategy for accessing the genetic and functional diversity of uncultured microorganisms. Appl Environ Microbiol 66:2541–2547

Rondon MR, Raffel SJ, Goodman RM et al (1999) Toward functional genomics in bacteria: analysis of gene expression in *Escherichia coli* from a bacterial artificial chromosome library of *Bacillus cereus*. Proc Natl Acad Sci U S A 96:6451–6455

Rusch DB, Halpern AL, Sutton G et al (2007) The Sorcerer II Global Ocean Sampling expedition: Northwest Atlantic through Eastern Tropical Pacific. PLoS Biol 5:398–431

Rutherford K, Parkhill J, Crook J et al (2000) Artemis: sequence visualization and annotation. Bioinformatics 16:944–945

Sabehi G, Kirkup BC, Rozenberg M et al (2007) Adaptation and spectral tuning in divergent marine proteorhodopsins from the eastern Mediterranean and the Sargasso Seas. ISME J 1:48–55

Salzberg SL, Yorke JA (2005) Beware of mis-assembled genomes. Bioinformatics 21:4320–4321

Sambrook J, Russel DW (2001) Molecular cloning: a laboratory manual, 3rd edn. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York

Sandberg R, Winberg G, Branden CI et al (2001) Capturing whole-genome characteristics in short sequences using a naive Bayesian classifier. Genome Res 11:1404–1409

Sanger F, Nicklen S, Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A 74:5463–5467

Schirmer A, Gadkari R, Reeves CD et al (2005) Metagenomic analysis reveals diverse polyketide synthase gene clusters in microorganisms associated with the marine sponge *Discodermia dissoluta*. Appl Environ Microbiol 71:4840–4849

Schlegel HG (1996) Winogradsky discovered a new Modus vivendi. Anaerobe 2:129–136

Schleper C, Delong EF, Preston CM et al (1998) Genomic analysis reveals chromosomal variation in natural populations of the uncultured psychrophilic archaeon *Cenarchaeum symbiosum*. J Bacteriol 180:5003–5009

Schleper C, Swanson RV, Mathur EJ et al (1997) Characterization of a DNA polymerase from the uncultivated psychrophilic archaeon *Cenarchaeum symbiosum*. J Bacteriol 179: 7803–7811

Schloss PD, Handelsman J (2008) A statistical toolbox for metagenomics: assessing functional diversity in microbial communities. BMC Bioinformatics 9:34

Schmeisser C, Stockigt C, Raasch C et al (2003) Metagenome survey of biofilms in drinking-water networks. Appl Environ Microbiol 69:7298–7309

Schmidt TM, Delong EF, Pace NR (1991) Analysis of a marine picoplankton community by 16s ribosomal-RNA gene cloning and sequencing. J Bacteriol 173:4371–4378

Schwartz M (2001) The life and works of Louis Pasteur. J Appl Microbiol 91:597–601

Schübbe S, Kube M, Scheffel A et al (2003) Characterization of a spontaneous nonmagnetic mutant of *Magnetospirillum gryphiswaldense* reveals a large deletion comprising a putative magnetosome island. J Bacteriol 185:5779–5790

Sebat JL, Colwell FS, Crawford RL (2003) Metagenomic profiling: microarray analysis of an environmental genomic library. Appl Environ Microbiol 69:4927–4934

Seshadri R, Kravitz SA, Smarr L et al (2007) CAMERA: a Community Resource for Metagenomics. PLoS Biol 5:e75

Sheng Y, Mancino V, Birren B (1995) Transformation of *Escherichia coli* with large DNA molecules by electroporation. Nucleic Acids Res 23:1990–1996

Shizuya H, Birren B, Kim UJ et al (1992) Cloning and stable maintenance of 300-kilobase-pair fragments of human DNA in *Escherichia coli* using an F-factor-based vector. Proc Natl Acad Sci U S A 89:8794–8797

Sogin ML, Morrison HG, Huber JA et al (2006) Microbial diversity in the deep sea and the underexplored "rare biosphere". Proc Natl Acad Sci U S A 103:12115–12120

Stahl DA, Amann R (1991) Development and application of nucleic acid probes. In: Stackebrandt E, Goodfellow M (eds) Nucleic acid techniques in bacterial systematics. John Wiley & Sons Ltd., Chichester, UK, pp 205–248

Stein JL, Marsh TL, Wu KY et al (1996) Characterization of uncultivated prokaryotes: isolation and analysis of a 40-kilobase-pair genome fragment front a planktonic marine archaeon. J Bacteriol 178:591–599

Stepanauskas R, Sieracki ME (2007) Matching phylogeny and metabolism in the uncultured marine bacteria, one cell at a time. Proc Natl Acad Sci U S A 104:9052–9057
Stoffels M, Ludwig W, Schleifer KH (1999) rRNA probe-based cell fishing of bacteria. Environ Microbiol 1:259–271
Stothard P, Wishart DS (2006) Automated bacterial genome analysis and annotation. Curr Opin Microbiol 9:505–510
Strous M, Pelletier E, Mangenot S et al (2006) Deciphering the evolution and metabolism of an anammox bacterium from a community genome. Nature 440:790–794
Tao Q, Zhang HB (1998) Cloning and stable maintenance of DNA fragments over 300 kb in *Escherichia coli* with conventional plasmid-based vectors. Nucleic Acids Res 26:4901–4909
Tatusov RL, Koonin EV, Lipman DJ (1997) A genomic perspective on protein families. Science 278:631–637
Teeling H, Meyerdierks A, Bauer M et al (2004) Application of tetranucleotide frequencies for the assignment of genomic fragments. Environ Microbiol 6:938–947
Telenius H, Carter NP, Bebb CE et al (1992) Degenerate oligonucleotide-primed PCR – General amplification of target DNA by a single degenerate primer. Genomics 13:718–725
Thornhill DJ, Wiley AA, Campbell AL et al (2008) Endosymbionts of *Siboglinum fiordicum* and the phylogeny of bacterial endosymbionts in *Siboglinidae* (Annelida). Biol Bull 214:135–144
Torsvik V, Goksoyr J, Daae FL (1990) High diversity in DNA of soil bacteria. Appl Environ Microbiol 56:782–787
Treusch AH, Kletzin A, Raddatz G et al (2004) Characterization of large-insert DNA libraries from soil for environmental genomic studies of Archaea. Environ Microbiol 6:970–980
Tringe SG, Rubin EM (2005) Metagenomics: DNA sequencing of environmental samples. Nat Rev Genet 6:805–814
Tringe SG, von Mering C, Kobayashi A et al (2005) Comparative metagenomics of microbial communities. Science 308:554–557
Turnbaugh PJ, Ley RE, Mahowald MA et al (2006) An obesity-associated gut microbiome with increased capacity for energy harvest. Nature 444:1027–1031
Tyson GW, Chapman J, Hugenholtz P et al (2004) Community structure and metabolism through reconstruction of microbial genomes from the environment. Nature 428:37–43
Uchiyama T, Abe T, Ikemura T et al (2005) Substrate-induced gene-expression screening of environmental metagenome libraries for isolation of catabolic genes. Nat Biotechnol 23:88–93
Urbach E, Vergin KL, Giovannoni SJ (1999) Immunochemical detection and isolation of DNA from metabolically active bacteria. Appl Environ Microbiol 65:1207–1213
Vallenet D, Labarre L, Rouy Z et al (2006) MaGe: a microbial genome annotation system supported by synteny results. Nucleic Acids Res 34:53–65
Van Domselaar GH, Stothard P, Shrivastava S et al (2005) BASys: a web server for automated bacterial genome annotation. Nucleic Acids Res 33:W455–W459
Venter JC, Remington K, Heidelberg JF et al (2004) Environmental genome shotgun sequencing of the Sargasso Sea. Science 304:66–74
von Mering C, Hugenholtz P, Raes J et al (2007) Quantitative phylogenetic assessment of microbial communities in diverse environments. Science 315:1126–1130
Ward DM, Weller R, Bateson MM (1990) 16s ribosomal-RNA sequences reveal numerous uncultured microorganisms in a natural community. Nature 345:63–65
Warnecke F, Luginbuhl P, Ivanova N et al (2007) Metagenomic and functional analysis of hindgut microbiota of a wood-feeding higher termite. Nature 450:U560–U565
Watabe K, Leusch M, Ito J (1984) Replication of bacteriophage phi-29 DNA in vitro – the roles of terminal protein and DNA-polymerase. Proc Natl Acad Sci USA-Biol Sci 81:5374–5378
Webster G, Newberry CJ, Fry JC et al (2003) Assessment of bacterial community structure in the deep sub-seafloor biosphere by 16S rDNA-based techniques: a cautionary tale. J Microbiol Meth 55:155–164
Wheeler DA, Srinivasan M, Egholm M et al (2008) The complete genome of an individual by massively parallel DNA sequencing. Nature 452:U872–U875

Wild J, Hradecna Z, Szybalski W (2002) Conditionally amplifiable BACs: switching from single-copy to high-copy vectors and genomic clones. Genome Res 12:1434–1444

Wilmes P, Bond PL (2006) Metaproteomics: studying functional gene expression in microbial ecosystems. Trends Microbiol 14:92–97

Wommack KE, Bhavsar J, Ravel J (2008) Metagenomics: read length matters. Appl Environ Microbiol 74:1453–1463

Woo SS, Jiang JM, Gill BS et al (1994) Construction and characterization of a bacterial artificial chromosome library of *Sorghum bicolor*. Nucleic Acids Res 22:4922–4931

Woyke T, Teeling H, Ivanova NN et al (2006) Symbiosis insights through metagenomic analysis of a microbial consortium. Nature 443:950–955

Ye YZ, Osterman A, Overbeek R et al (2005) Automatic detection of subsystem/pathway variants in genome analysis. Bioinformatics 21:I478–I486

Yokouchi H, Fukuoka Y, Mukoyama D et al (2006) Whole-metagenome amplification of a microbial community associated with scleractinian coral by multiple displacement amplification using phi 29 polymerase. Environ Microbiol 8:1155–1163

Yooseph S, Sutton G, Rusch DB et al (2007) The Sorcerer II Global Ocean Sampling expedition: expanding the universe of protein families. PLoS Biol 5:432–466

Yutin N, Suzuki MT, Teeling H et al (2007) Assessing diversity and biogeography of aerobic anoxygenic phototrophic bacteria in surface waters of the Atlantic and Pacific Oceans using the Global Ocean Sampling expedition metagenomes. Environ Microbiol 9:1464–1475

Zengler K, Toledo G, Rappe M et al (2002) Cultivating the uncultured. Proc Natl Acad Sci U S A 99:15681–15686

Zhang L, Cui XF, Schmitt K et al (1992) Whole Genome Amplification from a single cell – implications for genetic-analysis. Proc Natl Acad Sci U S A 89:5847–5851

Zhang K, Martiny AC, Reppas NB et al (2006) Sequencing genomes from single cells by polymerase cloning. Nat Biotechnol 24:680–686

Zhou J, Bruns MA, Tiedje JM (1996) DNA recovery from soils of diverse composition. Appl Environ Microbiol 62:316–322

Zimmer R, Verrinder GA (1997) Construction and characterization of a large-fragment chicken bacterial artificial chromosome library. Genomics 42:217–226

Zwirglmaier K, Ludwig W, Schleifer KH (2004) Improved method for polynucleotide probe-based cell sorting, using DNA-coated microplates. Appl Environ Microbiol 70:494–497

Chapter 3
Populations and Pathways: Genomic Approaches to Understanding Population Structure and Environmental Adaptation

Melody S. Clark, Arnaud Tanguy, Didier Jollivet, François Bonhomme, Bruno Guinand, and Frédérique Viard

Abstract The field of Genomics has essentially been fuelled by medical research with developments in human gene therapy, such as the Human Genome Project. This major international undertaking resulted in a significantly increased sequencing capacity, a dramatic decrease in the time and cost of sequencing and also the computational effort required for the analysis. Marine biologists are taking advantage of this high throughput technology, hence, the tools are now available to answer questions that would have not been possible even five years ago. Genomics, in terms of studying DNA, can effectively define the genetic structure of populations and as a consequence the mapping of species boundaries, approximate drift in populations and accurately measure biodiversity. Studying transcribed sequences (RNA) enables the identification of changes at the cellular level associated with the adaptation of species to particular habitats and now, in our changing environment, predicts their ability to survive perturbation.

The aim of this chapter is to familiarize the reader with the most commonly used genomic techniques that are available for population and adaptation studies. These encompass both DNA based methodologies for population studies and RNA based techniques for expression studies. Which technique is used largely depends on the species under study and the resources available. In environmental research it is important to understand that "resources" does not just refer to money for sequencing and library production, but also access to starting material and the ability to store the material successfully, often under difficult conditions. For example, a cruise to investigate a particular hydrothermal vent may only happen once in a researcher's lifetime and so material will be scarce, numbers will be limited and it may not be possible to store the material at a low enough temperature to prevent RNA degradation (thereby excluding expression studies). Also species availability tends to be on what is there at the time, rather than being able to perform a calculated choice for which species is the best to study. In other studies such as aquaculture or invasive

M.S. Clark (✉)
Ecosystems, British Antarctic Survey, Natural Environment Research Council, Cambridge CB3 0ET, UK
e-mail: mscl@bas.ac.uk

J.M. Cock et al. (eds.), *Introduction to Marine Genomics*,
Advances in Marine Genomics 1, DOI 10.1007/978-90-481-8639-6_3,

species, a means has to be found to work on a particular species even if the material is intractable as there is a defined requirement for work in that area. So having outlined the techniques, the question is how to use them?

In this chapter specific examples will be used to demonstrate how such techniques are being used to address these important ecological issues in the marine environment, concentrating on lower vertebrates (fish) and invertebrates. These encompass both population analyses and gene expression (functional) studies to understand how populations have adapted to and interact with, their environment.

Ultimately, the challenge for the marine biologist is to utilize the tools produced for the study of model organisms, where significant amounts of sequence data exist (i.e. resource rich) and to develop these for non-model species, essentially from a zero base-line. It is not an easy task.

Abbreviations

AFLP	Amplified fragment length polymorphism
BAC	Bacterial artificial chromosome
cDNA	copy/complementary DNA
COI	Cytochrome C oxidase
dbEST	database for ESTs: http://www.ncbi.nlm.nih.gov/dbEST/
DNA	Deoxyribonucleic acid
ER	Endoplasmic reticulum
EST	Expressed sequence tag
GH	Growth hormone
GHRH	Growth hormone releasing hormone
GMPD	Glycosylase mediated polymorphism detection
HSP	Heat shock protein
IPCC	Intergovernmental panel on climate change
MPA	Marine protected area
MPSS	Massively parallel sequencing signature
PR	Prolactin
Q-PCR	Quantitative or Real Time PCR
QTL	Quantitative trait locus/loci
RNA	Ribonucleic acid
SNP	Single nucleotide polymorphism
SSR	Simple sequence repeat
SYBR Green	Fluorescent green dye
UTR	Untranslated region

3.1 Tools

This first section will concentrate on the tools available, some of which can be used in either DNA and/or RNA studies and these are described below. The tools are

prefaced in the subtitles to indicate whether they are mainly used in either DNA-based (population based analyses) or RNA-based (functional/expression) studies. A brief description of each tool is given, with reference to more detailed texts if required.

3.1.1 DNA and RNA Studies: EST Libraries

EST is the abbreviation for "Expressed Sequence Tag". An EST is a single sequencing read from a cloned piece of cDNA. The production of a cDNA occurs when an RNA molecule corresponding to the expressed part of a gene is converted into a DNA copy to increase stability. Each clone is then usually sequenced from one end (generally the 5′ end of the gene to avoid the complications of sequencing errors associated with the polyA tail. It also allows more direct access to the open reading frame and thus increases the chance of finding potential gene matches using sequence similarity searching of the databases). This data is entered into a public database, dbEST (Boguski et al. 1993) (http://www.ncbi.nlm.nih.gov/dbEST/) on the understanding that the sequencing quality may not be 100% and errors may be present. ESTs can be produced from any cell/tissue/species and are a good method for discovering and identifying genes in non-model organisms. They can also be used to data-mine for markers for DNA studies (microsatellites and single nucleotide polymorphisms (SNPs)). For DNA-based studies, the source of the EST library is, not important however, this is an important consideration in expression work, along with the type of library used.

Production of lots of ESTs from a particular tissue/cell type creates a library. Libraries can be produced for just a few hundred sequences, or many thousands (largely dependent on cost and ease of access to large-scale sequencing facilities). There are essentially two approaches for producing EST libraries:

- Non-subtracted methodologies: cDNA is made directly from the cells or tissues under investigation. So the number of times a sequence is present in a library is a direct reflection on the quantity of the RNA quantity corresponding to that gene in the cell. This may mean that highly expressed sequences (cf. actin in muscle) may mask the detection of rarely expressed clones. Sequencing a library containing many highly expressed sequences is also not ideal as sequencing the same gene many times over is not efficient and only generally of utility when assessing the nucleotide polymorphism of a specific locus). This issue is particularly important when only a small number of clones are generated and sequenced, but knowing the relative quantities of an RNA molecule (gene) across in various tissues can be very useful.
- Subtracted and/or normalised methodologies: This involves hybridization steps to compete out the most highly expressed sequences in the cell/tissue and to increase the relative number of copies of rare differentially expressed sequences in the cell (cf. Suzuki et al. 1997, Carninci et al. 2002, Otsuka et al. 2003). So with this latter

technique, a greater number of different sequences are obtained in each library (gene discovery is more efficient), but no estimate of relative gene copy number is possible. This technique is more complicated than non-subtracted methodologies. It can often require larger amounts of RNA or if amplification techniques are used, sequence bias can result during the amplification process.

Both techniques of library production have their own advantages and disadvantages, the main point is to be aware of these and conduct the *in silico* analyses accordingly. The ESTs can then be used either as a data source for DNA markers, or the clones physically used for the production of gene chips or targeted gene analyses and to a large extent, it is the former where EST libraries have been put to best use to date (see Sections 3.1.2 and 3.1.3). Whilst there are numerous papers documenting the production of EST libraries (e.g. Douglas et al. 2007, Govoroun et al. 2006), they are invariably gene lists or catalogues for future studies. EST libraries per se cannot provide accurate information on environmental adaptation and any inferred findings need to be validated with in-depth functional analyses and experimental manipulation. Also non-model species suffer from lack of functional information associated with EST data, i.e. many sequences are designated as "unknown" or "putative protein" and this impacts on their usefulness for both functional and population studies (see Section 3.2.1.1).

3.1.2 DNA Studies: Microsatellites

Microsatellites or SSRs (simple sequence repeats) are small DNA stretches of a repeated core sequence of few base pairs (e.g. GT repeat units). Because they are highly polymorphic in length, they have been extensively used since the 1990s to produce genetic linkage maps and have been used in population assignments, paternity analyses and fine-scale dispersal analyses. The major drawback of this technique has long been the tremendous effort needed to generate a statistically relevant number of such polymorphic loci in non-model organisms (Zane et al. 2002).

The availability of genome databases for model organisms has considerably enhanced the possibility of finding microsatellites in non-model organisms by defining primers in regions conserved across a large set of organisms, but there is also the increasing resource of EST libraries where such markers are present in 4% of the bivalve cDNA sequences in GenBank (Saavedra and Bachere 2006). The frequency of SSR EST-based data is highly variable across species. However, microsatellites have been found in numerous ESTs libraries constructed for a wide range of ecologically and/or economically important marine species including invertebrates (e.g. the European clam *Ruditapes decussatus*, the blue mussel *Mytilus edulis*, the Japanese oyster *Crassostrea gigas* and the deep-sea vent mussel, *Bathymodiolus azoricus*, (Tanguy et al. 2008), the bay scallop *Argopecten irradians* (Roberts et al. 2005)), fishes (e.g. cod, *Gadus morhua*; halibut, *Hippoglossus hippoglossus*, Douglas et al.

2007, Weiss et al. 2007) or marine plants (e.g. eelgrass *Zostera marina*; Oetjen and Reusch 2007b). A tremendous number of EST-SSRs were also found in the 3′UTR regions of cDNAs in a collection of 100,000 sequence reads from the deep-sea vent polychaete *Alvinella pompejana* (Daguin and Jollivet, unpublished data). Microsatellites derived from genomic DNA have different properties to those derived from expressed sequences (ESTs), Oetjen and Reusch (2007b) has indicated that microsatellites derived from ESTs should be used with caution for population genetics analyses as they are likely to be in strong linkage with selected genes. However as these markers are found in known genes, they present a viable alternative to those methods that utilize anonymous genetic markers and there are several other advantages:

- They can specifically be used for studying selection processes. For example, based on the analysis of 58,146 Atlantic salmon EST sequences available in the GenBank database, 75 EST-linked microsatellites were used by Vasemägi et al. (2005) to examine the signature of selection in the Atlantic salmon *Salmo salar*.
- EST-derived markers are often conserved across species enabling comparative studies across a wide range of non-model species. There is no comprehensive review of such a transferability among aquatic organisms, but in plants, such an approach has been very effective (Ellis and Burke 2007). A potential pitfall of cross-amplification of markers across species is the increased occurrence of null alleles which will bias estimations of allele frequencies, reduce any observed heterozygosity, and increase the apparent levels of inbreeding (DeWoody et al. 2006). However, as primers flanking EST-SSRs are generally identified in more conserved sequences than those of anonymous microsatellites, null alleles will be less of a problem. Again, we are not aware of any comparative data on the relative occurrence of null alleles among sets of markers in aquatic organisms, but this has been documented in other organisms (plants: Rungis et al. 2004; beetle: Kim et al. 2008).
- Contrasting the diversity and levels of population differentiation of microsatellite loci spread throughout non-coding DNA (i.e. “traditional” anonymous microsatellites) against produced those from coding regions (e.g. EST-SSRs; UTRs, introns) potentially allows us to gain a better estimation of population genetics parameters and of the relative strength of selective pressures acting on each type of marker (e.g. Luikart et al. 2003, Oetjen and Reusch 2007b).

3.1.3 DNA Studies: Single Nucleotide Polymorphisms (SNPs)

SNPs are considered to be the most abundant type of genetic variation (polymorphism). In coding regions, they can be used to compute the ratio of synonymous to non-synonymous substitution that is of prime interest in evolutionary studies of selection (reviewed in Ford 2002, Vasemägi and Primmer 2005). Similar

to the situation for microsatellites, the development of SNPs from a zero baseline in non-model organisms can be particularly time-consuming and expensive (Kim and Misra 2007). However, these markers have also benefited from the ever-expanding public genome databases (Kim and Misra 2007, Phillips 2007, Hayes et al. 2007). For example, SNPs were successfully obtained in the Pacific salmon, a non-model species, by using sequences from two sister taxa, the rainbow trout and the Atlantic salmon (Smith et al. 2005, Campbell and Narum 2008, respectively; see also Ryynanen and Primmer 2006 for a review). Also 318 segregating SNPs were recently isolated from 17,056 EST sequences in cod (Moen et al. 2008) and the first results based on cDNA library screening at a population level have also been obtained in marine mollusc species (Tanguy et al. 2008, Faure et al. 2007, 2008). Identification of SNPs is not solely reliant upon screening EST databases and they can be generated from scratch. In-depth methodologies for the generation of SNP markers will not be presented here (reviewed in Kim and Misra (2007) and Hudson 2008). However, one particularly useful method is that of GMPD technology (glycosylase mediated polymorphism detection; O'Leary et al. 2006, Vaughan 2000). This approach seems to be particularly useful for targeting specific genes with SNPs, and as such, the results can contrast with patterns of genetic diversity using other marker types.

As with all marker types, there are disadvantages as well as advantages: Luikart et al. (2003) highlighted the possible drawbacks associated with the use of SNPs in population genomics as they are prone to severe ascertainment bias (bias in estimating population parameters) due to the criteria used for their selection (i.e. polymorphism) and the way they are usually tested (e.g. few individuals). Indeed, in Chinook salmon (*Oncorhynchus tshawytscha*) some ascertainment bias has been reported for SNPs compared to anonymous microsatellites and allozymes (Smith et al. 2007). The main point is to be aware of the advantages and limitations of using a specific technique and construct the experiment and analyses accordingly (see Section 3.2.1 for further discussion).

3.1.4 DNA Studies: Amplified Fragment Length Polymorphisms (AFLPs)

AFLP data sets are now relatively easy to produce and reliable. This technique involves the use of restriction enzymes to cut genomic DNA. Adaptors are then ligated onto the sticky ends and a sub-set of these amplified using primers designed to the adaptor and part of the restriction site. The resulting DNA fragments are separated by size (length of sequence) using polyacrylamide gel electrophoresis or capillary sequencers. Only two allelic states are counted (presence or absence), and a band of a specific length represents the presence of such an allele at an AFLP locus.

The AFLP technique is a particularly well-established molecular technique in plants (Meudt and Clarke 2007). It is relatively easy to produce >100 AFLP markers, compared to the microsatellite studies in most non-model organisms where often

only <20 unlinked loci are used. The number of loci used is then of primary importance to correctly assess the impact of selection (e.g. Beaumont and Nichols 1996), or bottlenecks (Luikart et al. 1998). Moreover, AFLPs are useful markers for the investigation of other issues such as inbreeding (an important parameter in managing populations) (Dasmahapatra et al. 2008), and analysing systems characterized by weak population structuring (Campbell et al. 2003), as is the case in marine organisms (Ward et al. 1994, DeWoody and Avise 2000). Although much less used for marine invertebrates compared to terrestrial plants or animals, AFLPs have recently proved to be a powerful tool to tackle numerous ecological issues (Table 1).

As regards applicability, artefacts may occur in some cases when using AFLPs with, for example, the probability of amplifying bands of the same size at distinct loci (Gort et al. 2006, see also Pompanon et al. 2005) and possible confusion

Table 1 Examples of uses of AFLPs. This list is not restrictive and most population genetics and associated ecological issues can theoretically be addressed using such markers

Characteristic	Species	References
Divergence between different ecotypes	*Littorina saxatilis* periwinkle	Oetjen and Reusch (2007a)
Zones of secondary contact	*Crassostrea virginica* oyster	Murray and Hare (2006)
	Anguilla spp. eel species	Albert et al. (2006)
Stock definition	*Cyclina sinensis* clams	Zhao et al. (2007)
	Crangon crangon shrimp	Weetman et al. (2007)
	Solea vulgaris common sole	Garoia et al. (2007)
Anadromous species	*Oncorhynchus keta* chum salmon	Flannery et al. (2007)
Connectivity among populations	*Haliotis rutescens* Californian red abalone	Gruenthal et al. (2007)
	Portunus pelagicus blue swimming crab	Klinbunga et al. (2007)
	Asterina gibbosa sea star	Baus et al. (2005)
	Riftia pachyptila tubeworm	Shank and Halanych (2007)
Forensic issues	Various	Maldini et al. (2006)
Assortative mating	Florida Keys hamlets genus *Hypoplectrus*	Barretto and McCartney (2008)
Reproductive tactics	*Stichopus chloronotus* Holothurian *Heteroxenia fuscescens* coral	Uthicke and Conand (2005) and Fuchs et al. (2006)
Maximising genetic diversity when founding a hatchery population	*Salmo salar* Atlantic salmon	Hayes et al. (2006)

between alleles and loci across individuals. This leads to a form of homoplasy (similarity due to convergent evolution), a phenomenon that can be encountered for any genetic marker (e.g. Ellegren 2004). For instance, Wares and Blakeslee (2007) found this issue was the major drawback to their analysis of the invasive status of the periwinkle *Littorina littorea* along northern American coasts. As AFLP markers are dominant markers, they do not really comply with detection methods of outlier loci developed so far. Such methods (reviewed in Guinand et al. 2004) rely on the estimation of allelic frequencies assuming Hardy-Weinberg proportions. Dominant markers do not allow verification of this assumption, but, as noted by Bonin et al. (2006), locus-specific deviations from the Hardy-Weinberg equilibrium may arise and therefore bias the results of outlier detection in an unpredictable way. Finally, a major drawback of AFLP markers is that they are anonymous markers in non-coding regions. As they are not identified per se and not easily linked to genes it is difficult to use them to fill in the gaps between phenotype and genotype and therefore between observed diversity and fitness. AFLPs are best used to improve knowledge in some ecological areas such as better stock or species delineation and the topics mentioned in Table 1. However they do not particularly allow links be made with regard to individual fitness, as investigating this issue with any type of genetic marker is controversial (Balloux et al. 2004, Pemberton 2004, Slate et al. 2004, DeWoody and DeWoody 2005). The AFLP technique is now being combined with new technologies based on pyrosequencing of AFLP fragments to develop SNPs and microsatellites.

3.1.5 DNA Studies: High Through-Put Sequencing

Sanger sequencing has become a routine laboratory technique, but recent major advances with the next-generation of sequencing technologies such as massively parallel sequencing signature (MPSS; Brenner et al. 2000) and pyrosequencing (reviewed in Margulies et al. 2005, Langaee and Ronaghi 2005) (also called 454 sequencing) are revolutionising this technique, allowing simultaneous processing of millions of short sequence reads. Although challenging from a bioinformatic perspective, such technologies offer several opportunities for addressing ecology and evolution issues, among them the possibility to carry out biodiversity analysis (Venter et al. 2004). Moreover, such methods are less prone to errors due to mishandling, recovery of missing or rare transcripts or clones that are unstable when cloned into bacteria. Technical improvements make such methods increasingly reliable (e.g. Hamady et al. 2008), and expression of most transcripts, including rare variants, can be accurately and precisely quantified (e.g. Stolovitzky et al. 2005). This methodology will become increasingly used as the length of the sequence read increases (450 bp reads are now possible with 454 pyrosequencing), increasing the chance of identifying genes in non-model species (Hudson 2008). The initial technology was restricted to tens of bases which is really only of use in sequenced model organisms (Hudson 2008).

The most powerful use of short-read sequence data is comparison to other genomes that are sequenced at high quality. For instance, Vera et al. (2008) used the silkworm *Bombyx mori* to BLAST sequences and to further establish useful contigs for the Nymphalid butterfly (*Melitaea cinxia*). The *B. mori*'s genome has been estimated to have about 18,000 genes, and assuming that this is representative for Lepidoptera, Vera et al. (2008) obtained 9,000 non redundant hits in their study. This may suggest that at least half of all genes in *M. cinxia* are now at least partially identified using a single pyrosequencing run. As Ellegren (2008) noted about Vera et al.'s (2008) study, this technology offers support for association mapping, QTL and population genetics studies, and candidate gene studies (e.g. Saastamoinen and Hanski 2008). Even though the *M. Cinxia* genome data is still a draft, such a study and that of Toth et al. (2007) on the *Polistes* wasp proved that high-throughput techniques of cDNA sequencing can be reliably used in species that are of ecological and evolutionary relevance. A further example of the use of such technology is given in Section 3.3.4 and the analysis of hybrid vigour. As stated earlier, the use of this technique will increase and, as costs decrease, it becomes a viable alternative to gene chip expression analyses.

3.1.6 DNA and RNA Studies: Targeted Gene Analyses

Targeted gene analyses or the candidate gene approach (Tabor et al. 2002) are where a particular gene or a small set of genes is intensively investigated in an organism or population under different conditions (either natural or artificially induced). It is difficult to perform targeted gene analyses on non-model organisms where there is either little or no sequence data available. Gene sequences can be produced via EST libraries and then specific primers designed for the gene of interest, or a more random approach can be adopted using degenerate PCR. This latter technique involves identifying the gene of interest in several different species, hopefully including data from the same taxa as the species of interest and designing primers from the amino acid sequence incorporating codon usage degeneracy.

In expression work, targeted gene analyses are currently used where there is some knowledge of which genes may change in expression levels between environmental conditions/treatments and knowledge of the actual gene sequences. These days, analysis of target genes in expression studies is carried out using Q-PCR. DNA studies have historically focused on phylogenetic analyses and a specific set of genes such as cytochrome c oxidase subunit I (COI) and the ribosomal genes (18s, 16s and 28s) (http://www.barcoding.si.edu/; http://rdp.cme.msu.edu/, http://bioinformatics.psb.ugent.be/webtools/rRNA/). The COI gene is the gene of choice for barcoding, a technique that is described in more detail in Chapter 1 However, developments in sequencing technologies have made it possible to re-sequence a target gene in many individuals from a population in parallel. The resulting haplotype information is then used to determine either whether specific alleles can be associated with traits of interest or coalescence-based methods to tackle selective and hybridizing processes in speciation of closely-related species

(Faure et al. 2008). In a variant of this technique (called ecoTILLING), rather than using a particular marker to follow a gene, the researcher begins with a gene (usually selected based on genome sequence) which is potentially related to a phenotype of interest (Comai et al. 2004). The ecoTILLING method allows the rapid detection of natural sequence variants of this gene within a population of genotypes by PCR amplification of a population of alleles followed by mismatch detection using the CEL1 or ENDO1 endonucleases, which cleave DNA at polymorphic sites (Comai et al. 2004). Variant alleles can then be followed using molecular markers and associated with particular traits of adaptive significance.

3.1.7 DNA Studies: Barcoding

The development of rapid sequencing methods including new developments based on pyro-sequencing technology may also be applied to barcoding approaches. Barcoding is a populist name given to the technique of assigning a unique identifier to every known species. This technique is not only useful for population and taxonomy studies, but also marine forensic analyses. However, the application of this technology has already been comprehensively explained in Chapter 1and will not be explored further here, as the emphasis will remain on large-scale population analyses.

3.1.8 RNA Studies: Microarrays or Gene Chips

Microarrays represent a method of analysing the expression of many genes at the same time. Thousands of sequences are individually attached (spotted) to a small glass slide (typically the size of a microscope slide) with each spot on the slide representing a gene sequence. These sequences can either be cDNAs or oligonucleotides designed using EST or genomic sequence information. Hybridization of samples to the microarray can be performed using either a single or double fluorescent dye system, the choice of which depends on the technology used to manufacture the chip (cf. Gibson 2002). Statistical analyses are then applied to the colour and level of intensity of fluorescence to determine how the applied treatment has affected gene expression. Generally genes that show more than a $2\times$ increase in fluorescence under any conditions are regarded as having significantly changed under the assumption that the signal is normally distributed between replicates within the two samples under scrutiny. Whilst this is clearly a powerful technique for mass screening of gene expression, experiments have to be carefully designed with rigorous consideration of what constitutes a "control" and planning sample replications. If too many different variables are introduced into the experiments, then the data can become too noisy and no meaningful information can be extracted. Currently it is essential to validate microarray results using other methods of transcriptional analysis, such as Q-PCR.

3.1.9 RNA Studies: Q-PCR

Q-PCR stands for "Quantitative PCR", also sometimes known as "Real-Time PCR". With this technique, the gene of interest is assayed in control and "treated" animals using fluorescently labelled primers (often SYBR Green) and incorporation of these primers is monitored during the PCR run by laser excitation. The level of fluorescence directly reflects the amount of PCR product generated. By comparing the point in the PCR reaction where the amplification enters the log phase for both control and treated samples and determining the difference between the two, a measure of the change in relative levels of transcript abundance caused by the treatment can be made. There are a number of ways to control and validate this technique (see Pfaffl 2001, Pfaffl et al. 2002, Radonic et al. 2004). This technique is often used in tandem with microarray analyses, as the estimate of the change in expression level produced via Q-PCR is more accurate than that produced via microarray analysis. However, designing and testing specific primers is a time-consuming process. Microarrays are therefore best adapted for the initial, global overview of transcriptional activity, whereas Q-PCR is used subsequently to analyse the expression of specific candidate genes identified by the microarray analysis.

In the previous sections, the main tools that are available for ecological genomics studies have been briefly described. Note however, that this list of tools is continually expanding, fuelled by technological spin-offs from the Human Genome Project. It is now important to outline some of their uses in the marine environment and how they can contribute to our knowledge on population dynamics, biodiversity, and environmental issues.

3.2 Population Genomics

Population genomics differs from population genetics in that it involves significantly increased coverage of the genome and the use of bioinformatics facilities to deliver molecular tools as well as to analyze large datasets. Both neutral and selected markers are targeted in population genomics in order to simultaneously analyze and discriminate demographic (i.e. change in population size) and selective processes (i.e. effects of environmental constraints).

It combines the methodology developments derived from genome analyses with the conceptual framework of population genetics (for a review see Luikart et al. 2003). It can be broadly defined as the screening and analysis of a large number of DNA regions in order to address evolutionary or ecological issues (Black et al. 2001, Luikart et al. 2003, Schlotterer 2003, Feder and Mitchells-Olds 2003). Methodologies and approaches are highly interdisciplinary. For example, it is obvious that any relationship between a phenotype and an environment can be tracked to both the level of the population or of the individual in order to define the genetic basis of this phenotype. However, this relationship can also be analysed from a more functional perspective by examining a set of co-expressed genes

in a network involved in the physiological regulation of genes that led to the expression of this phenotype at the individual or population levels (Feder 2007, see also Crawford and Oleksiak 2007, Marden 2008). Some authors bring together population and functional genomics to define environmental/ecological genomics (e.g. Ungerer et al. 2008), whereas others define ecological genomics as a distinct field (e.g. Wilson et al. 2005). We do not disagree with these different approaches and it is clear that genomics *sensu lato* profoundly influences current biology. Here we describe the use of the conceptual frameworks of population and/or quantitative genetics that have the potential for identifying and studying the genetic basis of those traits affecting fitness that are key to natural selection (e.g. Ellegren and Sheldon 2008). Such research is now commonplace in model species, and is increasingly used to study fundamental questions in ecology, phenotypic plasticity and gene-environment interactions (reviewed in, e.g., Morin et al. 2004, Nielsen 2005).

Among the promising fields derived from population genomics is the understanding of complex evolutionary processes like patterns and rates of adaptation. By examining large portions of the genome, population genomics can theoretically disentangle locus specific-effects such as recombination, selection, epistatic interactions etc. that affect one or a few loci, from genome-wide effects affecting the whole genome (bottlenecks, founder events, inbreeding). (e.g. Mitchell-Olds et al. 2007, Stinchcombe and Hoekstra 2008, Ellegren and Sheldon 2008). Some selective processes may also act on the whole genome as a result of the long-term adaptation of organisms over the course of evolution (i.e. thermal or pressure effects on protein stability). Population genomics can be used to address ecological and environmental issues such as biodiversity studies and marine ecosystems surveys with applications to fisheries, monitoring of invasive species or design of Marine Protected Areas (MPAs). They also offer great promise to address the phenomenon of trade-offs that are expressed at the phenotypic level and dependent on the modulation of gene expression and molecular processes (Roff 2007).

3.2.1 Analysis: Choices, Limitations and Considerations

3.2.1.1 Marker Type

The use of several genetic marker types including SNPs on the analysis of genetic data and estimation of population parameters has been reviewed by Vasemägi and Primmer (2005) and Ryynanen et al. (2007). The study of Vasemägi et al. (2005) on Atlantic salmon can be used to illustrate issues about changes in heterozygosity, allele diversity, levels of population differentiation across sets of genetic markers, and number of loci potentially under selection. They identified genetic signatures of divergent selection by screening 95 genomic and 78 EST derived mini- and microsatellites for populations inhabiting contrasting natural environments (salt, brackish, and freshwater habitats). They detected no significant variations in heterozygosity, allele diversity or levels of population differentiation across anonymous or EST-SSRs loci when looking at all the populations surveyed as a whole.

Nevertheless, when grouping populations by type of habitats, significant differences were found regarding genetic and allelic diversities. They also found that tandem repeat markers in ESTs did not deviate more frequently from neutral expectations than anonymous genomic microsatellite loci. This could be due to the unbalanced number of markers of each type in this particular study, but similar results have been reported in other organisms (e.g. Woodhead et al. 2005). This result can be explained by the tiny fraction of EST-SSR markers (more generally gene-linked) that are influenced by selection. Indeed, Vasemägi et al. (2005) finally identified only nine putative EST-SSRs (12%) that were potentially under selective pressure.

Moen et al. (2008) carried out a similar study on cod using SNPs markers. However, the cod study did not include other types of markers to survey gene or allele diversities and levels of population differentiation. In cod, forty-eight SNPs out of a total of 318 (≈15%) had levels of genetic differentiation significantly different from zero, and twenty-nine were found to be potentially selected (≈9%). Of the latter, seventeen were associated with genes of known function.

Together these two studies indicate that the percentage of markers under selection is >9% for studies based mainly on EST-SSRs or SNPs. This figure can be compared to estimates obtained in the few genome scans that have been carried out using AFLP markers. These reported a range of ≈1.5–12.5% of loci that could have been or were influenced by selective constraints, and subsequently conferred or are actually prone to confer a potential fitness gain within a given environment (Wilding et al. 2001, Campbell and Bernatchez 2004, Bonin et al. 2006, Gruenthal and Burton 2008). The highest value reported for AFLPs (12.5%, 49 loci of a total of 392) is debatable as indicated by Bonin et al. (2006), as it depends on the statistical method used to detect loci under selection, and on the data sampling design (i.e. the geographical scale and/or the main ecological factor considered in the hierarchical data analysis). For AFLPs, this number is probably overestimated and the correct value in this example is more likely ≈2% (Bonin et al. 2006). More studies are needed so that rigorous comparisons of marker types can be carried out, but current data indicates that coding sequences involved in EST-SSRs or gene-based SNPs are (somewhat logically) better sources of adaptive polymorphisms. This number probably does not exceed 15%, indicating that a huge effort is required when establishing EST-libraries to recover a "significant" number of potentially selected genes. However, this does not negate EST libraries as a substantial resource for mining adaptive traits, particularly because they are usually in the public domain and available for anyone to exploit.

To date, very few population studies have been carried out to look at genetic variation in aquatic organisms based solely on EST-SSR markers, but the study of Vasemägi et al. (2005) does reveal some pitfalls to such an approach in non-model organisms. For example, in this study, a large number of ESTs did not match any known gene in the databases. The percentage of "unknown" genes has been estimated to be as high as 70% in the European sea bass (*Dicentrarchus labrax* (Boutet et al. 2006, Chini et al. 2006) or the polychaete *Alvinella pompejana* (Alvinella Consortium) and as low as ≈13% in halibut (Douglas et al. 2007). Hence, obtaining EST-SSRs by random EST database mining for any non-model species could

produce similar results to those by Vasemägi et al. (2005), which in their case did not directly shed light on the physiological basis of local adaptation. This could possibly be improved if EST-SSRs were only used from ESTs that matched known genes. This would improve our knowledge of how physiological regulation shapes a given phenotype and/or pattern of habitat distribution. However, this can dramatically decrease the number of such markers available for population studies, reducing the benefit of genomics in providing ecological and population tools. In consequence, functional information should not be an a priori prerequisite for choosing EST-SSR markers.

3.2.1.2 Differentiating Selective and Demographic Effects

One challenge when studying the genetic basis of adaptation is that the pattern of variation produced by genetic hitch-hiking or selective sweeps may also be the result of demographic changes. Therefore a common approach in population genetics studies is to analyse a very large number of unlinked loci (the so-called "genome scan" method; see Luikart et al. 2003, Storz 2005) in an attempt to disentangle selective and demographic effects (see also Teshima et al. 2006). As stated by, e.g., Wenne et al. (2007), the outlier loci in genome scans – i.e. loci that do not statistically conform to the neutral theory – offer the potential for the generation of informative markers which are suited for specific questions raised in management scenarios of marine species. According to the theory first developed by Maynard-Smith and Haigh (1974), outlier loci are most likely the causal loci, but are either physically linked or in linkage disequilibrium (LD) with the site(s) that undergo or have undergone selection. The extent of LD between the marker locus and the functionally relevant mutation involved in one ecological adaptation can vary dramatically across the genome and also between study systems. This will be affected by population history, mating system, recombination rate, the age of the selected allele, the strength of selection and many other factors (Nordborg and Tavare 2002). Several marker types can be used for genome scans such as anonymous microsatellites (amphibian: Bonin et al. 2006), SNPs (man: Akey et al. 2002), or a combination of several markers (cf. Chinook salmon: Smith et al. 2007). Dominant (see Bensch et al. 2002) AFLP (Vos et al. 1995) markers are also appropriate tools for carrying out such analyses as these can be produced at a moderate cost and with no previous genomic knowledge.

3.2.1.3 Identifying Adaptive Traits

The use of molecular markers, such as anonymous microsatellites, EST-SSRs and AFLPs, is not necessarily the best approach for addressing all ecological issues. This is because they are mostly neutral and do not enable the understanding of fitness differences among wild individuals that could drive adaptation in a given environment. Generally, it is not known if they are associated with a gene and, even if this is known, the role(s) of such an associated gene in metabolic/gene network(s) can be very obscure in a wide range of marine organisms. The continued development of

such markers, however, is essential for the production of fine-scale linkage maps, which ultimately can lead to identification of adaptive traits and/or genes.

Linkage analysis is the traditional method of identifying chromosomal regions containing QTLs (Quantitative Trait Loci) in model organisms. The methodology relies on following the inheritance of segregating traits in pedigrees and seeking to find (co)inheritance of traits and numerous genetic markers. Hence, statistical analyses of genome-wide molecular markers such as microsatellites and AFLPs, and phenotypes measured in the progeny of controlled crosses can be used to identify chromosomal regions contributing to phenotypic differentiation for a trait or a suite of traits of interest (reviewed in Mackay 2001, Erickson et al. 2004). If this can be established, trait loci are inferred to map in the vicinity of marker loci and markers such as microsatellites, AFLPs and even SNPs can lose their purely anonymous status. The determinism of the trait under study can then theoretically be investigated by looking at the polymorphism of a few markers. The use of inbred lines in genetic crosses is the most powerful method for QTL analysis, because it maximizes linkage disequilibrium between markers and trait loci (Lynch and Walsh 1998). Analysis is then mostly pedigree-based.

The power of linkage analysis increases with number of meioses that can be studied. Linkage maps have so far mainly been constructed for species that can be bred in captivity, including fishes, insects and mammals. In marine organisms, studies have concentrated on aquaculture species including the European sea bass (Chistiakhov et al. 2005), the sea bream (*Sparus aurata*; Franch et al. 2006), the European flat oyster (*Ostrea edulis*; Lallias et al. 2007a), the Pacific oyster (*Crassostrea gigas*; Hubert and Hedgecock 2004, Li and Guo 2004), the Eastern oyster (*Crassostrea virginica*; Yu and Guo 2003); the bay scallop (*Argopecten irradians*; Qin et al. 2007), the blacklip abalone (*Haliotis rubra*; Baranski et al. 2006), and the blue mussel (*Mytilus edulis*; Lallias et al. 2007b). However, the number of linkage maps available for aquatic organisms is rapidly increasing (Wenne et al. 2007). Most of the QTL studies to date have concerned growth related traits, and to a lesser extent disease resistance traits. Even though growth is one important proxy of fitness, other ecologically relevant traits (e.g. reproductive performance, larval duration, gene expression response to various stresses) should certainly be investigated as these could explain performances of individuals.

At present, the genetic maps available for aquatic organisms are not sufficiently fine to map adaptations as they rarely comprise more than >1,000 markers disseminated throughout the genome. An average marker distance of 20 cM is required for the location of a QTL to the correct chromosome arm (e.g. Rogers et al. 2001, Chistiakhov et al. 2005), but, for finer mapping, a marker distance of 1 cM or less is needed. Mapping chromosomal regions that co-segregate with traits of interest represents just the first step towards the identification of the causative genetic variants that shape the phenotype. Numerous genes usually reside within a targeted chromosomal region and the eventual identification of such variants requires refined mapping and nomination of candidate genes. There are only a few examples of such studies so far in natural populations, notably the mapping and subsequent identification of ectodysplasin (*Eda* gene) the armour plate patterning gene in different

populations of the three-spined stickleback (*Gasterosteus aculeatus*) (cf. Cresko et al. 2004, Colosimo et al. 2005). Allelic variants at this locus have led to the reduction of pelvic plates in the freshwater ecotype of this species, a process that has been associated with changes in predation risk.

For QTL mapping to become a commonplace methodology for studies of outbred natural populations would require sampling of either large pedigrees, or extensive series of sibling-pairs and the components of fitness measured in these individuals (e.g. mammals: Beraldi et al. 2006). This is generally not possible without a massive input of resources and a huge targeted breeding programme and hence is not really an option, particularly if no commercial advantage is gained. The nearest example related to this issue was a pilot study in the European sea bass. This reported crosses among individuals with a poorly known pedigree and allowed the characterisation of QTLs for body shape (Chatziplis et al. 2007), a trait of potential benefit to both the aquaculture industry and *par hasard* the study of wild populations.

An alternative method to QTL mapping is association or linkage disequilibrium mapping. This relies on the analysis of linkage between markers and trait loci that are in linkage disequilibrium by genome scans of population samples, rather than by pedigree analysis using QTLs. This approach has typically a much higher resolution than conventional pedigree analysis, but the efficiency of association mapping depends on the number and distribution of markers used to scan the genome, and the extent of linkage disequilibrium. This is a parameter itself that varies (Jorde 2000) and is strongly reliant on population/species history (e.g. Backström et al. 2006) and model of population structure (Yu et al. 2006). We are not aware of marine species where such a method could be used in a near future to identify causative sequence variants, as long as the number of markers in genome scans is low.

The use of genome scans using any of the markers cited above is important in confirming that QTLs identified in contemporary populations have played a part in adaptive phenotypic differentiation, driven by directional selection. By definition, QTLs may be used to infer the genetic basis of adaptive traits underlying species or population differences, but they do not rely per se on the effects selection may have on corresponding adaptive traits (Hoekstra and Nachman 2003, Rogers and Bernatchez 2005). Whilst genome scans rely on detecting potentially selected markers, the function and role of these loci on the construction of a given phenotype needs to be ascertained. Nevertheless, if signs of selection as revealed by scans for reduced within-population variability or increased between-population divergence coincides with the chromosomal location of QTLs, this highlights the significance of genes within these regions in adaptive evolution (Ellegren and Sheldon 2008). Such a result has been recently presented by Rogers and Bernatchez (2005). They integrated QTL mapping and genome scanning methods to analyse diverging sympatric pairs of the lake whitefish (*Coregonus clupeaformis*) species complex. This was to test the hypothesis that differentiation between dwarf and normal ecotypes at a growth associated QTL was maintained by selection. Indeed, their objective consisted of evaluating growth as a phenotype–environment association, determining its genetic basis with QTL mapping, screening natural populations for outlier levels of differentiation, and finally assessing observed patterns of divergence for

outliers over multiple environments. They were able to determine whether the loci closest to a growth rate QTL were the same as loci showing elevated differentiation in genome-wide scans of natural populations (Campbell and Bernatchez 2004). They found that eight AFLP loci (so-called "QTL homologues") closest to the QTL for growth rate showed values outside the empirically determined 95% confidence limits for genetic differentiation estimated from 440 AFLP loci. Thus suggesting that differentiation at these loci was due to selection on nearby growth rate loci. The authors were able to show that one AFLP locus corresponding to a growth rate QTL exhibited significantly higher levels of genetic differentiation between ecotypes than expected under neutrality. This exemplifies a similar genetic basis for adaptations of the normal and dwarf ecotypes in each lake, and how particular portions of the genome are related to dramatic phenotypic changes for a multigenic trait as growth. Hence, such a study reinforces the potential of a complementary QTL/genome scan approach towards studies of adaptive divergence. As Campbell and Bernatchez's (2004) study did not use a dense linkage map (25 of the 40 presumptive linkage groups were covered), and "only" 440 AFLP markers were used, it also means that our understanding of the basis of adaptive divergence across sister species and/or ecotypes may be improved even with the sparse QTL mapping that is presently possible in marine species.

Hence, although time intensive, costly and challenging, developing linkage maps, searching for QTLs, and combining information with genome scans arguably represents a comprehensive way of identifying genomic regions contributing to adaptive variation, especially for multigenic traits (Price 2006, Stinchcombe and Hoekstra 2008). Given this, there are still drawbacks to this approach, which are exemplified by an example in stickleback.

A microsatellite-based linkage map had established that the portion of the genome containing a large-effect QTL contributing to adaptive variation in pelvic morphology between oceanic and lake populations was shown to cover a region of $\approx$10 Mb (Shapiro et al. 2004). Associated molecular knowledge including sequencing of a BAC (bacterial artificial chromosome) covering this region of the stickleback genome enabled the identification of a gene (*Pitx1*), which showed expression differences associated with a reduced pelvis in lake populations (Shapiro et al. 2004). However, the nature of the precise molecular change driving the phenotypic change – certainly in *cis*-regulatory modules in this particular case – has yet to be identified. Such a situation raises at least three questions:

- How is a similar type of question approached without BACs or some substantial fraction of genome sequence?
- What would happen if this QTL was not a large-effect QTL? It is clear that the developmental processes of the hind limb and pelvis are under the control of a few major genes in vertebrates (e.g. Marcil et al. 2003) that translate into large-effect QTLs. QTLs affecting most ecologically important traits certainly do not belong to such a category. Hence, such clear genotype-phenotype coupling cannot be expected in most cases.

- Why so much effort (e.g. development of numerous markers, pedigrees, linkage map) for only one gene? Is this useful? Why not a candidate gene approach that deals with a few well-known genes that would come from the available biochemical and physiological literature or other techniques such as microarrays (see below, and, e.g., Ellegren and Sheldon 2008)?

Hence, although combining genome scan and QTL approaches can be useful in identifying genomic regions of interest, the relevance of individual genes and their surrounding *cis*- and *trans*-regulatory regions do necessitate extensive sequencing of resources (e.g. BACs), and comparative genomics to identify genes of interest in marine organisms.

3.3 Practical Application of Population Genomics in the Marine Environment

There are a number of applications for the tools described above and population genomics approaches, which inter-link with ecological issues:

3.3.1 Dispersal in the Sea: From Larval Development to Local Adaptation and Speciation Processes

Understanding the extent of connectivity between marine populations is crucial, not only for conservation purposes but also for fisheries management, as well as being a means to increase our understanding of the functioning and evolution of populations in the marine system. Comprehensive methodologies exist which means this issue can be addressed from larval tracking through to gene analyses. Chapter 1 can be consulted for more information on the specific area of larval identification, whilst here three more general examples are provided of genome-based approaches that fuel the debate about the patterns and extent of dispersal in the sea:

3.3.1.1 Pelagic Larval Studies

In species with a bentho-pelagic life cycle (the most frequent cycle in marine invertebrates) and in which adults are non-mobile, larvae are the major dispersal vector. Larvae which are able to metamorphose (i.e. competent larvae) can delay their metamorphosis in response to biotic and abiotic factors (Hadfield 1998, Hadfield et al. 2001). This delayed metamorphosis may enhance the dispersal potential and the connectivity between populations with implications on local adaptation, maintenance of species cohesiveness and population dynamics. Numerous studies have nevertheless shown that evolutionary trade-offs exist between characteristics linked to the dispersal abilities of the species and its cost, e.g. between the avoidance of

competition between related individuals and the increased mortality during dispersal (Pechenik 1999).

Despite its biological significance, only a limited number of studies have attempted to elucidate the mechanisms underlying the timing of acquisition of competence in marine invertebrates. Although the metabolic pathways and the substances inducing or preventing larval development, metamorphosis and settlement are studied in numerous species, the molecular and genetic bases of these processes as well as their variations in response to selective pressures are still poorly documented (Hadfield 1998). Until recently, genomic tools were poorly available for non-model organisms. As pointed out by Medina (2009), the study of non-model organisms can now provide lights on major issue about life-cycle evolution, for instance, on-going studies based on micro-arrays and Q-PCR carried out in the invasive gastropod, *Crepidula fornicata* (Taris et al. 2009) demonstrating the involvement of NO signalling pathway in larval metamorphosis. Another major illustration comes from several recent studies by Degnan and collaborators on gastropods of the genus *Haliotis* (Degnan et al. 1997, Jackson et al. 2005, Williams et al. 2009). Through the identification of cDNA fragments representing differentially-expressed genes, they obtained a suite of genes likely to be involved in metamorphosis and competence in *H. asinina* (Jackson et al. 2005). Results of an extensive study carried out by means of cDNA microarrays then suggested that pathways operating at larval metamorphosis may regulate the expression of novel genes specific to abalone and molluscs metamorphosis (Williams et al. 2009). Analyses of the conservation of these genes and of their associated polymorphisms should help in understanding the selective processes acting on larval development pathways.

3.3.1.2 Genetic Basis of Adaptive Differentiation in High Gene Flow Species

Marine species often exhibit high gene flow through larval dispersal. Such dispersal ability may prevent local adaptation in a heterogeneous environment in the absence of spawning shifts between the inter-connected populations. On the other hand, both the very high fecundities (allowing differential mortalities) and large population sizes (enabling a high selection efficiency through the parameter $2 N_e s$) provide opportunities for counteracting the homogenising effects of gene flow. The question is to determine the actual extent of adaptive polymorphism, which accounts for the presence of the same marine species in contrasting environments, and to examine under which set of conditions such an adaptation may arise and persist.

Population genomics and more specifically genome scan analyses can provide relevant tools to address these questions. In the case of divergent selection according to habitats, new beneficial alleles may sweep to fixation in a single habitat where such an allele is advantageous. Subsequently, the frequency of some neutral alleles at closely linked loci should be increased in that region, due to genetic hitchhiking. Therefore, the genetic divergence should be greater for neutral loci, which are closely linked to the selected locus than for other neutral loci. This is the rationale for locating outlier loci and then candidate genes that may have been targets of selection, with the various types of markers and techniques described above.

The outlier status of a locus having undergone a selective sweep has been demonstrated among 11 loci that are otherwise undifferentiated between two patches of *Mytilus edulis* (Faure et al. 2008). Vasemägi et al. (2005) identified nine loci in the Atlantic salmon, *Salmo salar* with highly significant deviations from the neutral expectations and proposed them as promising candidate genes for adaptation to different habitats. A similar method was also implemented in the eelgrass *Zostera marina*: using a genome-scan, over 25 anonymous or gene-linked microsatellite loci appeared to be under selection in subtidal and intertidal populations of the eelgrass. Oetjen and Reusch (2007a) identified three outlier loci, one of them being linked to a nodulin gene involved in water transport regulation. This suggested a functional significance in the genetic divergence observed at the outlier locus that may reflect habitat differences.

3.3.1.3 Study of Hybrid Zones and the Speciation Processes

As in terrestrial ecosystems, ample opportunity for secondary contact between differentiated gene pools also exists at sea. This creates hybridisation or tension zones where the two genomes confront each other. These zones are maintained by a balance between an afflux of parental genomes mediated mostly through larval dispersal and the removal of less fit hybrid gene combinations. Because of the demographic characteristics of most marine species, which allows a very fine selective sieving of the various genotypes, one expects that these zones settle precisely where the two differentiated gene pools (e.g. subspecies) display habitat specificities. The combination of habitat specialisation genes (exogenous selection) and under-dominant "speciation" genes (endogenous selection) gives rise to locally coincident genetic clines where all the genes are in linkage disequilibrium and may roughly resemble primary differentiation along an environmental gradient. These clines are the result of the interaction of "congealed" genomes (i.e. genomes that cannot completely remix with each other because of the existence of these "speciation" genes). One of the best studied examples of this is the mosaic zone of the blue mussel *Mytilus edulis* and *M. galloprovincialis* (Bierne et al. 2002, 2003), but it is likely that many other cases have gone unreported or misinterpreted. Recent surveys of *Bathymodiolus* populations using coalescence-based methods over multiple genes both along the mid-Atlantic Ridge and the South East Pacific Rise indicated that such secondary contact with hybridization are frequent occurrences in the deep ocean with old introgression events that cover large portions of the oceanic ridges (Faure et al. 2009). If genome scans are performed along those clines, many markers will co-segregate, yet only a few may really reveal differential adaptation to varying environmental conditions. This should not be underestimated when studying places where many species show genetic clines such as at the mouth of the Baltic Sea (Johannesson and André 2006). Ecological genomics represents a great potential for disentangling the complex interactions of exogenous and endogenous selective forces in the future. Both categories of genetic effects are pertinent to studying the speciation process in the marine realm.

3.3.2 Marine Bio-Invasions: Using Genomic Resources to Study Invasive Species

Biological invasions, tightly linked to human activities, are one of the major threats to marine biodiversity and ecosystem stability. Although one of the main properties of biological invasions is the unpredictable nature of their appearance, introductions of non-indigeneous species have been shown to be highly correlated with commercial shipping routes (Ricciardi and MacIsaac 2000) and aquaculture (Wolff and Reise 2002). For instance, shipping between the USA and Great Britain may have caused the introduction of 20% of the foreign species in British waters. Foreign species may also be capable of taking advantage of coastal water pollution (e.g., the introduced alga *Ulva fascicata* has supplanted the autochthonous species *Ulva pertusa* in polluted environments in Japan; Morand and Briand 1996) or climate change (Stachowicz et al. 2002). In this context, several authors have underlined the usefulness and lack of molecular and genomic data for the study of biological invasions (Holland 2000, Roman and Darling 2007, Sax et al. 2007, Caroll 2008, Darling and Blum 2008). Population genomics and DNA-based approaches may provide important information about biological invasion processes in several ways. For example, using the completely sequenced mitochondrial genome of *Laminaria digitata* (AJ344328; Oudot et al. 2002) and *Pylaiella littoralis* (AJ277126; Oudot-Le Secq et al. 2001), Engel et al. (2008) identified seven polymorphic intergenic regions which were conserved over a range of brown algae species from the Laminariale (*sensu lato*) and to a lesser extent Fucaceae. The sequences of two of these genomic regions were used in a worldwide population survey of circa 500 individuals of the invasive brown alga *Undaria pinnatifida,* which showed that different introduction processes were responsible for the successful introduction of this Japanese kelp in different parts of the world (Voisin et al. 2005).

Darling and Blum (2008) recently advocated the use of genomic resources such as microarrays and quantitative PCR as potentially powerful tools for examining microbial or planktonic diversity. The extent of biological invasions involving these organisms is certainly neglected due to their small size and difficulties with identification. Another important and unresolved question in invasion biology studies is the tempo and mode of evolution of invasive species in their new range. Genome scan approaches based on AFLPs, SNPs or full genome studies can potentially provide tools for analyzing on-going adaptation processes. In the specific case of introduced species, such analyses may be used to examine the possibility for selection on standing genetic variation (Barrett and Schluter 2008).

3.3.3 Uncovering the Genetic Basis of Hybrid Vigour in Aquaculture Populations

Oysters are a highly prized commercial species. Investment into understanding their genetics is of paramount importance for their maintenance as a cash-rich

aquaculture crop. Using the new MPSS (Massive Parallel Signature Sequencing) technology, Hedgecock et al. (2007) studied the genetic and physiological causes of heterosis (hybrid vigor) and its converse, inbreeding depression in the Pacific oyster, *Crassostrea gigas*, the causes of which have remained elusive for nearly a century (e.g. Crow 1998). Explanations proposed for heterosis have included:

- Overdominance (the superiority of heterozygotes at genes affecting fitness traits)
- Dominance (the masking of deleterious recessive mutations in hybrids by dominant alleles inherited from one or the other inbred parent)
- Epistasis (the interaction of alleles at different loci).

In previous investigations, the evidence for heterosis in bivalves was indirect. An observed correlation between a heterozygous marker and fitness-related traits, such as growth was established (e.g. Hedgecock et al. 1995). Also controlled crosses including the study of F2 hybrid populations revealed a high load of deleterious recessive mutations concordant with the dominance hypothesis (Launey and Hedgecock 2001).

The physiological causes of growth heterosis are far less studied than the genetic origins, but it was shown that both larval and adult hybrid oysters have higher feeding rates and efficiencies than their inbred counterparts (Bayne et al. 1999, Pace et al. 2006). The two main aims of the Hedgecock et al. (2007) study were:

- To estimate the number of genes implicated in the hybrid vigour of oyster
- To provide a genome-wide scan for the causes for heterosis.

Gene-expression patterns underlying growth heterosis were analysed in two partially inbred ($f = 0.375$) and two hybrid larval populations produced by a reciprocal cross between the two inbred families. cDNAs were cloned and 4.5 Mb of sequence tags were generated. The sequences contained 23,274 distinct signatures (i.e. short read sequence tags) that were expressed at non-zero levels, and showed a highly positively skewed distribution with median and modal counts of 9.25 million and three transcripts per million, respectively. For nearly half (57%) of these signatures, expression levels were shown to depend on genotype. Results demonstrated that this phenomenon is predominantly non-additive (hybrids deviate from the inbred average as in the over- or underdominance hypotheses), and that overdominance was prevalent in explaining expression patterns in the Pacific oyster as opposed to results reported in maize or *Drosophila* (Swanson-Wagner et al. 2006, Gibson et al. 2004).

The genetic basis of overdominant phenotypes may be due to interactions between *cis*-acting regulatory elements or differences in levels of *trans*-acting factors. As noted by Hedgecock et al. (2007), it should be possible to distinguish between *cis*- and *trans*- regulation of expression levels by contingency tests on the linkage between expression and genotype in the next generation:

- *cis* regulation: the linkage between heterozygosity at the candidate locus and expression level should remain unchanged.
- *trans* regulation: recombination between the candidate locus and *trans*-acting elements (i.e. a possible cause of epistasis) should eliminate or reduce over-expression in heterozygotes.

The likely importance of *trans*-acting elements in regulating many non-additive patterns of gene expression would suggest that growth heterosis may largely result from epistatic gene interactions. Besides the importance of this study in identifying the main causes of hybrid vigour, further analysis suggested ≈350 candidate genes were involved in growth heterosis and exhibited concordant non-additive expression in reciprocal hybrids. This represented only ≈1.5% of the total transcripts and approximately matched the number of genes that regulate the physiology of life span and lipid accumulation in *C. elegans* (Ashrafi et al. 2003). This is clearly only one example of the investigation of heterosis, but given the success using new technologies, additional studies, in different organisms, are certain to follow, particularly driven by the commercial interests of the aquaculture industry.

3.3.4 Gene Polymorphism and Population Adaptation

Sequencing the same gene many times in different individuals of the same species does not correspond to the traditional view of population genetics (Ellegren and Sheldon (2008) and Stinchcombe and Hoekstra (2008)). Indeed it would more generally be viewed as phylogeny or as a sort of bar-coding. However, nucleotide changes in a particular gene (polymorphisms) can explain within- and among-population differences related to the observed phenotypic scope (or within and among species differences) (reviewed in Yang and Bielawski 2002).

For example, Streelman and Kocher (2002) reported a microsatellite polymorphism in the proximal promoter of the prolactin (*prl*) gene in the Nile tilapia (*Oreochromis niloticus*). They demonstrated that distinct microsatellite genotypes were associated with differences in *PRL* expression and that the growth performance of tilapia challenged by various salinities was, at least partly, reflected by changes in growth patterns (assessed by body mass) across a salinity gradient. Also in the sea bass, prolactin gene expression was demonstrated to differ among habitats (Boutet et al. 2007). This suggests that that other regulatory mechanisms (potentially due to polymorphisms) at this locus could be involved in more generalised adaptations rather than purely salinity as described in the case of tilapia. Similarly, Almuly et al. (2008) also reported results that described the role of polymorphisms and minisatellites in the regulation of the growth hormone gene (GH). GH genotypes of one of these loci in seabream (*Sparus aurata*) have been shown to correlate with populations inhabiting distinct environments (open sea vs lagoon) (L. Chaoui and F. Bonhomme, *pers. obs.*). Finally, using the basis of a pedigree-based analysis, Tao and Boulding (2003) reported that a SNP marker located in one intron of

the growth hormone-releasing hormone (*GHRH*) locus explained a significant fraction (≈10%) of the phenotypic variation in early growth rate in the Arctic charr (*Salvelinus alpinus*). They also reported marginally significant results for another SNP marker located in the promoter of the GH gene. Contrasting nucleotide polymorphisms and species divergence in closely-related species at various gene loci may also help in delineating the role of selection at a given locus. Recently, Faure et al. (2007) showed that the second intron of the EF1α gene was under strong purifying selection in the vent mussel *Bathymodiolus* genus whereas it evolves neutrally in other bivalve molluscs. Discriminating between selective sweeps, gene hitchhiking or population expansion was only made possible by the combined analysis of polymorphisms in two distinct species.

The examples highlighted here mainly report polymorphisms in *cis*-regulatory regions rather than coding sequences as the genetic basis of an observed phenotype. As more data is generated from marine species, the utility of cross-species comparisons for identifying non-coding genomic regions potentially involved in regulation of gene expression and phenotypes increases (cf. Lennard Richard et al. 2007). Whilst the case for sequence variation in non-coding regions as drivers for phenotypic change is clear (e.g. Kashi et al. 1997, Britten et al. 2003), understanding if coding verses non-coding variation represents the major source supporting phenotypic variation is a highly debated issue (see Hoekstra and Coyne 2007, Wray 2007). For aquatic organisms data relating ecological issues to polymorphisms in the coding regions of candidate genes have already been reported. Examples have been shown in studies as diverse as understanding of the genetic basis of reproductive isolation within- and among-species (Palumbi 1999, Moy et al. 2008) and adaptation to toxicants (Cohen 2002).

3.4 Expression Studies and Environmental Genomics

Allied to population studies, which concentrate on the analysis of DNA, are the studies of RNA or expressed sequences. Whilst DNA analyses define a population (and indeed in some examples have indicated fitness traits), RNA studies specifically analyse the function and fitness of that population. For example how species adapt to extreme environments and how they cope with change, a particularly critical point to consider given the predicted changes to our climate over the coming years (IPCC 2007). The layering of functional information onto populations is termed "Environmental Genomics". This "omics" off-shoot is a real mix of laboratory based experimentation combined with environmental observations and sampling. The two approaches have to be used in tandem to produce meaningful functional data.

For example, the initial cloning and production of an assay for heat shock protein (HSP70) genes in Antarctic molluscs was conducted at the environmentally unrealistic temperatures of 15ºC. Indeed this was the lowest temperature at which these genes were expressed under laboratory conditions (Clark et al. 2008b). However,

given the success of the assay in the lab, the relative stress levels of environmentally sampled animals could then be examined. The reason being that if no HSP70 expression was detected in these animals, the logical conclusion would be that these genes were inactive under that particular set of circumstances, given the fact that the genes had been characterised and there was a working and accurate assay. In fact this proved not to be the case and indicated that control of these genes was more complex depending on whether the stress applied was short-term acute (experimental) or longer term chronic (environmental) (Clark et al. 2008d, Clark and Peck 2009).

Although this is a single gene example and would not be considered as "genomics" per se, it does demonstrate the complexity of environmental genomics and the requirement to integrate different disciplines: ecology, through physiology to gene expression and genomics. It also serves to highlight, that to date, the majority of environmental stress monitoring in invertebrate species has used the HSP70 gene family (to great effect), with relatively few examples of a genomics approach. Hence this following section will include some single gene examples, where they serve to demonstrate molecular approaches to an environmental question, which can then be expanded to a genomics approach.

3.4.1 Defining Habitat Limits: Biogeography

The question of "which animal inhabits which environment and why?" is fundamental to ecology. Population genetics is often seen as at least a partial answer to this question as this discipline can document DNA differences between populations, and say that population X is not the same as population Y even though they look very similar. The markers used in this type of analysis are generally neutral markers (i.e. do not code for genes) and hence not under selection pressure. Identifying differences in neutral DNA between populations does not answer the fundamental question of why species "chose" to live where they do and what particular adaptations they need to survive in their chosen environment (also see Section 3.2.1.3). In the past we have been able to document morphological, physiological and biochemical environmental adaptations, but with molecular biology, this can be taken to the more detailed scale of the cellular level, the level at which all these phenotypic adaptations are controlled. It is now possible to investigate the transcriptional and proteomic profiles of different populations living under different environmental conditions and manipulate these to investigate the nature of the underlying cellular changes. Such knowledge, apart from providing an answer to the question of habitat "choice", will also enable us to predict how animals will adapt in the face of perturbation, in particular climate change.

This type of work is of great relevant to ectotherms, of which those in the marine environment are prime examples, where one might expect tight linkage between the environmental temperature and species distribution patterns (Somero 2002). To a large extent this field of research has concentrated on inter-tidal species, as these

provide a natural laboratory with species living in clearly defined tidal zones with different stresses and organism thermal limits (Roberts et al. 1997, Tomanek and Somero 2000, Tomanek 2002, Halpin et al. 2002, Tomanek 2005). As stated previously, this work has largely concentrated on monitoring the production of heat shock genes/proteins, particularly the inducible form of HSP70 as these are regulated, not only by temperature but are also generalised stress proteins.

The general findings of this body of research indicates that a species ability to colonise the inter-tidal zone is at least partly dictated by its thermal tolerance. Although a certain amount of temperature acclimation can occur, the most exposed animals (i.e. those furthest up the shoreline) actually have a relatively lower capacity to further adjust their heat shock response than those closer to the inter-tidal region and therefore are the most vulnerable to climate change. This initially appears counter-intuitive, as these are the animals which inhabit the harshest environment and so might be expected to be the most robust in the face of change. This is almost certainly due to cellular energy budgets and protein production, so the requirement for increased HSP production on the high shoreline has to be traded off against other cellular processes. For example reciprocal transplant experiments in inter-tidal mussels resulted in those higher up the shoreline growing more slowly (Hofmann 2005).

Relatively local scale experiments are also mirrored on the larger scale (reviewed in Hofmann 2005), where HSP production is correlated with biogeography. Efforts have concentrated on studying species along latitudinal gradients (Halpin et al. 2002, Osovitz and Hofmann 2005) and species at their edge ranges (Hofmann 2005). This work has recently been expanded using microarray profiling. A 2,496 feature cDNA array was produced for the inter tidal mussel *Mytilus californianus* and probed with RNA from four different populations across a 17º latitudinal gradient along the west coast of North America (Place et al. 2008). Analysis concentrated on genes associated with environmental stress (protein folding, protein degradation and apoptosis) and several showed particularly high levels of expression including the heat shock cognate HSC71, the beta subunit of the proteasome, elf2-α, a stress regulated translation initiation factor and an integral membrane protein involved in the stress response in yeast. These genes potentially provide biomarkers of stress for future work in this organism. Overall, the four location-specific gene expression profiles revealed the complexity of local habitat environment overriding latitude. The expression of the majority of the genes varied significantly as a function of the collection site and provided support for the original hypothesis that the physical response of *M. californianus* to emersion and abiotic factors is population-specific.

In such work involving sampling over latitudinal gradients or global regions, it is important to work either on the same species or closely related congeners, so that the adaptive effects of genetic variation can be clearly demonstrated as being independent of phylogeny. With more genome data becoming available from non-model species, this work is now rapidly progressing from the analysis of single genes to thousands with the use of gene chips. Thus producing a more holistic genome level analysis of the trade-offs involved in habitat selection.

3.4.2 Microarrays: Identification of Biochemical Pathways Involved in Adaptation

This section is concerned with the identification of more complex pathways and how these change in relation to perturbation or natural environmental cycling. This specifically refers to gene chips. Work in the field is relatively limited so far, as considerable specialised molecular input is required: library production, generation of gene chips, hybridization of the chips and analysis. None of this is trivial and requires specialist skills only available in relatively few laboratories. It is possible to use gene chips produced for a model organism to ask questions in a non-model species, cf. Hogstrand et al. (2002) examining the response to zinc exposure in rainbow trout (*Oncorhynchus mykiss*) using high density spotted arrays from the Japanese pufferfish (*Takifugu rupripes*). Whilst this technique has been proved to work effectively (reviewed in Buckley 2007, Kassahn 2008), the detected magnitude of fold difference in gene expression decreases across phylogenetic distance (Renn et al. 2004) and this has to be considered as part of the experimental design. Also, another problem of working on non-model species, is a restricted ability to identify genes (also see Section 3.2.1.1) and more importantly putatively assign function, as this is all based on sequence similarity with database entries, most of which are vertebrates, in particular, mammals. As a result, gene chip analyses in non-model species has been dominated by fish species, in particular those involved in aquaculture and ecotoxicology (Wenne et al. 2007, Daib et al. 2008).

However, examples of environmentally-biased gene chip experiments are beginning to appear. These include the heat stress response of the inter-tidal porcelain crab (*Petrolisthes cinctipes*) (Teranishi and Stillman 2007), cold stress in the common carp (*Cyprinus carpio* L.) (Gracey et al. 2004), daily fluctuating temperatures in the annual killifish (*Austrofundulus limnaeus*) (Podrabsky and Somero 2004) and the purely environmental sampling example of *M. californianus* detailed above (Place et al. 2008). Whilst these experiments document gene changes associated with changing conditions and can for example, highlight pathways most readily affected cf. protein folding, protein degradation and protein synthesis with heat stress (Teranishi and Stillman 2006), the gene lists are long. Detailed analysis of all genes is simply not logistically possible and these experiments essentially provide candidate genes for future studies, e.g. stearoyl-CoA desaturase in cold adaptation of the carp (Gracey et al. 2004). However without the initial broad brush stroke approach of screening thousands of genes in the first place, it would not have been possible to identify such candidates, which ultimately may help us understand the fundamental nature of environmental adaptation.

3.4.3 Genome Plasticity and Seasonal Variation

One point to remember is that expression of the genome is not static. Unless a long-term study is undertaken, with many different samplings of the population, then

whatever measure is taken, is a snapshot at that particular point in time and reflects a particular set of environmental variables. Interpretation of such data does require an extensive background knowledge of the organism. For example events such as spawning or the effects of seasonal temperature variations (thermal history) and production of season-specific proteins (cf. antifreezes in winter) (Buckley et al. 2001, Tomanek 2002, Enevoldsen et al. 2003, Jin and DeVries 2006) can significantly alter gene expression patterns and will obviously bias any profiling work. This variation in gene expression according to organism history and environmental signals is termed plasticity and represents a range or expression window in which genes (and as a consequence species) can operate effectively and adapt.

This is another area where studies are largely confined to the individual gene level and even so, such data is limited, but it is emerging that this plasticity of response is highly gene dependant. In a comprehensive molecular study into how thermal history influences muscle development in fish, two temperature responsive genes were identified; Myogenin and FoxK1 (Fernandes et al. 2006, 2007), but at least an equal number of genes studied were unaffected by heat treatment (Mackenzie 2006). Most of the work on plasticity, to date, has concentrated on terrestrial species, as these generally experience far wider temperature ranges than marine species (cf. Deere and Chown 2006) or the standard model eukaryote, yeast (Stern et al. 2007). But, given predicted increased seawater temperatures under climate change senarios, this is clearly an emerging area of importance in the marine domain (Peck et al. 2009). This understanding of which genes are more "adaptable" will progress as analyses are scaled up to the genomic level using gene chips and new sequencing technologies.

3.4.4 Adaptation to Extreme Environments

The section on Population Genomics has presented several examples of where DNA polymorphisms effect gene expression (Section 3.3.4) and hence imply adaptation and specialisation. However, nowhere is adaptation observed as strongly, as in extreme environments. These require far more large-scale genomic changes to adapt to what is often a very hostile environment. This is an area, not just of academic inquiry, but also of potential commercial interest, with investigations ranging from characterisation of novel proteins to identifying enzyme variants that work "better" under different circumstances that could be used in, for example, food processing or detergent production (Clark et al. 2004, Peck et al. 2005). Whilst many types of extreme environments exist, this section will concentrate on adaptations to two of the most commonly described natural environments: hydrothermal vents and the Polar regions. The aim is to provide an overview of research into two opposite extremes: the hot and the cold, followed by the relatively "new" challenge of anthropomorphic change (toxicology/pollution). Adaptations to hydrothermal vents and the Polar regions have taken place over thousands/millions of years, whilst organisms have had to adjust/adapt to pollutants over periods of only tens to hundreds of years.

3.4.4.1 Hydrothermal Vents

Hydrothermal vents are characterized by very specific physical and chemical properties, such as elevated pressure (up to 420 atm), high and abruptly changing temperature (from 2–4°C to 400°C) that can occur both spatially (within tens of cm: Piccino et al. 2004) and temporally (10–50°C within a minute: Le Bris et al. 2005), high levels of sulphide and/or methane that can fuel endosymbioses (Childress and Fisher 1992) and chemical toxicity (heavy metals and radionuclides: Cherry et al. 1992, Luther et al. 2001) and the complete absence of light. However, numerous living organisms such as shrimps, clams, mussels, giant tubeworms, crabs and fishes have been discovered in those environments. These organisms have developed different adaptive strategies, which ensure their exploitation of the hydrothermal vent fluid. The most studied adaptations are:

- Symbiosis as a response to the absence of photosynthesis that has led to a food chain based on primary production of energy and organic molecules by chimiolithoautotrophic bacteria (Minic and Herve 2004, Stewart and Cavanaugh 2006, Duperron et al. 2007).
- Adaptation to high temperatures (Gaill et al. 1995, Sicot et al. 2000)
- Adaptation to toxicants (Company et al. 2004)
- Adaptation to hypoxia/anoxia (Hourdez and Weber 2005).

Studies of adaptation to high temperatures have essentially concentrated on bacteria. These show some general features, such as an increase in charged amino acids, proline residues and replacement of some lysines by arginine, which increases hydrogen bonds (Kumar et al. 2000, Nishio et al. 2003, Robinson et al. 2006). Similar patterns are observed in eukaryote species such as *A. pompejana,* which lives in a hotter part of this environment than *P. grasslei*. Analyses of amino-acid composition have revealed a significant increase in positively charged residues together with an increase in protein hydrophobicity (Jollivet et al. in prep). Some specific proteins have been carefully studied in terms of their thermostability, for example collagen (Sicot et al. 2000), mitochondrial (Dahlhoff et al. 1991) and cytoplasmic (Jollivet et al. 1995) respiratory chain proteins and haemoglobins (see for review Hourdez and Weber 2005). In the specific case of collagen, proline hydroxylation seems to play a crucial role in enhancing molecular thermostability and therefore it is suggested that post-translational processes are also key factors in adaptation to high thermal regimes.

Regarding the adaptation processes linked to the presence of toxicants and hypoxia/anoxia, very few studies have been conducted and most are concerned with the mussel genus and the effect of heavy metals or oxidative stress on enzymatic activities (Company et al. 2008) or specific gene expression such as metallothioneins (Hardivillier et al. 2006). Pruski and Dixon (2003) also showed a positive correlation between the levels of DNA strand breakage and HSP70 protein expression in response to decompression and to oxidative stress. The molecular analyses of the giant HBL-Haemoglobin of deep-sea vent annelids are a good example typifying

how this fauna adapted to the highly poisoning sulphide vent environment. In this particular case, the respiratory pigment is able to reversibly bind sulphide onto two distinct cysteine residues of the A2 and B2 globins (Bailly et al. 2002). Bailly et al. (2003) demonstrated that such an ability was probably lost by positive Darwinian selection during the course of evolution in modern annelids living in non-reducing habitats.

There is a distinct lack of sequence data from hydrothermal species. However, a recent effort has concentrated on the sequencing of the transcriptomes of *Bathymodiolus azoricus*, *Paralvinella grasslei, Alvinella pompejana* and *Riftia pachyptyla*) (Sanchez et al. 2007, Tanguy et al. 2008, Alvinella Consortium Project). The first cDNA microarrays have also been developed for *B. azoricus* and *P. grasslei* and the aim is to use these as models in order to study the response of these organisms to temperature challenges, heavy metal effects and symbiosis. A recent study has already demonstrated that deep-sea vent mussels can endure a global depression in gene expression associated with short-term exposures (30–120 min) to temperatures higher than 20°C, suggesting that these animals are more likely adapted to cold temperatures (Boutet et al. 2008).

3.4.4.2 Polar Environments

Living in the cold obviously has resulted in specialist genetic adaptations, of which there are some classic single gene investigations:

- Plasma antifreeze: this was first discovered in Antarctic fish (DeVries 1970), but now recognised as a standard adaptation to the cold marine environment (DeVries 1982).
- Specific protein modifications resulting in increases in molecule flexibility, essential for efficient functioning in the cold (Fields and Somero 1998, Detrich et al. 1989, 1992, Fields et al. 2002,Römisch et al. 2003).
- Deletion of haemoglobin genes and the ability to produce functional erythrocytes in the Channichthyidae (icefish) (Moylan and Sidell 2000, di Prisco et al. 2002).
- Lack of the classical heat shock response (upregulation of the inducible form of HSP70) in a number of fish and invertebrate species (Hofmann et al. 2000, Clark et al. 2008a, c).

This latter "adaptation" has been recently investigated via a pilot microarray analysis hybridizing RNA from the Antarctic fish *Trematomus bernacchii* onto an array produced for the eurythermal goby fish *Gillichthys mirabilis* (Buckley and Somero 2009). In this study they showed that the Antarctic fish, although not displaying the classical heat shock response, did indeed show enhanced expression of many genes associated with central aspects of the evolutionary conserved cellular stress response. Whilst this experiment used a heterologous array approach, organism-specific arrays are now becoming available for some polar terrestrial insects (cf. Purac et al. 2008) and it is expected that this will soon also be the case with polar marine species.

In general, with the exception of antifreezes, most work on cold adaptation of fish and invertebrates has been carried out on Antarctic species as these have been effectively geographically isolated in a constantly cold environment for around 25–15 Ma BP and comprise a high percentage of endemic species (reviewed in Clarke and Johnston 1996). In contrast the Arctic marine benthic fauna comprises a relatively young assemblage characterised by species from either the Pacific or Atlantic with notably few endemics (Dunton 1992). Hence comparisons of Polar species can provide important information to separate phylogeny from cold adaptation (cf. Verde et al. 2007).

Intimately linked to the subject of Polar environments is that of climate change. Scientific opinion now largely supports the view of man-induced climate change with a number of predicted scenarios for our environment in the future (IPCC 2007) and the Polar regions are where warming is happening more rapidly than most other places on the planet. Oceanic temperatures are predicted to rise by 2°C over the next 100 years (Murphy and Mitchell 1995) faster than in any period over the past million years or on record over the last Pleistocene glacial cycle (Zachos et al. 2001). But along the Antarctic Peninsula regional climate change has been rapid with temperature rises in the Bellingshausen surface ocean sea of 1°C in 50 years (Meredith and King 2005). But the situation is more complex and the effect of CO_2 emissions is not only linked to temperature, but also ocean acidification and pH changes. In the 250 years since the onset of the industrial revolution, atmospheric CO_2 levels have risen from 280 to 381 ppm (Canadell et al. 2007) and are still rising rapidly. CO_2 is taken up by the ocean and forms carbonic acid, thus reducing ocean pH, and decreasing the saturation state (= increasing solubility) of calcium carbonate. Ocean pH has fallen from an average 8.16–8.05 (Caldeira and Wicket 2003) and models predict that pH at the ocean surface will continue to fall by an estimated 0.2–0.4 units by the year 2100 (Caldeira and Wicket 2003, 2005, Royal Society 2005, Cao et al. 2007). These predicted changes in ocean pH are greater, and far more rapid, than any experienced in the past 300 million years. Again, the Polar Regions will be particularly affected as the Southern Ocean has among the lowest present-day $CaCO_3$ saturation rate of any ocean region, and will therefore be among the first to become undersaturated (Orr et al. 2005).

So clearly there is an urgent requirement to determine the effect of increased seawater temperatures and acidification on marine life. Not all organisms are affected equally and while all those studied to date have been shown to be affected, it is clear that some will adapt and survive (Dupont et al. 2008). We currently know almost nothing about the genetic mechanisms and genomic basis of this phenoptypic plasticity. It is with this requirement that Polar species, in particular those in the Antarctic are of great use, for example in the development of biomarkers for climate change. The advantage of working on Antarctic species is that they are very thermally sensitive (Peck et al. 2004) and their response is not confounded by pollution, which affects the rest of the planet. Therefore Antarctic organisms offer us the cleanest signals for the effects of climate change and should be regarded as environmental sentinels. Integrated genomic and functional studies on Polar organisms are in their infancy (Clark et al. 2004, Peck et al. 2005), but such studies are now

being specifically targeted to the problems associated with climate change and their effects on functional biodiversity.

3.4.4.3 Ecotoxicology Monitoring

Marine ecosystems, particularly coasts, estuaries and coral reefs, are currently experiencing major crises worldwide as environmental change places significant physiological stress (metals, pesticides, chemicals but also parasites) on organisms. Understanding ecosystem resilience and predicting the impact of such environmental stresses on marine organisms depends on knowing the physiological status and plasticity of these organisms in these ecosystems. Traditional ecotoxicology in marine species involves the study of "biological biomarkers" such as enzyme activities and stress protein quantification, immunological parameters evaluation and life trait (growth, reproductive stage) evaluation. Recent progresses in genomic and proteomic techniques have lead to the emergence of new approaches in ecotoxicology called "ecotoxicogenomics". This is defined as the study of gene and protein expression in non-model organisms that is important in understanding responses to environmental toxicant exposures. Ecotoxicogenomics is a technology that has been made possible mainly by the advent of DNA microarray analyses. Microarrays have aided our understanding of relationships between global gene expression profiles, physiological states of an organism, and traditional toxic endpoints (Irwin et al. 2004, Lee et al. 2003).

Genomics provides a detailed view of physiological diversity and function, and thus a mechanistic insight into how organisms respond to environmental stress.

Genomic approaches in ecotoxicology will:

- Improve our understanding of toxicant/stress/infection mechanisms.
- Develop a physiological perspective on the environmental facilitation of toxicant/stress/infection within organisms
- Produce tools to predict the spread of toxicant/stress/infection, leading to better assessment and prediction of environmental health.

In particular, the application of ecotoxicogenomics will extrapolate from experimental in vitro to in vivo systems and across the species barrier. It will aid in the understanding of specific molecular events underlying the mode of action of toxicants, and therefore these can be developed as biomarkers to identify exposure to environmental stressors.

Gene expression profiling can be used to show that the specific genes repressed or induced upon exposure to a toxic stress vary depending on the cell type and the type of toxicants to which the cells were exposed (Troester et al. 2004). The major challenge of ecotoxicogenomic approaches will be to take into account intrinsic sources of variability in gene expression levels due to different physiological states,

age, sex, and genetic polymorphisms in natural populations. Several specific aspects could be then considered with more attention:

- Identification of conserved genes that are up-regulated in response to toxicant exposure (according to exposure time and toxicant concentration).
- Determination of how these gene expression profiles can be used to diagnose stressors.
- Identification of genes that are most informative to incorporate into more specific stress gene arrays for monitoring purposes.

In marine species, a particular effort has been focused on fishes (Sheader et al. 2006) and is underway in marine mollusc species. The toxicogenomic approach will certainly present new opportunities to improve understanding of the molecular mechanisms underlying toxic responses to environmental contaminants (Bradley and Theodorakis 2002, Moore 2001).

3.5 Summary and Future Issues

This chapter summarises the current molecular tools available to the marine biologist for studying ecological questions. What is clear is that whilst new techniques are continually being brought in (e.g. MPSS and 454 sequencing), their utility is reliant upon a robust understanding of the underlying ecology and physiology of the species under study. This often means identifying populations of the same species with clear phenotypic contrasts, such that the contrast between phenotype and genotype reveals a set of candidate genes for the further functional and genetic analyses of a biological process. Equally adopting a rigorous experimental design using known physiological knowledge will obviate many data interpretation problems. Hence a holistic overview of ecological adaptation can only be obtained by a combination of approaches. These are not restricted to genomics and must incorporate ecology and physiology; indeed these are the disciplines that dictate the questions, populations and experimental approaches used, with the molecular biology as an additional tool in the armoury.

As regards future perspectives, undoubtedly the use of MPSS and 454 will increase, as these techniques rapidly generate large amounts of genome data on non-model species. Thus the traditional view of a non-model species, as a gene-poor resource will change dramatically over the next few years. Each community should concentrate on generating a number of “new model species” that are particularly useful for answering ecological questions in their specific domain, be it aquaculture, hydrothermal vents or climate change and polar species. These species should originate from a range of taxa across different feeding guilds, so that understanding adaptation to a particular set of conditions is not limited to a single species. Indeed this should ideally stretch from microbes through to higher predators, producing a real gene to ecosystem understanding of our marine environment.

References

Akey JM, Zhang G, Zhang K, Jin L, Shriver MD (2002) Interrogating a high-density SNP map for signatures of natural selection. Genome Res 12:1805–1814

Albert V, Jonsson B, Bernatchez L (2006) Natural hybrids in Atlantic eels (*Anguilla anguilla, A. rostrata*): evidence for successful reproduction and fluctuating abundance in space and time. Mol Ecol 15:1903–1916

Almuly R, Skopal T, Funkenstein B (2008) Regulatory regions in the promoter and first intron of Sparus aurata growth hormone gene: Repression of gene activity by a polymorphic minisatellite. Comp Biochem Physiol D 3:43–50

Ashrafi K, Chang FY, Watts JL, Fraser AG, Kamath RS, Ahringer, Ruvkun G (2003) Genome-wide RNAi analysis of *Caenorhabditis elegans* fat regulatory genes. Nature 421:268–272

Backström N, Ovarström A, Gustafsson L, Hellegren H (2006) Levels of linkage disequilibrium in a wild bird population. Biol Lett 2:435–438

Bailly X, Jollivet D, Vanin S, Deutsch J, Zal Z, Lallier FH, Toulmond A (2002) Evolution of the sulfide-binding function within the globin multigenic family of the deep-sea hydrothermal vent tubeworm *Riftia pachyptila*. Mol Biol Evol 19:1421–1433

Bailly X, Leroy R, Carney S, Collin O, Zal F, Toulmond A, Jollivet D (2003) The loss of the hemoglobin H_2S-binding function reveals molecular adaptation driven by Darwinian positive selection in annelids from sulfide-free habitats. Proc Natl Acad Sci USA 100: 5885–5890

Balloux F, Amos W, Coulson T (2004) Does heterozygosity estimate inbreeding in real populations?. Mol Ecol 13:3021–3031

Baranski M, Loughnan S, Austin CM et al. (2006) A microsatellite linkage map of the blacklip abalone, *Haliotis rubra*. Anim Genet 37:563–570

Barrett RD, Schluter D (2008) Adaptation from standing genetic variation. Trends Ecol Evol 23:38–44

Barretto FS, McCartney MA (2008) Extraordinary AFLP fingerprint similarity despite strong assortative mating between reef fish color morphospecies. Evolution 62:226–233

Baus E, Darrock DJ, Bruford MW (2005) Gene-flow patterns in Atlantic and Mediterranean populations of the Lusitanian sea star *Asterina gibbosa*. Mol Ecol 14:3373–3382

Bayne BL, Hedgecock D, McGoldrick D, Rees R (1999) Feeding behaviour and metabolic efficiency contribute to may growth heterosis in Pacific oysters [*Crassostrea gigas* (Thunberg)]. J Exp Mar Biol Ecol 233:115–130

Beaumont MA, Nichols RA (1996) Evaluating loci for use in the genetic analysis of population structure. Proc R Soc Lond B 263:1619–1626

Bensch S, Helbig AJ, Salomon M, Seibold I (2002) Amplified fragment length polymorphism analysis identifies hybrids between two subspecies of warblers. Mol Ecol 11:473– 481

Beraldi D, McRae AF, Gratten J et al (2006) Development of a linkage map and mapping of phenotypic polymorphisms in a free-living population of Soay sheep (*Ovis aries*). Genetics 173:1521–1537

Bierne N, Borsa P, Daguin C, Jollivet D, Viard F, Bonhomme F, David P (2003) Introgression patterns in the mosaic hybrid zone between *Mytilus edulis* and *M. galloprovincialis*. Mol Ecol 12:447–462

Bierne N, David P, Langlade A, Bonhomme F (2002) Can habitat specialisation maintain a mosaic hybrid zone in marine bivalves?. Mar Ecol Progr Ser 245:157–170

Black WC, Baer CF, Antolin MF, DuTeau NM (2001) Population genomics: genome-wide sampling of insect populations. Annu Rev Entomol 46:441–469

Boguski MS, Lowe TMJ, Tolstoshev CM (1993) dbEST – database for 'expressed sequence tags'. Nat Genet 4:332–333

Bonin A, Taberlet P, Miaud C, Pompanon F (2006) Explorative genome scan to detect candidate loci for adaptation along a gradient of altitude in the common frog (*Rana temporaria*). Mol Biol Evol 23:773–783

Boutet I, Ky CL, Bonhomme F (2006) A transcriptomic approach of salinity response in the euryhaline teleost, *Dicentrarchus labrax*. Gene 379:40–50

Boutet I, Nebel C, De Lorgeril J, Guinand B (2007) Molecular characterisation and extrapituitary prolactin expression in the European sea bass *Dicentrarchus labrax* under salinity stress. Comp Biochem Physiol D 2:74–83

Boutet I, Tanguy A, Le Guen D, Piccino P, Hourdez S, Ravaux J, Shillito B, Legendre P, Jollivet D (2008) Global depression in gene expression as a response to rapid changes of temperature in the hydrothermal vent mussel *Bathymodiolus azoricus*. PLOS Biol 276:3071–3079

Bradley B, Theodorakis C (2002) The post-genomic era and ecotoxicology. Ecotoxicology 11:7–9

Brenner S, Johnson M, Bridgham J, Golda G, Lloyd DH, Johnson D, Luo S, McCurdy S, Foy M, Ewan M et al. (2000a) Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. Nat Biotech 18:630–634

Britten RJ, Rowen L, Williams J, Cameron RA (2003) Majority of divergence between closely related DNA samples is due to indels. Proc Natl Acad Sci USA 100:4661–4665

Buckley BA (2007) Comparative environmental genomics in non-model species: using heterologous hybridisation to DNA-based arrays. J Exp Biol 210:1602–1606

Buckley BA, Owen M-E, Hofmann GE (2001) Adjusting the thermostat: the threshold induction temperature for the heat-shock response in intertidal mussels (genus *Mytilus*) changes as a function of thermal history. J Exp Biol 204:3571–3579

Buckley BA, Somero GN (2009) cDNA analysis reveals the capacity of the cold adapted Antarctic fish *Trematomus bernacchii* to alter gene expression in response to heat stress. Polar Biol 32:403–415

Caldeira K, Wicket ME (2003) Oceanography: anthropogenic carbon and ocean pH. Nature 425:365

Caldeira K, Wicket ME (2005) Ocean model prediction of chemistry changes from carbon dioxide emission to the atmosphere and ocean. Geophy Res Lett 110:C09S04. doi:10.1029/2004JC002671

Company R, Serafim A, Cosson RP, Fiala-Medioni A, Camus L, Colaco A, Serrao-Santos R, Bebianno MJ (2008) Antioxidant biochemical responses to long-term copper exposure in Bathymodiolus azoricus from Menez-Gwen hydrothermal vent. Sci Total Environ 389: 407–417

Campbell D, Bernatchez L (2004) Generic scan using AFLP markers genes as a means to assess the role of directional selection in the divergence of sympatric whitefish ecotypes. Mol Biol Evol 21:945–956

Campbell D, Duchesne P, Bernatchez L (2003) AFLP utility for population assignment studies: analytical investigation and empirical comparison with microsatellites. Mol Ecol 12:1979–1991

Campbell NR, Narum SR (2008) Identification of novel SNPs in Chinook salmon and variation among life history types. Trans Am Fish Soc 137:96–106

Canadell JG, Le Quere C, Raupach MR, Field CB, Buitenhuis ET, Ciais P, Conway TJ, Gillett NP, Houghton RA, Marland G (2007) Contributions to accelerating atmospheric CO_2 growth from economic activity, carbon intensity, and efficiency of natural sinks. Proc Natl Acad Sci USA doi:10.1073/pnas.0702737104

Cao L, Caldeira K, Jain AK (2007) Effects of carbon dioxide and climate change on ocean acidification and carbonate mineral saturation. Geophys Res Lett 34:L05607. doi:10.1029/2006GL028605

Carninci P, Shibata Y, Hayatsu N, Sugahara Y, Shibata K, Itoh M, Konno H, Okazaki Y, Muramatsu M (2002) Normalization and subtraction of cap-trapper-selected cDNAs to prepare full-length cDNA libraries for rapid discovery of new genes. Genome Res 10:1617–1630

Caroll SP (2008) Facing change: forms and foundations of contemporary adaptation to biotic invasions. Mol Ecol 17:361–372

Chatziplis D, Batargias C, Tsigenopoulos CS, Magoulas A, Kollias S, Kotoulas G, Volckaert FAM, Haley CS (2007) Mapping quantitative trait loci in European sea bass (*Dicentrarchus labrax*): The BASSMAP pilot study. Aquaculture 272:S172–S182

Cherry R, Desbruyeres D, Heyraud M, Nolan C (1992) High levels of natural radioactivity in hydrothermal vent polychaetes. C R Acad Sci Paris ser III 315:21–26

Childress JJ, Fisher CR (1992) The biology of hydrothermal vent animals: physiology, biochemistry, and autotrophic symbioses. Oceanogr Mar Biol 30:337–441

Chini V, Rimoldi S, Terova G, Saroglia M, Rossi F, Bernardini G, Gornati R (2006) EST-based identification of genes expressed in the liver of adult seabass (*Dicentrarchus labrax*, L.). Gene 376:102–106

Chistiakhov DA, Hellemans B, Haley CS, Law AS, Tsigenopoulos CS, Kotoulas G, Bertotto D, Libertini A, Volckaert FAM (2005) A microsatellite linkage map of the European sea bass *Dicentrarchus labrax* L. Genetics 170:1821–1826

Clark MS, Clarke A, Cockell CS, Convey P, Detrich IIIHW, Fraser KPP, Johnston I, Methe B, Murray AE, Peck LS, Romisch K, Rogers A (2004) Antarctic genomics. Comp Func Genom 5:230–238

Clark MS, Fraser KPP, Burns G, Peck LS (2008a) The HSP70 heat shock response in the Antarctic fish *Harpagifer antarctic*us. Polar Biol 31:171–180

Clark MS, Fraser KPP, Peck LS (2008b) Antarctic marine molluscs do have an HSP70 heat shock response. Cell Stress Chaperones 13:39–49

Clark MS, Fraser KPP, Peck LS (2008c) Lack of an HSP70 heat shock response in two Antarctic marine invertebrates. Polar Biol 31:1059–1065

Clark MS, Geissler P, Waller C, Fraser KPP, Barnes DKA, Peck LS (2008d) Low heat shock thresholds inwild Antarctic inter-tidal limpets (*Nacella concinna*). Cell Stress Chaperones 13:51–58

Clark MS, Peck LS (2009) Triggers of the HSP70 stress response: environmental responses and laboratory manipulation in an Antarctic marine invertebrate (*Nacella concinna*). Cell Stress Chaperones (in press)

Clarke A, Johnston IA (1996) Evolution and adaptive radiation of Antarctic fishes. Trends Ecol Evol 11:212–218

Cohen S (2002) Strong positive selection and habitat-specific amino acid substitution patterns in Mhc from an estuarine fish under intense pollution stress. Mol Biol Evol 19:1870–1880

Colosimo PF, Hosemann KE, Balabhadra S, Villarreal G Jr, Dickson M, Grimwood J et al (2005) Widespread parallel evolution in sticklebacks by repeated fixation of ectodysplasin alleles. Science 307:1928–1933

Comai L, Young K, Till BJ et al. (2004) Efficient discovery of DNA polymorphisms in natural populations by ecotilling. Plant J 37:778–786

Company R, Serafim A, Bebianno MJ, Cosson R, Shillito B, Fiala-Médioni A (2004) Effect of cadmium, copper and mercury on antioxidant enzyme activities and lipid peroxidation in the gills of the hydrothermal vent mussel *Bathymodiolus azoricus*. Mar Environ Res 58:377–381

Crawford DL, Oleksiak MJ (2007) The biological importance of measuring individual variation. J Exp Biol 210:1613–1621

Cresko WA, Amores A, Wilson C, Murphy J, Currey M, Philips P, Bell MA, Kimmel CB, Postlewaith JH (2004) Parallel genetic basis for repeated evolution of armorloss in Alaskan threespine stickleback populations. Proc Natl Acad Sci USA 101:6050–6055

Crow JF (1998) 90 years: the beginning of hybrid maize. Genetics 148:923–928

Dahlhoff E, O'Brien J, Somero GN, Vetter RD (1991) Temperature effects on mitochondria from hydrothermal vent invertebrates: evidence for adaptation to elevated and variable habitat temperatures. Physiol Zool 64:1490–1508

Daib AM, Williams TD, Sabine VS, Chipman JK, George SG (2008) The GENIPOL European flounder *Platichthys flesus* L. toxicogenomics microarray: application for investigation of the response to furunculosis vaccination. Fish Biol 72:2154–2169

Darling JA, Blum MJ (2008) DNA-based methods for monitoring invasive species: a review and prospectus. Biol Invasions 9:751–765

Dasmahapatra KK, Lacy RC, Amos W (2008) Estimating levels of inbreeding using AFLP markers. Heredity 100:286–295

DeVries AL (1970) Freezing resistance in Antarctic fishes. In: Holdgate MW (ed) Antarctic biology. Academic Press, New York
DeVries AL (1982) Biological antifreeze agents in coldwater fishes. Comp Biochem Physiol 73A:627–640
DeWoody JA, Avise JC (2000) Microsatellite variation in marine, freshwater and anadromous fishes compared with other animals. J Fish Biol 56:461–473
DeWoody YD, DeWoody JA (2005) On the estimation of genome-wide heterozygosity using molecular markers. J Hered 96:85–88
DeWoody JA, Nason JD, Hipkins VD (2006) Mitigating scoring errors in microsatellite data from wild populations. Mol Ecol Notes 6:951–957
Deere JA, Chown SL (2006) Testing the beneficial acclimation hypothesis and its alternatives for locomotor performance. Am Nat 5:630–644
Degnan BM, Degnan SM, Fentenany G, Morse DE (1997) A Mox homeobox gene in the gastropod mollusc Haliotis rufescens is differentially expressed during larval morphogenesis and metamorphosis. FEBS Lett 411:119–122
Detrich HWIII, Fitzgerald TJ, Dinsmore JH, Marchese-Ragona SP (1992) Brain and egg tubulins from Antarctic fishes are functionally and structurally distinct. J Biol Chem 267:18766–18775
Detrich HWIII, Johnson KA, Marchese-Ragona SP (1989) Polymerization of Antarctic fish tubulins at low temperatures: energetic aspects. Biochem 28:10085–10093
di Prisco G, Cocca E, Parker S, Detrich HW III (2002) Tracking the evolutionary loss of hemoglobin expression by the white blooded Antarctic icefishes. Gene 295:185–191
Douglas SE, Knickle LC, Kimball J, Reith ME (2007) Comprehensive EST analysis of Atlantic halibut (*Hippoglossus hippoglossus*), a commercially relevant aquaculture species. BMC Genom doi:10.1186/1471-2164-8-144
Dunton K (1992) Arctic biogeography: the paradox of the marine benthic fauna and flora. Trends Ecol Evol 7:183–189
Duperron S, Sibuet M, MacGregor BJ, Kuypers MMM, Fisher CR, Dubilier N (2007) Diversity, relative abundance and metabolic potential of bacterial endosymbionts in three *Bathymodiolus* mussel species from cold seeps in the Gulf of Mexico. Environ Microbiol 9:1423–1438
Dupont S, Havenhand J, Thorndyke W, Peck LS, Thorndyke M (2008) CO_2-driven ocean acidification radically affects larval survival and development in the brittlestar *Ophiothrix fragilis*. Mar Ecol Prog Ser 373:285–294
Ellegren H (2004) Microsatellites: simple sequences with complex evolution. Nat Rev Genet 5:435–445
Ellegren H (2008) Sequencing goes 454 and takes large-scale genomics into the wild. Mol Ecol 17:1629–1631
Ellegren H, Sheldon BC (2008) Genetic basis of fitness differences in natural populations. Nature 452:169–175
Ellis JR, Burke JM (2007) EST-SSRs as a resource for population genetic analyses. Heredity 99:125–132
Enevoldsen LT, Heiner I, DeVries AL, Steffensen JF (2003) Does fish from the Disko Bay area of Greenland possess antifreeze proteins during the summer?. Polar Biol 26:365–370
Engel C, Billard E, Voisin M, Viard F (2008) Conservation and polymorphism of mitochondrial intergenic sequences in brown algae. Eur J Phycol 43:195–205
Erickson DL, Fenster CB, Stenoien HK, Price D (2004) Quantitative trait locus analyses and the study of evolutionary process. Mol Ecol 13:2505–2522
Faure B, Bierne N, Tanguy A, Bonhomme F, Jollivet D (2007) Evidence for a slightly deleterious effect of intron polymorphisms at the EF1a gene in the deep-sea hydrothermal vent bivalve *Bathymodiolus*. Gene 406:99–107
Faure M, David P, Bonhomme F, Bierne N (2008) Genetic hitchhiking in a subdivided population of *Mytilus edulis*. BMC Evol Biol 8:164. doi:10.1186/1471-2148-8-164
Faure B, Jollivet D, Tanguy A, Bonhomme F, Bierne N (2009) Secondary contact zone in the deep sea: a multi-locus analysis of divergence and gene flow between two closely-related species of *Bathymodiolus*. Genetics 4:e6485

Feder ME (2007) Evolvability of physiological and biochemical traits: evolutionary mechanisms including and beyond single-nucleotide mutation. J Exp Biol 310:1653–1660

Feder ME, Mitchell-Olds T (2003) Evolutionary and ecological functional genomics. Nat Rev Genet 4:649–655

Fernandes JMO, MacKenzie MG, Kinghorn JR, Johnston IA (2007) FoxK1 splice variants show developmental stage-specific plasticity of expression with temperature in the tiger pufferfish. J Exp Biol 210:3461–3472

Fernandes JMO, MacKenzie MG, Wright PA, Steele SL, Suzuki Y, Kinghorn JR, Johnston IA (2006) Myogenin in model pufferfish species: comparative genomic analysis and thermal plasticity of expression during early development. Comp Biochem Physiol D 1:35–45

Fields PA, Kim Y-S, Carpenter JF, Somero GN (2002) Temperature adaptation in *Gillichthys* (Teleost: Gobiidae) A4-lactate dehydrogenases: identical primary structures produce subtly different conformations. J Exp Biol 205:1293–1303

Fields PA, Somero GN (1998) Hot spots in cold adaptation: localised increases in conformational flexibility in lactate dehydrogenase A4 orthologs of Antarctic Notothenioid fishes. Proc Natl Acad Sci USA 95:11476–11481

Flannery BG, Wenburg JK, Gharrett AJ (2007) Variation of amplified fragment length polymorphisms in yukon river chum salmon: population structure and application to mixed-stock analysis. Trans Am Fish Soc 136:911–925

Ford MJ (2002) Applications of selective neutrality tests to molecular ecology. Mol Ecol 11: 1245–1262

Franch R, Louro B, Tsalavouta M, Chatziplis D, Tsigenopoulos CS, Sarropoulou E, Antonello J, Magoulas A, Mylonas CC, Babbucci M, Patarnello T, Power DM, Kotoulas G, Bargelloni L (2006) A genetic linkage map of the hermaphrodite teleost fish *Sparus aurata* L.. Genetics 174:851–861

Fuchs Y, Douek J, Rinkevich B, Ben-Shlomo R (2006) Gene diversity and mode of reproduction in the brooded larvae of the coral *Heteroxenia fuscescens*. J Hered 97:493–498

Gaill F, Mann K, Wiedemann H, Engel J, Timpl R (1995) Structural comparison of cuticle and interstitial collagens from annelids living in shallow sea-water and at deep-sea hydrothermal vents. J Mol Biol 246:284–294

Garoia F, Guarniero I, Grifoni D, Marzola S, Tinti F (2007) Comparative analysis of AFLPs and SSRs efficiency population genetic structure of Mediterranean *Solea vulgaris*. Mol Ecol 16:1377–1387

Gibson G (2002) Microarrays in ecology and evolution: a preview. Mol Ecol 11:17–24

Gibson G, Riley-Berger R, Harshman L, Kopp A, Vacha S, Nuzhdin S, Wayne M (2004) Extensive sex-specific nonadditivity of gene expression in *Drosophila melanogaster*. Genetics 167: 1791–1799

Gort G, Koopman WJM, Stein A (2006) Fragment length distributions and collision probabilities for AFLP markers. Biometrics 62:1107–1115

Govoroun M, Le Gac F, Guiguen Y (2006) Generation of a large-scale repertoire of Expressed Sequence Tags (ESTs) from normalised rainbow trout cDNA libraries. BMC Genom doi:10.1186/1471-2164-7-196

Gracey AY, Fraser EJ, Li W, Fang Y, Taylor RR, Rogers J, Brass A, Cossins AR (2004) Coping with cold: An integrative, multitissue analysis of the transcriptome of a poikilothermic vertebrate. Proc Natl Acad Sci USA 101:16970–16975

Gruenthal KM, Acheson LK, Burton RS (2007) Genetic structure of natural populations of California red abalone (*Haliotis rufescens*) using multiple genetic markers. Mar Biol 152: 1237–1248

Gruenthal KM, Burton RS (2008) Genetic structure of natural populations of the California black abalone (*Haliotis cracherodii* Leach, 1814), a candidate for endangered species status. J Exp Mar Biol Ecol 355:47–58

Guinand B, Lemaire C, Bonhomme F (2004) How to detect polymorphisms undergoing selection in marine fishes? A review of methods and case studies, including flatfishes. J Sea Res 51:167–182

Hadfield MG (1998) Research on settlement and metamorphosis of marine invertebrate larvae: past, present and future. Biofouling 12:9–29
Hadfield MG, Carpizo-Ituarte EJ, del Carmen K, Nedved BT (2001) Metamorphic competence, a major adaptive convergence in marine invertebrate larvae. Am Zool 41:1123–1131
Halpin PM, Sorte CJ, Hofmann GE, Menge BA (2002) Patterns of variation in levels of Hsp70 in natural rocky shore populations from microscales to mesoscales. Am J Physiol Integ Comp Biol 42:815–824
Hamady M, Walker JJ, Harris JK, Gold NJ, Knight R (2008) Error-correcting barcoded primers for pyrosequencing hundreds of samples in multiplex. Nat Meth 5:235–237
Hardivillier Y, Denis F, Demattei MV, Bustamante P, Laulier M, Cosson R (2006) Metal influence on metallothionein synthesis in the hydrothermal vent mussel *Bathymodiolus thermophilus*. Comp Biochem Physiol C Toxicol Pharmacol 143:321–332
Hayes BJ, Gjuvsland A, Omholt S (2006) Power of QTL mapping experiments in commercial Atlantic salmon populations, exploiting linkage and linkage disequilibrium and effect of limited recombination in males. Heredity 97:19–26
Hayes B, Laerdahl JK, Lien S, Moen T, Berg P, Hindar K, Davidson WS, Koop BF, Adzhubei A, Høyheim B (2007) An extensive resource of single nucleotide polymorphism markers associated with Atlantic salmon (*Salmo salar*) expressed sequences. Aquaculture 265:82–90
Hedgecock D, Lin JZ, DeCola S, Haudenschild CD, Meyer E, Manahan DT, Bowen B (2007) Transcriptomic analysis of growth heterosis in larval Pacific oysters (*Crassostrea gigas*). Proc Natl Acad Sci USA 104:2313–2318
Hedgecock D, McGoldrick DJ, Bayne BL (1995) Hybrid vigor in Pacific oysters: an experimental approach using crosses among inbred lines. Aquaculture 137:285–298
Hoekstra HE, Coyne JA (2007) The locus of evolution: evo-devo and the genetics of adaptation. Evolution 61:995–1016
Hoekstra HE, Nachman MW (2003) Different genes underlie adaptive melanism in different populations of rock pocket mice. Mol Ecol 12:1185–1194
Hofmann GE (2005) Patterns of gene expression in ectothermic marine organisms on small to large-scale biogeographical patterns. Intergr Comp Biol 45:247–255
Hofmann GE, Buckley BA, Airaksinen S, Keen JE, Somero GN (2000) Heat-shock protein expression is absent in the Antarctic fish *Trematomus bernacchii* family Nototheniidae. J Exp Biol 203:2331–2339
Hogstrand C, Balesaria S, Glover CN (2002) Application of genomics and proteomics for study of the integrated response to zinc exposure in a non-model fish species, the rainbow trout. Comp Biochem Physiol Mol Biol 133:523–535
Holland BS (2000) Genetics of marine bioinvasions. Hydrobiologia 420:63–71
Hourdez S, Weber RE (2005) Molecular and functional adaptations in deep-sea hemoglobins. J Inorg Biochem 99:130–131
Hubert S, Hedgecock D (2004) Linkage maps of microsatellite DNA markers for the pacific oyster *Crassostrea gigas*. Genetics 168:351–362
Hudson ME (2008) Sequencing breakthrough for genomic ecology and and evolutionary biology. Mol Ecol Res 8:3–17
IPCC (2007) Climate change 2007: synthesis report. Contribution of work groups I, II and III to the 4th Assessment Report of the Intergovernmental Panel on Climate Change. Core writing team: Pachauri RK and Reisinger A (eds). IPCC, Geneva, Switzerland.
Irwin RD, Boorman GA, Cunningham ML, Heinloth AN, Malarkey DE, Paules RS (2004) Application of toxicogenomics to toxicology: basic concepts in the analysis of microarray data. Toxicol Pathol 32:72–83
Jackson DJ, Ellemor N, Degnan BM (2005) Correlating gene expression with larval competence, and the effect of age and parentage on metamorphosis in the tropical abalone *Haliotis asinina*. Mar Biol 147:681–697
Jin Y, deVries AL (2006) Antifreeze glycoprotein levels in Antarctic notthenioid fishes inhabiting different thermal environments and the effect of warm acclimation. Comp Biochem Physiol B 144:290–300

Johannesson K, Andre C (2006) Life on the margin: genetic isolation and diversity loss in a peripheral marine ecosystem, the Baltic Sea. Mol Ecol 15:2013–2029

Jollivet D, Desbruyeres D, Ladrat C, Laubier L (1995) Evidence for differences in the allozyme thermostability of deep-sea hydrothermal vent polychaetes (Alvinellidae):a possible selection by habitat. Mar Ecol Prog Ser 123:125–136

Jorde LB (2000) Linkage disequilibrium and the search for complex disease genes. Gen Res 10:1435–1444

Kashi Y, King D, Soller M (1997) Simple sequence repeats as a source of quantitative genetic variation. Trends Genet 13:74–78

Kassahn KS (2008) Microarrays for comparative and ecological genomics: beyond single-species applications for array technologies. J Fish Biol 72:2407–2434

Kim A, Misra A (2007) SNP genotyping: technologies and biomedical applications. Ann Rev Biomed Eng 9:289–320

Kim KS, Ratcliffe ST, French BW, Liu L, Sappington TW (2008) Utility of EST-Derived SSRs as population genetics markers in a beetle. J Hered 99:112–124

Klinbunga S, Khetpu K, Khamnamtong B, Menasveta P (2007) Genetic heterogeneity of the blue swimming crab (*Portunus pelagicus*) in Thailand determined by AFLP analysis. Biochem Genet 45:725–736

Kumar S, Tsai CJ, Nussinov R (2000) Factors enhancing protein thermostability. Protein Eng 13:179–191

Lallias D, Beaumont AR, Haley CS, Boudry P, Heurtebise S, Lapègue S (2007a) A first-generation genetic linkage map of the European flat oyster *Ostrea edulis* (L.) based on AFLP and microsatellite markers. Anim Genet 38:560–568

Lallias D, Lapegue S, Hecquet C, Boudry P, Beaumont AR (2007b) AFLP-based genetic linkage maps of the blue mussel (*Mytilus edulis*). Anim Genet 38:340–349

Langaee T, Ronaghi M (2005) Genetic variation analyses by Pyrosequencing. Mut Res 573:96–102

Launey S, Hedgecock D (2001) High Genetic Load in the Pacific Oyster *Crassostrea gigas*. Genetics 159:255–265

Le Bris N, Zbinden M, Gaill F (2005) Processes controlling the physico-chemical micro-environments associated with Pompeii worms. Deep-Sea Res 52:1071–1083

Lee M, Kwon J, Kim SN, Kim JE, Koh WS, Kim EJ, Chung MK, Han SS, Song CW (2003) cDNA microarray gene expression profiling of hydroxyurea, paclitaxel, and p-anisidine, genotoxic compounds with differing tumorigenicity results. Environ Mol Mutagen 42:91–97

Lennard Richard ML, Bengten E, Wilson MR, Miller NW, Warr GW, Hikima J (2007) Comparative genomics of transcription factors driving expression of the immunoglobulin heavy chain locus in teleost fish. J Fish Biol 71(Suppl. B):153–173

Li L, Guo X (2004) AFLP-based genetic linkage maps of the Pacific oyster C*rassostrea gigas* Thunberg. Mar Biotech 6:26–36

Luikart G, Allendorf FW, Cornuet JM, Sherwin WB (1998) Distortion of allele frequency distributions provides a test for recent population bottlenecks. J Hered 89:238–247

Luikart G, England PR, Tallmon D, Jordan S, Taberlet P (2003) The power and promise of population genomics: from genotyping to genome typing. Nat Rev Genet 4:981–994

Luther GW, Rozan TF, Taillefert M, Nuzzio DB, Meo CD, Shank TM, Lutz RA, Cary SC (2001) Chemical speciation drives hydrothermal vent ecology. Nature 410:813–816

Lynch M, Walsh B (1998) Genetics and analysis of quantitative traits. Sinauer Associates, Sunderland, MA

Mackay TFC (2001) Quantitative trait loci in *Drosophila*. Nat Rev Genet 2:11–20

Mackenzie MG (2006) Characterisation of genes regulating muscle development and growth in two model pufferfish species (*Takifugu rubripes* and *Tetraodon nigroviridis*). Thesis St Andrews University, Scotland, UK

Maldini M, Marzano FN, Fortes GG, Papa R, Gandolfi G (2006) Fish and seafood traceability based on AFLP markers: Elaboration of a species database. Aquaculture 261: 487–494

Marcil A, Dumontier E, Chamberland M, Camper SA, Drouin J (2003) *Pitx1* and *Pitx2* are required for development of hindlimb buds. Development 130:45–55
Marden JH (2008) Quantitative and evolutionary biology of alternative splicing: how changing the mix of alternative transcripts affects phenotypic plasticity and reaction norms. Heredity 100:111–120
Margulies M et al. (2005) Genome sequencing in microfabricated high-density picolitre reactors. Nature 437:376–380
Maynard Smith J, Haigh J (1974) The hitchhiking effect of a favorable gene. Genet Res 23:23–35
Medina M (2009) Functional genomics opens doors to understanding metamorphosis in nonmodel invertebrate organisms. Mol Ecol 18(5):763–764
Meredith MP, King JC (2005) Rapid climate change in the ocean west of the Antarctic Peninsula during the second half of the 20th century. Geophys Lett 32:L19604–L19609
Meudt HM, Clarke AC (2007) Almost forgotten or latest practice? AFLP applications, analyses and advances. Trends Plant Sci 12:106–117
Minic Z, Herve G (2004) Biochemical and enzymological aspects of the symbiosis between the deep-sea tubeworm *Riftia pachyptila* and its bacterial endosymbiont. Eur J Biochem 271: 3093–3102
Mitchell-Olds T, Willis JH, Goldstein DB (2007) Which evolutionary processes influence natural genetic variation for phenotypic traits?. Nat Rev Genet 8:845–856
Moen T, Hayes B, Nilsen F, Delghandi M, Fjalestad K, Fevolden S-E, Berg PR, Lien S (2008) Identification and characterization of novel SNP markers in Atlantic cod: evidence for directional selection. BMC Genet 9:18
Moore MN (2001) Biocomplexity: the post-genome challenge in ecotoxicology. Aquat Toxicol 59:1–15
Morand P, Briand X (1996) Excessive growth of macroalgae: a symptom of environmental disturbance. Botanica Marina 39:491–516
Morin PA, Luikart G, Wayne RK SNP Workshop Group (2004) SNPs in ecology, evolution and conservation. Trends Ecol Evol 19:208–216
Moy GW, Springer SA, Adams SL, Swanson WJ, Vacquier VD (2008) Extraordinary intraspecific diversity in oyster sperm bindin. Proc Natl Acad Sci USA 105:1993–1998
Moylan TJ, Sidell BD (2000) Concentrations of myoglobin and myoglobin mRNA in heart ventricles from Antarctic fishes. J Exp Biol 203:1277–1286
Murphy JM, Mitchell JFB (1995) Transient-response of the Hadley-Center coupled ocean-atmosphere model to increasing carbon-dioxide. 2. Spatial and temporal structure of response. J Climate 1:57–80
Murray MC, Hare MP (2006) A genomic scan for divergent selection in a secondary contact zone between Atlantic and Gulf of Mexico oysters, *Crassostrea virginica*. Mol Ecol 15:4229–4242
Nielsen R (2005) Molecular signatures of natural selection. Ann Rev Genet 39:197–218
Nishio Y, Nakamura Y, Kawarabayasi Y, Usuda Y, Kimura E, Sugimoto S, Matsui K, Yamagishi A, Kikuchi H, Ikeo K, Gojobori T (2003) Comparative complete genome sequence analysis of the amino acid replacements responsible for the thermostability of Corynebacterium efficiens. Genome Res 13:1572–1579
Nordborg M, Tavare S (2002) Linkage disequilibrium: what history has to tell us. Trends Genet 18:83–90
Oetjen K, Reusch TBH (2007a) Genome scans detect consistent divergent selection among subtidal vs. intertidal populations of the marine angiosperm *Zostera marina*. Mol Ecol 16:5156–5157
Oetjen K, Reusch TBH (2007b) Identification and characterization of 14 polymorphic EST-derived microsatellites in eelgrass (Zostera marina). Mol Ecol Notes 7:777–780
O'Leary DB, Couglan J, McCarthy TV, Cross TF (2006) Application of a rapid method of SNP analysis (glycosylase mediated polymorphism detection) to mtDNA and nuclear DNA of cod *Gadus morhua*. J Fish Biol 69(Suppl. A):145–153
Orr JC, Fabry VJ, Aumont O, Bopp L, Doney SC, Feely RA, Gnandesikan A, Gruber N, Ishida A, Joos F, Key RM, Lindsay K, Plattner GK, Rodgers KB, Sabine CL, Sarmiento JL, Schlitzer

R, Slater RD, Totterdell IJ, Weirig MF, Yamanaka Y, Yool A (2005) Anthropogenic ocean acidification over the twenty-first century and its impact on calcifying organisms. Nature 437:681–686

Osovitz CJ, Hofmann GE (2005) Thermal history-dependant expression of the hsp70 gene in purple sea urchins: biogeographic patterns and the effect of temperature acclimation. J Exp Mar Biol Ecol 327:134–143

Otsuka M, Arai M, Mori M, Kato M, Kato N, Yokosuka O, Ochiai T, Takiguchi M, Omata M, Seki N (2003) Comparing gene expression profiles in human liver, gastric, and pancreatic tissues using full-length-enriched cDNA libraries. Hepato Res 1:76–82

Oudot M-P, Kloareg B, de Goër S (2002) The complete sequence of the mitochondrial genome of *Laminaria digitata* (Laminariales). Eur J Phycol 37:163–172

Oudot-Le Secq M-P, Fontaine J-M, Rousvoal S, Kloareg B, Loiseaux-de Goër S (2001) The complete sequence of a brown algal mitochondrial genome, the ectocarpale *Pylaiella littoralis* (L.) Kjellm. J Mol Evol 53:80–88

Pace DA, Marsh AG, Leong P, Green A, Hedgecock D, Manahan DT (2006) Physiological bases of genetically determined variation in growth of marine invertebrate larvae: a study of growth heterosis in the bivalve *Crassostrea gigas*. J Exp Mar Biol Ecol 353:188–209

Palumbi SR (1999) All males are not created equal: fertility differences depend on gamete recognition polymorphisms in sea urchins. Proc Natl Acad Sci 96:12632–12637

Pechenik JA (1999) On the advantages and disadvantages of larval stages in benthic marine invertebrate life cycles. Mar Ecol Prog Ser 177:269–297

Peck LS, Clark MS, Clarke A, Cockell CS, Convey P, Detrich IIIHW, Fraser KPP, Johnston I, Methe B, Murray AE, Romisch K, Rogers A (2005) Genomics: Applications to Antarctic Ecosystems. Polar Biol 28:351–365

Peck LS, Clark MS, Morley SA, Massey A, Rosetti H (2009) Animal temperature limits: effects of size, activity and rates of change. Func Ecol 23:248–256

Peck LS, Webb KE, Bailey DM (2004) Extreme sensitivity of biological function to temperature in Antarctic marine species. Func Ecol 18:625–630

Pemberton JM (2004) Measuring inbreeding depression in the wild: the old ways are the best. Trends Ecol Evol 19:613–615

Pfaffl MW (2001) A new mathematical model for relative quantification in real-time RT-PCR. Nucl Acids Res 29:2002–2007

Pfaffl MW, Horgan GW, Dempfle L (2002) Relative expression software tool (REST©) for group-wise comparison and statistical analysis of relative expression results in real-time PCR. Nucl Acids Res 30:1–10

Phillips C (2007) Online resources for SNP analysis – a review and route map. Mol Biotechnol 35:65–97

Piccino P, Viard F, Sarradin PM, Le Bris N, Le Guen D, Jollivet D (2004) Thermal selection of PGM allozymes in newly founded populations of the thermotolerant vent polychaete *Alvinella pompejana*. Proc Roy Soc Lond B 271:2351–2359

Place SP, O'Donnell MJ, Hofmann GE (2008) Gene expression in the intertidal mussel Mytilus californianus: physiological response to environmental factors on a biogeographic scale. Mar Ecol Prog Ser 356:1–14

Podrabsky JE, Somero GN (2004) Changes in gene expression associated with acclimation to constant temperatures and fluctuating daily temperatures in an annual killifish *Austrofundulus limnaeus*. J Exp Biol 207:2237–2254

Pompanon F, Bonin A, Bellemain E, Taberlet P (2005) Genotyping errors: causes, consequences and solutions. Nat Rev Genet 6:847–859

Price AH (2006) Believe it or not, QTLs are accurate!. Trends Plant Sci 11:213–216

Pruski AM, Dixon DR (2003) Toxic vents and DNA damage: first evidence from a naturally contaminated deep-sea environment. Aquat Toxicol 64:1–13

Purać J, Burns B, Thorne MAS, Grubor-Lajšić G, Worland MR, Clark MS (2008) Cold hardening processes in the Antarctic springtail, *Cryptopygus antarcticus*: clues from a microarray. J Insect Physiol 54:1356–1362

Qin Y, Liu X, Zhang H, Zhang G, Guo X (2007) Genetic mapping of size-related quantitative trait loci (QTL) in the bay scallop (*Argopecten irradians*) using AFLP and microsatellite markers. Aquaculture 272:281–290

Radonic A, Thulke S, Mackay IM, Landt O, Siegert W, Nitsche A (2004) Guideline to reference gene selection for quantitative real-time PCR. Biochem Biophys Res Comm 313:856–862

Renn SCP, Aubin-Horth N, Hofmann HA (2004) Biologically meaningful expression profiling across species using hetrerologous hybridisation to a cDNA microarray. BMC Genom doi: 10.1186/1471-2164-5-42

Ricciardi A, MacIsaac HJ (2000) Recent mass invasion of the north american great lakes by ponto-caspian species. Trends Ecol Evol 15:62–65

Roberts DA, Hofmann GE, Somero GN (1997) Heat shock protein expression in *Mytilus californianus*: Acclimatization (seasonal and tidal height comparisons) and acclimation effects. Biol Bull 192:309–320

Roberts S, Romano C, Gerlach G (2005) Characterization of EST derived SSRs from the bay scallop, *Argopecten irradians*. Mol Ecol Notes 5:567–568

Robinson-Rechavi M, Alibes A, Godzik A (2006) Contribution of electrostatic interactions, compactness and quaternary structure to protein thermostability: Lessons from structural genomics of *Thermotoga maritime*. J Mol Biol 356:547–557

Roff DA (2007) Contributions of genomics to life-history theory. Nat Rev Genet 8:116–125

Rogers SM, Bernatchez L (2005) Integrating QTL mapping and genome scans towards the characterization of candidate loci under parallel selection in the lake whitefish (*Coregonus clupeaformis*). Mol Ecol 14:351–361

Rogers SM, Campbell D, Baird SJE, Danzmann RG, Bernatchez L (2001) Combining the analyses of introgressive hybridisation and linkage mapping to investigate the genetic architecture of population divergence in the lake whitefish (*Coregonus clupeaformis*, Mitchill). Genetica 111:25–41

Roman J, Darling JA (2007) Paradox lost: genetic diversity and the success of aquatic invasions. Trends Ecol Evol 22:454–464

Römisch K, Collie N, Soto N, Logue J, Lindsay M, Scheper W, Cheng C-H C (2003) Protein translocation across the endoplasmic reticulum membrane in cold adapted organisms. J Cell Sci 116:2875–2883

Royal Society (2005) Ocean acidification due to increasing atmospheric carbon dioxide. Policy Document 12/05, The Royal Society

Rungis D, Berube Y, Zhang J, Ralph S, Ritland CE, Ellis BE et al. (2004) Robust simple sequence repeat markers for spruce (*Picea* spp.) from expressed sequence tags. Theor Appl Genet 109:1283–1294

Ryynänen HJ, Primmer C (2006) Single nucleotide polymorphism (SNP) discovery in duplicated genomes: intron-primed exon-crossing (IPEC) as a strategy for avoiding amplification of duplicated loci in Atlantic salmon (*Salmo salar*) and other salmonid fishes. BMC Genom 7:192

Ryynänen HJ, Tonteri A, Vasemägi A, Primmer CR (2007) A comparison of the efficiency of single nucleotide polymorphisms (SNPs) and microsatellites for the estimates of population and conservation genetic parameters in Atlantic salmon (*Salmo salar*). J Hered 98:692–704

Saastamoinen M, Hanski I (2008) Genotypic and environmental effects on flight activity and oviposition in the Glanville fritillary butterfly. Am Nat 171:701–712

Saavedra C, Bachere E (2006) Bivalve genomics. Aquaculture 256:1–14

Sanchez S, Hourdez S, Lallier F (2007) Identification of proteins involved in the functioning of *Riftia pachyptila* symbiosis by Subtractive Suppression Hybridization. BMC Genomics 8:337

Sax DF, Stachowicz JJ, Brown JH, Bruno JF, Dawson MN, Gaines SD, Grosberg RK, Hastings A, Holt RD, Mayfield MM, O'Connor MI, Rice WR (2007) Ecological and evolutionary insights from species invasions. Trends Ecol Evol 22:465–471

Schlotterer C (2003) Hitchhiking mapping – functional genomics from the population genetics perspective. Trends Genet 19:32–38

Shank TM, Halanych KM (2007) Toward a mechanistic understanding of larval dispersal: insights from genomic fingerprinting of the deep-sea hydrothermal vent tubeworm *Riftia pachyptila*. Mar Ecol 28:25–35

Shapiro MD, Marks ME, Peichel CL, Blackman BK, Nereng BJ, Schluter D et al. (2004) Genetic and developmental basis of evolutionary pelvic reduction in threespine sticklebacks. Nature 428:717–723

Sheader DL, Williams TD, Lyons BP, Chipman JK (2006) Oxidative stress response of European flounder (Platichthys flesus) to cadmium determined by a custom cDNA microarray. Mar Environ Res 62:33–44

Sicot FX, Mesnage M, Masselot M, Exposito JY, Garrone R, Deutsch J, Gaill F (2000) Molecular adaptation to an extreme environment: origin of the thermal stability of the Pompeii worm collagen. J Mol Biol 302:811–820

Slate J, David P, Dodds KG, Veenvliet BA, Glass BC, Broad TE et al. (2004) Understanding the relationship between the inbreeding coefficient and multilocus heterozygosity: theoretical expectations and empirical data. Heredity 93:255–265

Smith CT, Antonovich A, Templin WD, Elfstrom CM, Narum SR, Seeb LW (2007) Impacts of marker class bias relative to locus-specific variability on population inferences in chinook salmon: A comparison of single-nucleotide polymorphisms with short tandem repeats and allozymes. Trans Am Fish Soc 136:1674–1687

Smith CT, Elfstrom CM, Seeb LW, Seeb JE (2005) Use of sequence data from rainbow trout and Atlantic salmon for SNP detection in Pacific salmon. Mol Ecol 14:4193–4203

Somero GN (2002) Thermal physiology and vertical zonation of intertidal animals: optima, limits and costs of living. Am J Physiol Integ Comp Biol 42:780–789

Stachowicz JJ, Terwin JR, Whitlatch RB, Osman RW (2002) Linking climate change and biological invasions: Ocean warming facilitates nonindigenous species invasions. Proc Natl Acad Sci USA 99:15497–15500

Stern S, Dror T, Stolovicki E, Brenner N, Braun E (2007) Genome-wide transcriptional plasticity underlies cellular adaptation to novel challenge. Mol Sys Biol Doi:10.1038/msb4100147

Stewart FJ, Cavanaugh CM (2006) Bacterial endosymbioses in Solemya (Mollusca: Bivalvia)—model systems for studies of symbiont–host adaptation. Antonie Leeuwenhoek 90:343–360

Stinchcombe JR, Hoekstra HE (2008) Combining population genomics and quantitative genetics: finding the genes underlying ecologically important traits. Heredity 100:158–170

Stolovitzky GA, Kundaje A, Held GA, Duggar KH, Haudenschild CD, Zhou D, Vasicek TJ, Smith KD, Aderem A, Roach JC (2005) Proc Natl Acad Sci USA 102:1402–1407

Storz JF (2005) Using genome scans of DNA polymorphism to infer adaptive population divergence. Mol Ecol 14:671–688

Streelman JT, Kocher TD (2002) Microsatellite variation associated with prolactin expression and growth of salt-challenged tilapia. Physiol Genom 9:1–4

Suzuki Y, Yoshitomo-Nakagawa K, Maruyama K, Suyama A, Sugano S (1997) Construction and characterization of a full length-enriched and a 5′-end-enriched cDNA library. Gene 1–2: 149–156

Swanson-Wagner RA, Jia Y, De Cook R, Borsuk LA, Nettleton D, Schnable PS (2006) All possible modes of gene action are observed in a global comparison of gene expression in a maize F-1 hybrid and its inbred parents. Proc Natl Acad Sci USA 103:6805–6810

Tabor HK, Risch NJ, Myers RM (2002) Candidate gene approaches for studying complex genetic traits: practical considerations. Nat Rev Genet 3:391–397

Tanguy A, Bierne N, Saavedra C, Pina B, Bachère E, Kube M, Bazin E, Bonhomme F, Boudry P, Boulo V, Boutet I, Cancela L, Dossat C, Favrel P, Huvet A, Jarque S, Jollivet D, Klages S, Lapegue S, Leite R, Moal J, Moraga D, Reinhardt R, Samain J-F, Zouros E, Canario A (2008) Increasing genomic information in bivalves through new EST collections in four species: Development of new genetic markers for environmental studies and genome evolution. Gene 408:27–36

Tao WJ, Boulding EG (2003) Associations between single nucleotide polymorphisms in candidate genes and growth rate in Arctic charr (*Salvelinus alpinus* L. Heredity 91:60–69

Taris N, Comtet T, Viard F (2009) Inhibitory function of nitric oxide on the onset of metamorphosis in competent larvae of *Crepidula fornicata*: a transcriptional perspective. Mar Genomics 2: 161–167

Teranishi KS, Stillman JH (2007) A cDNA microarray analysis of the response to heat stress in hepatopancreas tissue of the porcelain crab *Petrolisthes cinctipes*. Comp Biochem Physiol D 2:53–62

Teshima KM, Coop G, Preworski M (2006) How reliable are empirical genomic scans for selective sweeps?. Genome Res 16:702–712

Tomanek L (2002) The heat shock response: its variation, regulation and ecological importance in intertidal gastropods (genus *Tegula*). Am J Physiol Integ Comp Biol 42:797–807

Tomanek L (2005) Two-dimensional gel analysis of the heat shock response in marine snails (genus *Tegula*): interspecific variation in protein expression and acclimation ability. J Exp Biol 208:3133–3143

Tomanek L, Somero GN (2000) Time course and magnitude of synthesis of heat shock proteins in congeneric marine snails (Genus *Tegula*) from differentb tidal heights. Physiol Biochem Zool 73:249–256

Toth AL, Varala K, Newman TC, Miguez FE, Hutchison SK, Willoughby DA et al. (2007) Wasp brain gene expression supports an evolutionary link between maternal behavior and eusociality. Science 318:441–444

Troester MA, Hoadley KA, Parker JS, Perou CM (2004) Prediction of toxicant-specific gene expression signatures after chemotherapeutic treatment of breast cell lines. Environ Health Perspect 112:1607–1613

Ungerer MC, Johnson LC, Herman MA (2008) Ecological genomics: understanding gene and genome function in the natural environment. Heredity 100:178–183

Uthicke S, Conand C (2005) Amplified fragment length polymorphism (AFLP) analysis indicates the importance of both asexual and sexual reproduction in the fissiparous holothurian Stichopus chloronotus (Aspidochirotida) in the Indian and Pacific Ocean. Coral Reefs 24:103–111

Vasemägi A, Nilsson J, Primmer CR (2005) Expressed Sequence Tag-Linked Microsatellites as a Source of Gene-Associated Polymorphisms for Detecting Signatures of Divergent Selection in Atlantic Salmon (*Salmo salar* L). Mol Biol Evol 22:1067–1076

Vasemägi A, Primmer CR (2005) Challenges for identifying functionally important genetic variation: the promise of combining complementary research strategies. Mol Ecol 14:3623–3642

Vaughan P (2000) Use of uracil DNA glycosylase in the detection of known DNA mutations and polymorphisms. Glycosylase-mediated polymorphism detection (GMPD-check). Methods Mol Biol (Clifton, NJ) 152:169–177

Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA et al. (2004) Environmental genome shotgun sequencing of the Sargasso Sea. Science 304:66–74

Vera JC, Wheat CW, Fescemyer HW, Frilander MJ, Crawford DL, Hanski I, Marden JH (2008) Rapid transcriptome characterization for a nonmodel organism using 454 pyrosequencing. Mol Ecol 17:1636–1647

Verde C, Lecointre G, di Prisco G (2007) The phylogeny of polar fishes and the structure and molecular evolution of hemoglobin. Polar Biol 30:523–539

Voisin M, Engel C, Viard F (2005) Differential shuffling of native genetic diversity across introduced region in a brown alga: aquaculture vs. maritime traffic effects. Proc Natl Acad Sci USA 102:5432–5437

Vos P, Hogers R, Bleeker M, Reijans M, van de Lee T, Hornes M, Frijters A, Pot J, Peleman J, Kuiper M, Zabeau M (1995) AFLP: a new technique for DNA fingerprinting. Nucl Acids Res 23:4407–4414

Ward RD, Woodwark M, Skibinski DOF (1994) A comparison of genetic diversity levels in marine, freshwater, and anadromous fishes. J Fish Biol 44:213–232

Wares JP, Blakeslee AMH (2007) Amplified fragment length polymorphism data provide a poor solution to the *Littorina littorea* puzzle. Mar Biol Res 3:168–174

Weetman D, Ruggierro E, Mariani S, Shaw PW, Lawler AR, Hauser L (2007) Hierarchical population genetic structure in the commercially exploited shrimp *Crangon crangon* identified by AFLP analysis. Mar Biol 151:565–575

Weiss E, Bennie M, Hodgins-Davis A, Roberts S, Gerlach G (2007) Characterization of new SSR-EST markers in cod, *Gadus morhua*. Mol Ecol Notes 7:866–867

Wenne R, Boudry P, Hemmer-Hansen J, Lubieniecki KP, Was A, Kause A (2007) What role for genomics in fisheries management and aquaculture?. Aquatic Living Res 3:241–255

Wilding CS, Butlin RK, Grahame J (2001) Differential gene exchange between parapatric morphs of Littorina saxatilis detected using AFLP markers. J Evol Biol 14:611–619

Williams EA, Degnan BM, Gunter H, Jackson DJ, Woodcroft BJ, Degnan SM (2009) Widespread transcriptional changes pre-empt the critical pelagic–benthic transition in the vetigastropod Haliotis asinina. Mol Ecol 18:1006–1025

Wilson K, Thorndyke M, Nilsen F, Rogers A, Martinez P (2005) Marine systems: moving into the genomics area. Mar Ecol 26:3–16

Wolff WJ, Reise K (2002) Oyster imports as a vector for the introduction of alien species into northern and western European coastal waters. In: Leppäkoski E, Gollasch S, Olenin S (eds) Invasive aquatic species of Europe. Distribution, impacts and management. Kluwer Academic Publishers, Dordrecht/Boston/London, pp 193–205

Woodhead M, Russell J, Squirrell J, Hollingsworth PM, Mackenzie K, Gibby M et al. (2005) Comparative analysis of population genetic structure in *Athyrium distentifolium* (Pteridophyta) using AFLPs and SSRs from anonymous and transcribed gene regions. Mol Ecol 14:1681–1695

Wray GA (2007) The evolutionary significance of *cis*-regulatory mutation. Nat Rev Genet 8: 206–216

Yang Z, Bielawski J (2000) Statistical methods for detecting molecular adaptation. Trends Ecol Evol 15:496–503

Yu J et al. (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat Genet 38:203–208

Yu Z, Guo X (2003) Genetic linkage map of the Eastern oyster *Crassostrea virginica* Gmelin. Biol Bull 204:327–338

Zachos J, Pagani M, Sloan L, Thomas E, Billups K (2001) Trends, rhythms and aberrations in global climate 65 Ma to present. Science 292:686–693

Zane L, Bargelloni L, Patarnello T (2002) Strategies for microsatellite isolation: a review. Mol Ecol 11:1–16

Zhao YM, Li Q, Kong LF, Bao ZM, Zhang RC (2007) Genetic diversity and divergence among clam *Cyclina sinensis* populations assessed using amplified fragment length polymorphism. Fish Sci 73:1338–1343

Chapter 4
Phylogeny of Animals: Genomes Have a Lot to Say

Ferdinand Marlétaz and Yannick Le Parco

Abstract Multiple lines of evidence have been proposed to resolve the tree of metazoans. Views based on morphology and development were often questioned because they relied on characters whose evolutionary orientation is difficult. Molecules offer an independent perspective and the employment of some genes, such as ribosomal RNA subunits (SSU/LSU) or Hox genes, have led to a profound reshaping of the metazoan tree, leading to the "New View" of animal phylogeny. However, classical molecular approaches have not succeeded in settling some long-standing issues in animal relationships. Recently, extensive genome data have been collected for a large set of organisms including several evolutionarily relevant marine species. This has allowed the development of phylogenomic approaches, which have the potential to overcome the limitations of single-gene phylogenies by inferring trees on the base of whole genome evidence. This new approach has triggered several advances regarding metazoan relationships such as the reevaluation of chordate relationships and the re-positioning of some problematic minor phyla with improved accuracy. These studies have also raised new questions about the processes that underlie morphological evolution, as well as the future of molecular phylogenetics.

4.1 Introduction

The first attempt to establish a classification of animals was made by Aristotle in *De generatione animalium,* which defined five main groups according to the degree of perfection of their generation and the relative warmness of their blood: mammals, ovoviviparous sharks, birds and reptiles, marine animals (fishes, cephalopods & crustaceans) and finally insects (Aristotle 1965). Just one of these categories, the marine animals, contains representatives of the whole tree of animals, if defined

F. Marlétaz (✉)
Station Marine d'Endoume, UMR 6540 DIMAR, CNRS & Université de la Méditerranée, Marseille, France
e-mail: ferdinand.marletaz@me.com

J.M. Cock et al. (eds.), *Introduction to Marine Genomics*,
Advances in Marine Genomics 1, DOI 10.1007/978-90-481-8639-6_4,

from a modern phylogenetic perspective (Fig. 4.2). This anecdote illustrates the lack of attention that we – as narrow-minded terrestrial animals – have paid to our marine relatives but it also strengthens the major importance of studying the astounding diversity of marine animal forms in order to understand our own ancestry.

The major difficulty with the classification of metazoans is the broad diversity of animal forms and body organizations in the 36 commonly admitted phyla. This diversity attains a maximum in marine organisms where, for example, the closest relatives of vertebrates have recently been demonstrated to be the urochordates whose adult form looks markedly different from the classical chordate body organization (Fig. 4.1) (Delsuc et al. 2006). Numerous strategies, such as the recapitulation theory or the cladistic method, have been promoted to establish hierarchies of morphological characters, but none of them have completely succeeded in dealing with convergent or parallel evolution (Jenner 2004). Recent discoveries coming from the evolutionary developmental biology field (the so-called *Evodevo* field) have stressed this problem by demonstrating that homologous genetic pathways are often involved

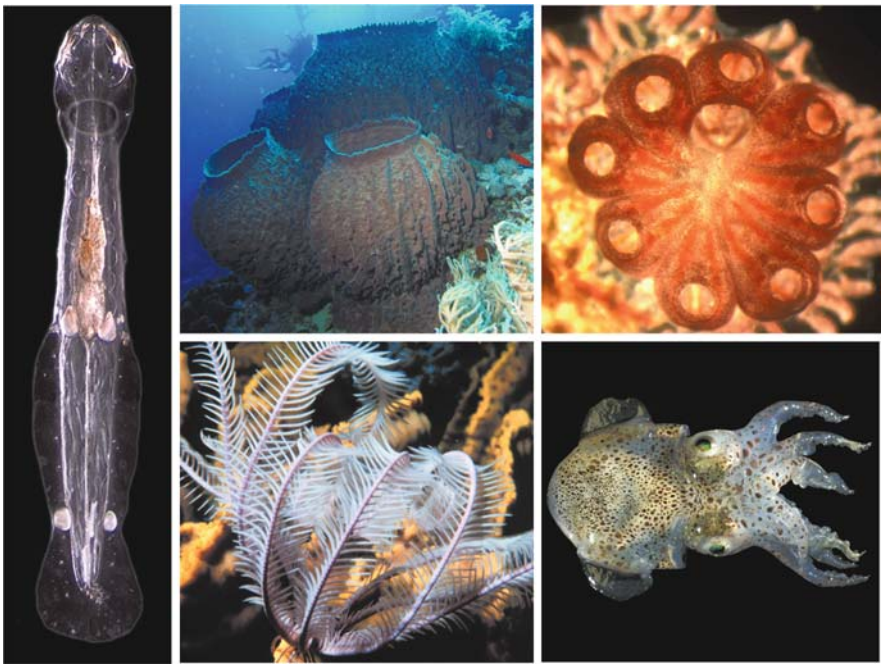

Fig. 4.1 Illustration of the diversity of metazoan body plans. (*left*) The chaetognath *Spadella cephaloptera* is representative of one of the most unique bilaterian phyla. (*middle top*) The massive barrel sponges *Xestospongia testudinaria.* (**c**) Several individuals of the colonial ascidian *Botrylloides leachi,* which belongs to urochordates, the closest relatives of the chordates (*top left*). (**d**) Another member of the deuterostomes, the crinoid *Antedon* (*bottom middle*). (**e**) The bobtail squid *Sepiola atlantica*, a cephalopod that displays numerous innovative features with respect to its body plan (*bottom left*)

in the development of unambiguously convergent organs such as limbs of arthropods and vertebrates (Shubin et al. 1997). Conversely, dissimilar body plans such as those of cnidarians and vertebrates share the extensive genetic networks that underlie anteroposterior and dorsoventral patterning (Martindale 2005). As a whole, this rephrases in molecular terms the problem of homology assignment that caused considerable problems in many morphological studies (Gould 2002, Wagner 2007). The accurate interpretation of developmental and morphological characters thus requires the use of an independent reference, which would help to determine the steps of animal evolution. The search for this reference has been pursued by the development of molecular inference methods and the study of signatures within some specific genes. This has led to a reevaluation of animal classification with the rise of the "new view" of animal phylogeny (Halanych 2004), which is based extensively on the use of 18S ribosomal RNA as a universal molecular marker. However, despite its indisputable success, this approach failed to solve some long-standing issues of animal phylogeny such as, for example, the resolution of relationships at the base of the metazoan tree or the branching of some *incertae sedis* (e.g. chaetognaths or acoel flatworms, Fig. 4.1). These limitations could be explained by two possible problems: (a) stochastic error related to the insufficient amount of information in the sampled marker genes and (b) systematic errors related to the impossibility for the molecular evolution model to fully account for the data.

The availability of a growing amount of genomic data for a broad range of organisms is now opening up a new field of investigation and providing new clues about animal relationships. These genomic data enable the assembly of datasets composed of a large number of protein coding genes. This can eliminate the problem of stochastic error, but can also help to identify more qualitative molecular characters, sometimes called "rare genomic changes" (Rokas and Holland 2000a). In this chapter, we will show how these genomic approaches have reshaped our view of animal relationships by attempting to describe both the successes and the pitfalls of these approaches. The development of improved inference methods but also the sequencing effort undertaken have led to some deep rearrangements of metazoan trees that will certainly change our view of animal evolution.

4.2 The Roots of Animal Phylogeny

Before the molecular biology era, the science of systematics had attempted to deal with the morphological complexity by proposing various methods to handle character interpretation.

4.2.1 Historical Schemes Are Based on the Coelom Evolution Hypotheses

The "theory of recapitulation" formulated by Haeckel in 1866 postulated that the successive steps of evolution had been accomplished through the addition of supplementary embryological stages (Gould 1977). Thus, the deepest levels of

evolutionary relationships between organisms could be inferred from the similarity of their earliest developmental stages. As a result, embryological characters were given prominent importance (Valentine 1997). Most historical hypotheses about animal relationships have thus been based on assumptions about the structure and formation of the coelomic cavities.

The so-called "traditional textbook" phylogeny, which is based to a large degree on the work of Libbie Hyman, mainly relied on the proposition that a coelom arose independently in protostomes and deuterostomes (Hyman 1940–1967). Accordingly, distinct mechanisms of coelom formation, schizocoely and enterocoely, would have respectively arisen from a triproblastic acoelomate ancestor of bilaterians (see Willmer 1990). This led some to consider that the acoelomate (platyhelminthes and nemertes) and pseudocoelomate (e.g. nematodes and rotifers) lineages constituted early offshoots during metazoans evolution (Fig. 4.2) (Barnes 1974). But such schemes should be distinguished from the original thinking of Libbie Hyman whose gradist approach was focused exclusively on the level of complexity in body plans and was not necessarily aimed at precisely deciphering evolutionary radiations (Jenner 2004).

However, a divergent interpretation of coelom evolution could lead to radically opposite views about metazoan relationships. For instance, the archecoelomate hypothesis, commonly adopted by German workers, was built upon the assumption that the bilaterian ancestor possessed a trimeric coelom developed by enterocoely. These multiple coelomic cavities would have been derived from the gastric pouches of an anthozoan-like ancestor (Remane 1963). The lineages that develop their coeloms through outpocketing of the archenteron such as hemichordates, chordates, lophophorates and chaetognaths were thus regarded as having diverged early among bilaterians (Siewing 1976) (Fig. 4.2). The segmented bodyplans encountered in annelids and arthropods would have subsequently originated through the segmentation of the posterior coelomic cavity after the loss of the anterior cavities (Tautz 2004). The body plans of acoelomates and pseudocoelomates were considered to represent various stages of coelomic degeneration from the ancestral type (Siewing 1976).

Alternative schemes have been proposed with respect to these examples and numerous refinements of each of these schemes have been formulated (Willmer 1990) but those cases perfectly illustrate the problem of character orientation that underlies the historical discussion of animal phylogeny.

4.2.2 Sorting More Characters Through a Cladistic Approach

Numerous developmental and morphological characters have been considered in the attempt to reconstruct metazoan relationships but they only contribute information about specific nodes of the tree (for an extensive review, see Willmer 1990). It has been argued, for example, that Bilaterians acquired a third germ layer (the mesoderm) and bilateral symmetry based on the two germ layers and radial symmetry described in basal metazoans: cnidarians, poriferans and ctenophores. But

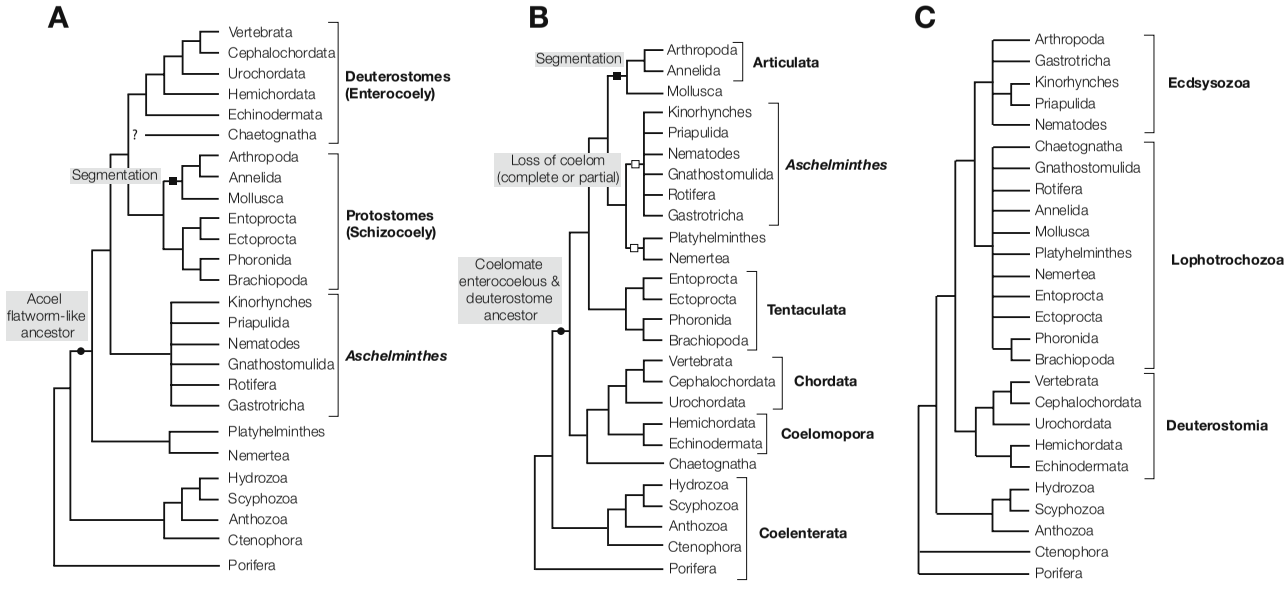

Fig. 4.2 Some historical schemes of metazoan phylogeny. (**a**) An extreme representation of the "traditional" textbook phylogeny assuming the progressive divergence of increasingly complex forms of bilaterian animals from a triploblastic acoelomate ancestor. The main character underlying this view is the establishment of body cavities, from Barnes (1974). (**b**) An archaecoelomate scheme of metazoan phylogeny. This view postulates an early origin of the coelom and enterocoely as the ancestral mode of coelom formation, which results in organisms with such developmental features being at the base of the tree (Marcus 1958). (**c**) The "new view" of metazoan phylogeny that splits all bilaterians into three main clades on the base of evidence from SSU molecular phylogeny and Hox genes, from Adoutte et al. (2000)

some characters are more difficult to interpret because they could lead to contradictory associations: common segmented bodies have historically been used to cluster annelids and arthropods into the "Articulés" group (Cuvier 1817) whereas other prominent characters such as a spiral cleavage pattern and trochophore-type larvae are shared between annelids and platyhelminthes, whose morphology is far simpler, but not arthropods (Nielsen 2001).

The cladistic method founded by Willi Hennig in 1950 provided an opportunity to deal more accurately with character evolution and to improve morphology-based classification (Hennig 1966). The cladistic method rejects groupings supported by the lack of a given character and conversely proposes that bona fide groupings (i.e. clades) should be supported by unambiguous shared derived characters (i.e. synapomorphies). This framework prompted extensive studies of morphological characters in animal phylogeny, in particular through the development of parsimony algorithms aimed at inferring a tree from a character matrix (Swofford 1990). For instance, the previous example of Spiralia versus Articulata was addressed in a landmark study, which rejected the Articulata hypothesis using a character matrix relying on careful definition of characters (Eernisse et al. 1992). This study thus led to the proposition that segmentation was established convergently in both annelids and arthropods, and uncovered novel new clades reminiscent of the "new view" of animal phylogeny (cf. infra).

Despite their successes, the studies based on morphological matrices were extensively criticized. Some authors simply proposed that morphology does not provide a sufficiently large set of unambiguous characters, in particular because of problems related to homology assessment (Scotland et al. 2003).

One may consider that such opinions are overstated. Methodological advances can bring generate new lines of morphological evidence and thereby increase the repertory of available characters. Ultrastructural studies performed using confocal laser scanning microscopy have recently shed new light on some debates, for example by providing evidence for a close relationship between sipunculids and annelids despite the apparent lack of segmentation in sipunculids (Wanninger et al. 2005). Similarly, 4D microscopy technique has offered opportunities to investigate cell lineages and fate maps with greater flexibility, and this has lead to some unexpected conclusions such as the potentially derived state of the appendicularian *Oikopleura dioica* with regard to tunicates (Stach et al. 2008).

However, numerous problems have been detected in the character matrices generally employed (Jenner 2001). In particular, the discrete coding of characters is rarely compatible with the subtle discussion of homology advocated by original cladistic thought (Hennig 1966, Jenner 2001). These issues could be especially misleading in the context of so-called total evidence studies that performed inference using a composite molecular and morphological dataset analyzed in a parcimony framework (Giribet et al. 2000, Jenner 2001). Nevertheless, this has debate stressed the importance of dealing with an alternative class of evidence – molecular data – in order to evaluate the phylogenetic accuracy of commonly discussed morphological characters.

4.2.3 Small Ribosomal RNA Gene and the "New View" of Animal Phylogeny

The inclusion of molecular data has led to a deep reorganization of the metazoan tree. The first step was the publication, twenty years ago, by Fields et al. (1988) of a 22 taxa dataset of SSU ribosomal RNA (also called 18S rRNA), which strikingly proposed the polyphyly of metazoans, with independent origins of bilaterians and cnidarians among eukaryotes. An early divergence of platyhelminthes within the bilaterians was also stressed (Field et al. 1988). However, despite the pioneering aspect of this work, these hypotheses were later rejected when the reinterpretation of the data demonstrated that the distance method employed had likely caused a long branch attraction artefact, a common systematic error related to the unequal rates of evolution among taxa (Felsenstein 1978). The use of alternative methods, such as evolutionary parsimony, applied to the same dataset, recovered the metazoans and deuterostomes as monophyletic clades and identified a close relationship between annelids and molluscs (Lake 1990). This setting up of new criteria for critical assessment of molecular phylogenies thus opened a new field of investigation that was going to strongly reshape our view of metazoan phylogeny.

The first major challenge to the historical views was the inclusion of the three lophophorate phyla within the protostomes. This reassignment was based on sequencing and analysis of SSU rRNA genes from phoronids and brachiopods by Halanych et al. (1995). The switch of the lophophorates phyla from deuterostomes to protostomes was not in the agreement with some of their characters, particularly the structure and development of their coelomic cavities (Eernisse et al. 1992, Emig 1982). This finding prompted the authors to coin the term Lophotrochozoa to describe this new clade uniting lophophorates and trochophore-bearing animals such as molluscs and annelids. However, the exact status of both pseudocoelomates (former aschelminthes) and arthropods with respect to this lophotrochozoan clade remained uncertain, especially because these lineages exhibit very fast evolutionary rates that still cause long branch attraction (see Bergsten 2005). In order to overcome this problem, Aguinaldo et al. (1997) selected the least diverged sequences of SSU genes for representative protostome animals and particularly for the very rapidly evolving nematodes. This approach allowed them to recover a clade of moulting animals, that they named Ecdysozoa, which included the arthropods and several former aschelminthes: the priapulids and the nematodes. Similarly, their analyses argued for the inclusion of the platyhelminthes within lophotrochozoans, which would indicate a split of the protostomes into two main clades: the ecdysozoans and the lophotrochozoans. This topology was later established as the "new view" of animal phylogeny because such a tree contrasts with former morphology-based schemes, particularly in the way it splits the former aschelminthes. This result unexpectedly indicated that organisms with very dissimilar body plans (e.g. nematodes and arthopods or platyhelminthes and annelids) could have close evolutionary relationships (Adoutte et al. 2000).

The "new view" of animal relationships was further reinforced by the discovery of molecular signatures derived from the homeodomain of Hox genes (de Rosa et al. 1999). The Hox transcription-factors constitute a multigenic family of about ten paralogous members present in all bilaterians. Those genes are generally tightly clustered in the genome and are expressed in the same order along the body plan of animals as the order of the genes along the chromosome following the so-called colinearity rule (Lemons and McGinnis 2006). The isolation of hox genes from priapulids and brachiopods by de Rosa et al. (1999) indicated that the posterior hox genes of ecdysozoans and lophotrochozans are likely to have independent origins, resulting in the respective Abd-B and Post1/2 classes that may be distinguished by some specific residues (de Rosa et al. 1999). Subsequent data have confirmed the interest of hox genes as molecular signatures of bilaterian evolution (Balavoine et al. 2002).

4.2.4 The Limits of the "New View"

The "new view" of animal phylogeny durably reshaped our understanding of animal evolution, especially through the reevaluation of the notion of body complexity (Adoutte et al. 2000). However, these early molecular data failed to convincingly resolve numerous nodes of the tree: (a) the relationships within the new protostome clades (Lophotrochozoa and Ecdysozoa) remained extremely elusive, (b) a set of phyla including chaetognaths, gastrotriches, rotifers, xenoturbellids and acoels were very difficult to position, often because of their fast evolutionary rates, and (c) the relationships at the base of the metazoan tree remained virtually unresolved. An attempt to deal with these issues was made by incorporating data derived from morphological matrices, in addition to the molecular data (Giribet et al. 2000, Peterson and Eernisse 2001). These "total evidence" studies succeeded in resolving problematic nodes and produced a convincing scheme of animal relationships but they were extensively criticized for having carelessly recycled previously published data matrices and thus reproducing their biases and errors (Jenner 2001).

Another approach taken to resolve this problem was the initiation of large-scale sequencing efforts to identify the least divergent species, but this approach did not always succeed. For instance, the chaetognaths exhibit very fast evolving SSU genes, which originally led to them being assigned a basal position among the metazoans (Telford and Holland 1993). However, further characterization of SSU genes from a large set of species did not lead to the identification of a slower evolving one (Papillon et al. 2006). Similarly, results suggesting that acoel flatworms had a basal position among bilaterians were sometimes criticized for possibly being biased by long-branch attraction (Deutsch 2008, Ruiz-Trillo et al. 1999). Increased taxon sampling at the base of the metazoan tree, with the addition of cnidarians, poriferans, ctenophores and also the enigmatic placozoans, led to interesting hypotheses such as sponge paraphyly but no consensus was drawn concerning the respective relationships of these lineages (Borchiellini et al. 2001, Collins 1998).

Another approach used was to to employ alternative marker genes such as the large subunit of ribosomal RNA (LSU or 28S), often in combination with other genes such as SSU (Mallatt and Winchell 2002, Winchell et al. 2002), elongation factor I alpha (EFIa) (Littlewood et al. 2001), heat-shock proteins (Borchiellini et al. 1998) or sodium-potassium ATPase beta-subunit (Anderson et al. 2004). This approach has provided valuable insights into the phylogeny of metazoans, but as a whole the trees based on each of the individual genes exhibited very incongruent topologies. The inference of an improved metazoan tree based on nuclear genes thus requires that this problem of incongruence be settled.

4.3 The Power and Pitfalls of Phylogenomics

Genome is the far side of all known organisms and represent inestimable source of phylogenetic information. Indeed, one can consider that morphological features are often shaped by strong evolutionary pressure whereas reliable arguments indicate that genome evolution is an essentially neutralistic process (see Kimura 1983, Lynch 2007). Two main classes of phylogenetic evidence could be drawn from genomic data: the first and most quantitative consists of inferring trees from the sequences of large sets of nuclear genes sampled in the genomic data. A second, more qualitative, type of evidence consists of detecting discrete molecular signatures by surveying whole genome features (Philippe et al. 2005a).

The first approach is based on the analysis of primary sequences from a large set of genes. This approach attempts to overcome the incongruence problem observed when the topologies obtained from independent marker genes are compared (cf. supra). It has been proposed that the concatenation of a large number of nuclear genes enables the detection of the *bona fide* species tree among the alternative topologies that are retrieved for different genes (Rokas et al. 2003). However, the large number of positions considered tends to increase systematic biases, which could lead to well-supported but inaccurate phylogenies (Jeffroy et al. 2006). Such systematic biases are mainly related to the heterogeneity of evolutionary rates and sequence composition, resulting in differences between taxa and sites of the alignment (Philippe et al. 2005a). It has been proposed that such biases could be limited by using improved models of substitution and by increasing the taxon sampling (Delsuc et al. 2005). For instance, the long-branch attraction problems could be reduced by using the CAT model that has recently been developed to account for the heterogeneity of sites along the alignment (Lartillot and Philippe 2004).

Two strategies can be employed to analyze such gene-rich datasets: the supertree or the supermatrix (Delsuc et al. 2005). The supertree approach infers an independent tree for each marker gene, with the subsequent computation of a supertree that summarizes all the independent topologies recovered. This approach limits

the systematic error because evolutionary models generally fit short and homogeneous one-gene alignments better. However, tree combination methods generally cause loss of information compared to full data phylogenetic estimates (de Queiroz and Gatesy 2007). The second approach, the supermatrix strategy, involves the concatenation of all marker genes in a single alignment. The analysis of such a large number of positions can be difficult in terms of models and computation but this approach has benefited considerably from recent improvements of inference methods. For instance, efficient maximum-likelihood heuristics (Guindon and Gascuel 2003, Stamatakis 2006) as well as accurate Bayesian reconstruction algorithms (Huelsenbeck et al. 2001) have recently been developed, advances which have enabled the implementation of realistic model of evolution. The supermatrix is thus currently considered as the method of choice for most studies focusing on animal relationships.

The second, more qualitative, phylogenomic approach mentioned at the start of this section relies on identification of genome-level features that could constitute molecular signatures valuable as informative characters. Some authors have coined the term "Rare Genomic Changes" to describe this new class of characters and they have emphasized their interest as a means of overcoming problems with large-scale sequence-based phylogenetic inference or at least as a mean of completing such studies (Rokas and Holland 2000a). Several kinds of genome features can provide valuable phylogenetic information, as depicted by the following examples. The structure of multigenic families and especially the loss and gain of genes have provided some very interesting clues for the resolution of some problematic nodes (Copley et al. 2004, Marlétaz et al. 2006). Alternatively, intron positions within homologous genes are broadly conserved in metazoans, as testified by the identification of conserved introns in the cnidarian Nematostella *vectensis* or the annelid *Platynereis dumerii* (Putnam et al. 2007, Raible et al. 2005). The pattern of intron conservation has thus been proposed to be suitable for phylogenetic inference (Roy and Gilbert 2005). Also, gene order and genome rearrangements have been extensively studied and have yielded very interesting findings, especially as far as mitochondrial gene arrangements are concerned (Boore 2006).

The growing amount of genomic data available for a large set of metazoans has made it possible to address the question of metazoan relationships using these different approaches and the contrasting results collected so far have made an important contribution towards renewing the debate about metazoan phylogeny.

4.4 Phylogenomics Resolves Animal Relationships

Early attempts to improve metazoan relationships using phylogenomic approaches often generated contrasting evidence and sometimes challenged the results of the "new view" of animal phylogeny. The settlement of these questions stressed the importance that should be given to taxonomic sampling.

4.4.1 Battle over the Coelomata and the Importance of Taxonomic Sampling

4.4.1.1 Early Phylogenomic Attempts Challenged the "New View"

The Ecdysozoa clade is the most controversial hypothesis of the "New View" of animal phylogeny and early multigene analyses of bilaterians therefore attempted to verify this hypothesis (Blair et al. 2002, Philip et al. 2005, Wolf et al. 2004). These studies employed from 100 to 780 nuclear genes which were analyzed following concatenation or independent approaches and they were based on species for which whole genome sequences were available (Blair et al. 2002, Philip et al. 2005, Wolf et al. 2004). Surprisingly, none of these studies recovered the ecdysozoan clade, indeed they all found that nematodes diverged prior to the radiation of coelomate animals. However, these results are thought to have been influenced by long-branch artefact because the genome of the nematode *C. elegans* exhibits very fast evolutionary rates (Aboobaker and Blaxter 2003). Indeed, these findings were challenged by the publication of the work of Philippe et al. (2005a, b), which successfully recovered the "new view" of animal phylogeny (Philippe et al. 2005a). By exploiting newly released EST data, these authors incorporated numerous new species belonging to groups such as the choanoflagellates, closest relatives of metazoans, the cnidarians, sister-group of the bilaterians, and a broad range of nematodes and platyhelminthes, some of which exhibit slower evolutionary rates (e.g. *Trichinella* for nematodes). The inclusion of those taxa made it possible to limit the phenomenon of long branch attraction – in phylogenetic jargon, to "break the long branches" – and thus produced support for the Ecdysozoa clade (Fig. 4.3). Notably, the platyhelminthes and the nematodes sometimes clustered together in this study, but the authors showed that the removal of fastest evolving genes enabled the recovery of the "new view" topology, with platyhelminthes being grouped with annelids and molluscs. This contradicts previous results, which estimated that gene sampling alone, and not taxon sampling, had the ability to increase the phylogenetic accuracy (Rokas and Carroll 2005).

4.4.1.2 Coelomata and the Interpretation of Rare Genomic Changes

The importance of dealing with accurate taxon sampling was also stressed in a recent debate about a possible revival of the Coelomata hypothesis. The origin of this debate was two studies that investigated two different qualitative genomic characters: "rare amino acid replacements" and patterns of intron conservation. Rare amino acid changes have been defined as amino acid substitutions that take place in a limited subset of taxa and that are caused by several nucleotide changes. Those substitutions are thus associated with a low probability of homoplasy, and 34 such positions were initially found to be in agreement with the Coelomata topology (Rogozin et al. 2007). However, subsequent addition of the cnidarian *N. vectensis* as

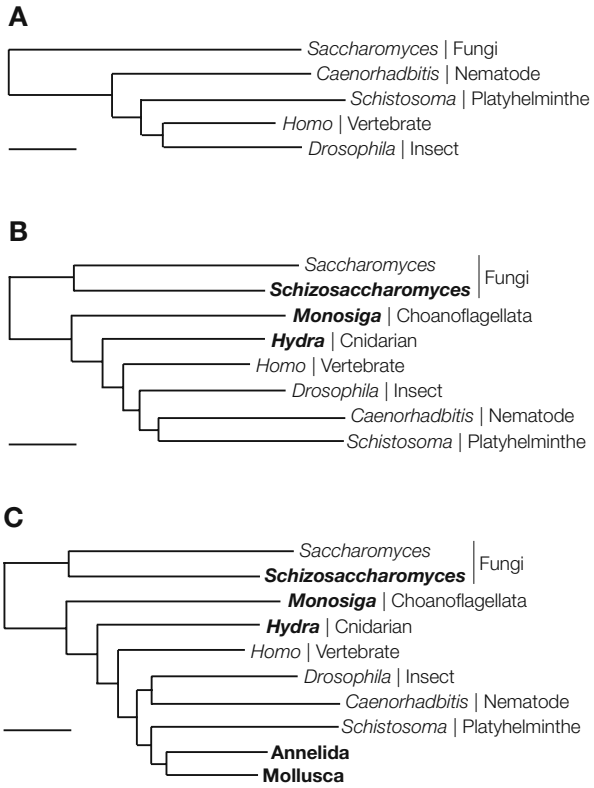

Fig. 4.3 The impact of taxonomic sampling and the demise of the *coelomata* hypothesis. These trees were inferred from 146 nuclear genes using Maximum likelihood inference. One branch length unit represents 0.1 substitutions per site. Progressive inclusion of intermediate taxa shows how long-branch attraction can be overcome. (**a**) The tree with limited taxon sampling supports the *coelomata* hypothesis with early divergence of nematodes and platyhelminthes. (**b**) The inclusion of intermediate bilaterian outgroups – *Hydra* and chaonoflagellates – results in nematodes and platyhelminthes being relocated back within the bilaterians but they still cluster together, in contradiction with the ecdysozoan hypothesis. (**c**) The selection of the least divergent marker genes (70 out of 146) and the addition of annelids and molluscs definitely recovers the "new view" of animal phylogeny (redrawn from Philippe et al. 2005b)

a metazoan outgroup in this dataset indicated that many of the signatures supporting the *coelomata* had been defined by using only non-metazoan outgroups such as plants and fungi. These signatures were thus likely to be ancestral among metazoans, since they were found to be shared between *Homo, Drosophila* and *Nematostella* (Irimia et al. 2007). In contrast, 13 rare amino acid changes support the Ecdysozoa clade when an accurate root is used.

Similarly, the pattern of intron conservation was employed as phylogenetic evidence and subtle methodological developments were carried out to account for the severe trend toward homoplasy in intron losses: an intron that is lost in a given

lineage has a high chance to be lost in another one (Zheng et al. 2007). Initially, support was recovered for Ecdysozoa (Roy and Gilbert 2005) but this result was subsequently criticized for not having sufficiently accounted for parallel intron losses, and contradictory analyses recovered support for Coelomata (Zheng et al. 2007). However, it has been suggested that this latter result was prone to the same rooting problems as the analyses of rare amino acid changes. Subsequent incorporation of the intron-rich cnidarian *N. vectensis* provided support for the Ecdysozoa again (Roy and Irimia 2008). Thus, although the importance of taxonomic sampling has been extensively discussed for conventional, sequence-based molecular phylogenies, there is clearly a need for similar attention to be paid to this problem for new types of phylogenetic evidence such as rare genomic changes (Rokas and Holland 2000b).

4.4.2 Is It Actually Possible to Decipher Animal Relationships?

A more general concern about the efficiency of phylogenomics has been raised by Rokas et al. (2003), who question its ability to fully resolve metazoan relationships (Rokas et al. 2005). To explain their failure to recover several classical metazoan clades using a 17 taxa and 50 gene dataset, the authors proposed that metazoan radiation was "compressed in time", which means that the time of divergence between animals lineages would have been too short for the deployment of an accurate phylogenetic signal. Despite its stimulating ideas, this work was contradicted by numerous other phylogenomic studies (Marlétaz et al. 2006, Matus et al. 2006, Philippe et al. 2005b). The problem of the lack of resolution observed by Rokas et al. (2005) was specifically addressed in subsequent work (Baurain et al. 2007) and two primary causes were identified: inappropriate taxon sampling and misleading evolutionary models. First, no effort was made to select slow evolving species within the diverging groups such as nematodes, although ecdysozoans were for example easily recovered when the nematode *Xiphinema* was employed instead of *Caenorhabditis* (Baurain et al. 2007). Also, it has been shown that the resolution of the tree may be improved by the use of better amino acid substitution models such as the CAT model or of better tree search algorithms such as SPR (subtree pruning and rebranching) (Hordijk and Gascuel 2005, Lartillot et al. 2007). Finally, the poor quality of raw sequence data could also hinder node resolution. Rokas et al. (2005) retrieved most of their sequences through PCR amplifications performed on conserved domains, thereby excluding the most variable and phylogenetically informative regions of the genes they analysed. Other phylogenomic studies of animal relationships have been mainly based on EST data that yield larger gene fragments.

The accuracy of phylogenomic reconstruction could be greatly enhanced by the selection of slowly diverging species, a strategy closely related to that proposed by Aguinaldo et al. for SSU-based phylogeny (Aguinaldo et al. 1997). However, in the multigene approach, it is not clear whether a single species possesses the least diverging copies of all marker genes. To overcome this problem, Marlétaz et al. (2006) have proposed the use of a new composite taxon strategy that selects the

least diverging copy of each selected gene from a pool of species representative of a monophyletic taxon (Marlétaz et al. 2006). This strategy allowed, for example, the building of a composite nematode taxon, whose branch is shorter than that of *Trichinella*, one of the slowest evolving nematodes.

The strength of the phylogenomic approach has been generally underlined by its ability to corroborate previous results based on multiple evidences. This suggests that the phylogenomic approach is a promising tool to address unsolved questions of animal relationships. However, the phylogenomic approach has also showed itself to be very sensitive to systematic errors. To set up an accurate analysis, it is thus necessary to carry out a careful examination of the taxonomic sampling, the inference methods used and the quality of the primary data.

4.5 Toward a Broad Phylogenomic Picture of Metazoan Relationships

The main limitation of single-gene phylogenies is that they are unable to handle some diverging taxa in terms of evolutionary rate or sequence composition. Such taxa are generally considered as *problematica* when morphological features do not allow this problem of phylogenetic affinity to be resolved. Among those taxa one can find what Libbie Hyman called "smaller coelomate groups", such as chaetognaths (Hyman 1959), and also the huge diversity of aschelminthes, from rotifers to kinorhynches, whose interpretation is one of the key questions of animal evolution (Jenner 2000). The possibility offered by phylogenomics to more accurately place these taxa in the metazoan tree has recently prompted a major sequencing effort, especially using EST methodology. The EST approach involves the sequencing of the extremity of a large set of clones from a cDNA library. The clustering of these transcript sequences can yield high quality partial transcriptome data. The lack of one or more markers in a given taxon was demonstrated not to prevent accurate reconstruction of the overall tree (Philippe et al. 2004) and EST collections of between 3,000 and 10,000 provide a sufficient set of markers to include a new taxon in a phylogenomic analysis (Marlétaz et al. 2008, Philippe and Telford 2006).

4.5.1 Challenging Well-Established Clades: The Case of Deuterostomes

Phylogenomics studies have lead to some longstanding phylogenetic hypotheses being reconsidered. The relationships between deuterostomes were the first to be challenged when urochordates were proposed to be the sister-group of vertebrates instead of cephalochordates (Delsuc et al. 2006). This grouping had strong statistical support obtained from the analysis of 146 nuclear genes and several key new taxa such as lampreys, hagfishes and the appendicularian *Oikopleura dioica.* This finding contradicts some well-established synapomorphies joining vertebrates

and the cephalochordate amphioxus, such as segmented myomeres (Schubert et al. 2006), but it is supported by some other features such as the discovery of migratory neural crest-like cells in tunicates (Jeffery et al. 2004). The close relationship between echinoderms and cephalochordates reported in this study was nevertheless surprising and required further confirmation. This confirmation was obtained by including the hemichordate taxon in the analysis: as a result, cephalochordates were branched back within the chordates and hemichordates were included with echinoderms in a new clade (called Ambulacraria) (Bourlat et al. 2006, Marlétaz et al. 2006) (Fig. 4.4). The new deuterostome status was completed by the addition of a new phylum: the xenoturbellids, which branches as sister-group of echinoderms and hemichordates (Bourlat et al. 2006).

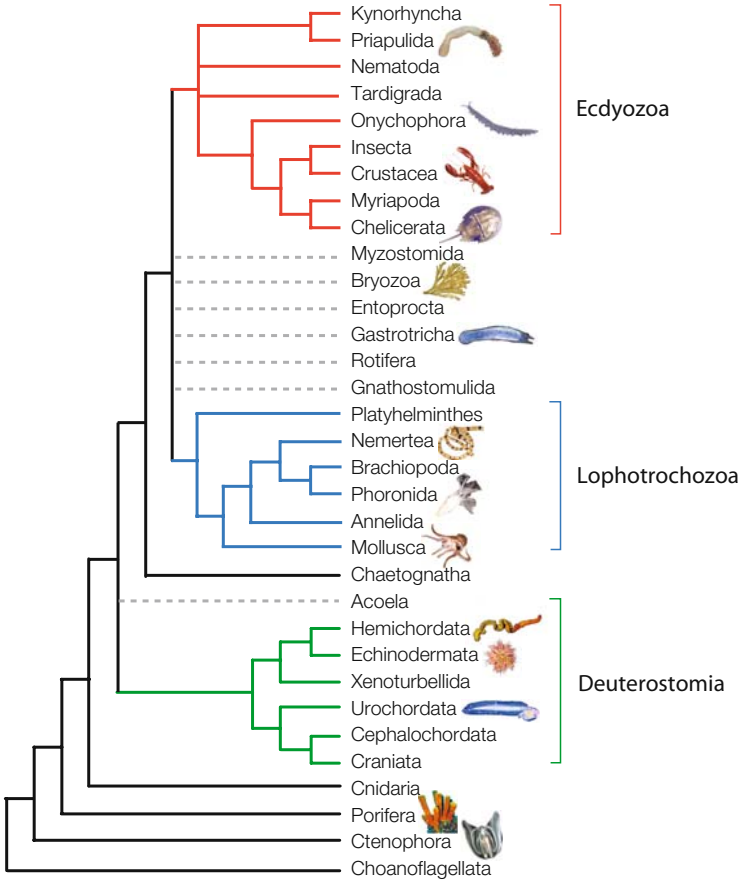

Fig. 4.4 Phylogenomic view of metazoan relationships. This tree summarizes the results of the most recent phylogenomic analyses focused on metazoans. All the nodes presented here are based on firm support values at least with the site-heterogenous CAT model. The overall "New View" of animal phylogeny is recovered with firm support for the deuterostomes, ecdysozoans and lophotrochozoans but a few clades such as chaetognaths, rotifers or acoel flatworms do not fit into this scheme. *Dashed* branches correspond to the most instable taxa whose position remains ambiguous

4.5.2 Chaetognaths Fit into the Bilaterian Tree

The phylogenomic approach has made it possible to more accurately determine the phylogenetic position of some groups that have represented longstanding problems. Chaetognaths have been puzzling for many years because their body plan exhibits several protostome features whereas their early development is typically deuterostomian, with enterocoelic formation of the coelom (Ball and Miller 2006). EST sequencing enabled phylogenomic analyses, which revealed that chaetognaths neither cluster within the ecdysozoans nor the lophotrochozoans and that they are probably a sister group of all the other protostomes (Marlétaz et al. 2008, Marlétaz et al. 2006). Their inclusion in protostomes had previously been supported by the analysis of their mitochondrial genome (Papillon et al. 2004). Alternative studies based on another chaetognath species also supported protostome affinities but rather argued for a basal position in lophotrochozoans (Matus et al. 2006) or alternatively, an unresolved position among protostomes (Dunn et al. 2008). Nevertheless, the basal branching of chaetognaths among protostomes was further supported by the recovery of an unusual molecular signature in chaetognath ESTs: the guanidinoacetate N-methyltransferase (GMT) gene was retrieved in all lineages predating the radiation of protostomes but in no current protostome group, which suggests that this gene was probably lost after the split of chaetognaths from protostomes (Marlétaz et al. 2008, Marlétaz et al. 2006). Despite alternative claims that chaetognaths should branch closer to lophotrochozoans, several studies independently support their position as a branching sister-group to the protostomes (Lartillot and Philippe 2008, Philippe et al. 2007). This peculiar phylogenetic position has two major implications. On one hand, this indicates that deuterostome-like embryology and especially enterocoelous formation of the coelom may be ancestral among bilaterians, a view which is reminiscent of the archecoelomate theory (Remane 1963) (Fig. 4.2). On the other hand, this constitutes the first clear contradiction of the "new view" of animal phylogeny since this phylum escapes the lophotrochozoan/ecdysozoan dichotomy (Adoutte et al. 2000).

4.5.3 Acoel Flatworms, Basal or Not?

Phylogenomics has also shed some light on the phylogenetic position of acoel flatworms, another highly debated topic since they were originally proposed to be the most basal bilaterian on the base of SSU analysis (Ruiz-Trillo et al. 1999). This phylogenetic position would suggest that the body organization of these triploblastic acoelomate flatworms could be reminiscent of those of bilaterian ancestor. Meanwhile, the status of acoels as an intermediate taxon has been extensively debated because their very fast evolutionary rates make long branch attraction artifacts possible (Deutsch 2008). Recently, a phylogenomic assessment of the position of acoels reached the conclusion that they are not platyhelminthes (Philippe et al. 2007). Their exclusion from protostomes is also supported by the occurrence of a

signature gene, the GMT enzyme, in this phylum. However, their position among bilaterians remained ambiguous since no clear phylogenetic signal was recovered for a precise branching of acoels among bilaterians (Fig. 4.4). The peculiar status of acoel flatworms and their exclusion from platyhelminths was then confirmed by the mean of phylogenomics.

4.5.4 Deeper into Protostome Relationships

The protostome clade presents the broadest diversity of phyla and body organizations and notably includes the most diverse kind of body organization (Adoutte et al. 2000). In particular, many minor protostome groups that were formerly placed within the aschelminthes display morphological features that are very hard to interpret. Moreover, they have been very difficult to position using classical molecular phylogenetics because of their fast evolutionary rates (e.g. Syndermates) (Passamaneck and Halanych 2006). Recently, an impressive sequencing effort was undertaken with the aim of solving this question of protostome relationships using the power of phylogenomics. It was expected that the resolution of lophotrochozoan phylogeny would be improved by the collection of EST data from several new phyla and from the most slowly diverging species of known phyla (e.g. platyhelminthes, molluscs) (Dunn et al. 2008, Hausdorf et al. 2007, Struck and Fisse 2008). The analysis of these new data allowed new relationships to be proposed within the lophotrochozoans: trochophore bearing animals, molluscs and an extended annelid clade that includes echiurans and sipunculids were found to be closely allied to nemerteans and lophophorates (Fig. 4.4) (Dunn et al. 2008). The nemerteans were surprisingly positioned as lophophorate sister-group. This last result corroborates the recent finding that the pilidium larva of palaeonemertean *Carinoma mutabilis* displays strong similarities with trochophore larvae (Maslakova et al. 2004). This clade of molluscs, annelids, lophophorates and nemerteans is supported by the presence of chitinous chaetae and may be corroborated by a palaetonlogical scenario that proposes a common origin for those chitinous chaetae and the calcareous spicules from which mollusc shell could have derived (Conway Morris and Peel 1995). Platyhelminthes are placed as sister-group of this lophotrochozoan assemblage, but the other phyla that were generally also associated with lophotrochozoans are more difficult to position accurately (Dunn et al. 2008). The bryozoa phylum (also called ectoprocta) and the entoprocts were proposed to cluster together, resurrecting an ancient hypothesis (Hausdorf et al. 2007), but this grouping was not recovered with different marker genes and taxonomic sampling. The position of these phyla remained thus very ambiguous, just as those of numerous other groups such as gastrotriches, rotifers or the enigmatic myzostomids, which exhibit strong “leaf instability”, a tendency to exhibit alternative branchings in the most likely trees and different bootstrap replicates (Dunn et al. 2008). The assembly of the protostome tree of life is thus far from achieved but these recent advances have demonstrated that increased taxon sampling leads to a strong improvement in phylogenetic resolution.

4.6 Conclusion: The Future of Animal Phylogeny

This chapter has described the impact of the most recent genome-based studies on the current view of metazoan relationships with respect to classical schemes based on morphology and single gene molecular phylogenies (mainly SSU rRNA). The new field of phylogenomics has been founded in the context of a strong debate about the status of the Coelomata clade, which finally ended up with the reassessment of the "New View" of animal phylogeny (Adoutte et al. 2000, Dunn et al. 2008). This debate mainly stressed the importance of dealing with extensive taxonomic sampling, even when examining qualitative molecular signatures or genome level characters. The extension of the phylogenomic approach has rephrased some longstanding questions in animal phylogeny such as the position of acoel flatworms, chaetognaths or relationships within the lophotrochozoan clade. It has also brought some surprising rearrangements, such as the positioning of the tunicates as vertebrate sister group (Delsuc et al. 2005). More surprisingly, the "new view" scheme was challenged by the branching of some phyla such as acoels or chaetognaths that do not fit within the ecdysozoan and lophotrochozoan clades. Recently the collection and analysis of a large amount of data from minor groups, described in a paper by Dunn et al. (2008), showed that that some phyla remain refractory to accurate positioning (Dunn et al. 2008). Several questions concerning metazoan tree of life thus remain open. For example, the exact relationships at the base of the metazoan tree remain elusive because of the difficulty of determining the branching order of bilaterians, poriferans, cnidarians and also ctenophores, all of which are groups that possibly diverge at the base of the metazoan radiation (Dunn et al. 2008). Answering this question would provide important insights into the nature of the metazoan ancestor: a complex organism that already possessed a mesoderm and bilateral symmetry, such as a ctenophore, or an organism of lesser complexity such as a sponge larva (Martindale and Henry 1999, Nielsen 2008)? Increased gene and taxonomic sampling as well as improvements to inference models are likely to be valuable tools for resolving these issues. Moreover, when the whole set of employable marker genes will have been exploited by molecular phylogeny, numerous other features such as gene order and synteny relationships might deserve further investigation (Philippe et al. 2005a).

The latter example of relationships at the base of metazoans emphasized the importance of phylogenetics as a means to understand the major evolutionary transitions of morphological characters. In particular, the power of molecular phylogenetics to ensure the orientation of characters has demonstrated that no general trends exist in the establishment of morphological complexity. The close association of nemerteans and annelids as well as that of vertebrates and tunicates has clearly shown that morphological simplifications have occurred repeatedly during the evolution of metazoans (Delsuc et al. 2005, Dunn et al. 2008). At a deeper evolutionary scale, the early divergence of ctenophores proposed by Dunn et al. (2008) suggested that the ancestor of all metazoans was far more complex than expected and that the morphological simplicity of poriferans could be secondarily derived.

However, the genomic processes that underlie such morphological switches remain to be uncovered. The relationship between the evolution of genomes and morphology constitutes one of the main current questions in evolutionary biology, especially because these two aspects of organisms are reputed to have undergone radically divergent modalities of evolution: strong selection for bodies versus neutralistic drift for genomes. The *Evodevo* field has recently uncovered the developmental genetics sources of macroevolutionary changes but these insightful studies remain limited to the molecular actors involved in specific genetic pathways (Muller 2007). On the other hand, the extensive study of both multigenic families and genome rearrangements that is underpinning phylogenomics could shed new light on developmental regulation at the genome scale, especially when it is associated with molecular genetics data acquired using model organisms. For instance, Domazet-Los et al. recently introduced the so-called "phylostratigraphic" approach involving comparison of the expression patterns and structures of multigenic families for a large number of genes involved in the development of germ layers (Domazet-Loso et al. 2007). They observed differences between the phylogenetic origins of all these genes and showed that those associated with the ectoderm, for example, are most likely more ancient. These promising findings plead for a renewed interest in neglected animal groups whose importance for comparative approaches and character orientation is undisputable. The extension of genome sequencing to some key organisms with key phylogenetic positions, paleontology and development, such as ctenophores and chaetognaths, constitutes the next step toward a better understanding of metazoan evolution.

References

Aboobaker AA, Blaxter ML (2003) Hox Gene Loss during Dynamic Evolution of the Nematode Cluster. Curr Biol 13:37–40

Adoutte A, Balavoine G, Lartillot N, Lespinet O, Prud'homme B, de Rosa R (2000) The new animal phylogeny: reliability and implications. Proc Natl Acad Sci U S A 97:4453–4456

Aguinaldo AM, Turbeville JM, Linford LS, Rivera MC, Garey JR, Raff RA, Lake JA (1997) Evidence for a clade of nematodes, arthropods and other moulting animals. Nature 387: 489–493

Anderson FE, Cordoba AJ, Thollesson M (2004) Bilaterian phylogeny based on analyses of a region of the sodium-potassium ATPase beta-subunit gene. J Mol Evol 58:252–268

Aristotle (1965) De Generatione animalium, tr. Arthur Platt, Clarendon Press, Oxford

Balavoine G, de Rosa R, Adoutte A (2002) Hox clusters and bilaterian phylogeny. Mol Phylogenet Evol 24:366–373

Ball EE, Miller DJ (2006) Phylogeny: the continuing classificatory conundrum of chaetognaths. Curr Biol 16:R593–R596

Barnes RD (1974) Invertebrate zoology. W.B. Saunders Company, Philadelphia.

Baurain D, Brinkmann H, Philippe H (2007) Lack of resolution in the animal phylogeny: closely spaced cladogeneses or undetected systematic errors?. Mol Biol Evol 24:6–9

Bergsten J (2005) A reviews of long-branch attraction. Cladistics 21:163–193

Blair JE, Ikeo K, Gojobori T, Hedges SB (2002) The evolutionary position of nematodes. BMC Evol Biol 2:7

Boore JL (2006) The use of genome-level characters for phylogenetic reconstruction. Trends Ecol Evol 21:439–446

Borchiellini C, Boury-Esnault N, Vacelet J, Le Parco Y (1998) Phylogenetic analysis of the Hsp70 sequences reveals the monophyly of Metazoa and specific phylogenetic relationships between animals and fungi. Mol Biol Evol 15:647–655

Borchiellini C, Manuel M, Alivon E, Boury-Esnault N, Vacelet J, Le Parco Y (2001) Sponge paraphyly and the origin of Metazoa. J Evol Biol 14:171–179

Bourlat SJ, Juliusdottir T, Lowe CJ, Freeman R, Aronowicz J, Kirschner M, Lander ES, Thorndyke M, Nakano H, Kohn AB, Heyland A, Moroz LL, Copley RR, Telford MJ (2006) Deuterostome phylogeny reveals monophyletic chordates and the new phylum Xenoturbellida. Nature 444:85–88

Collins AG (1998) Evaluating multiple alternative hypotheses for the origin of Bilateria: an analysis of 18S rRNA molecular evidence. Proc Natl Acad Sci U S A 95:15458–15463

Conway Morris S, Peel JS (1995) Articulated Halkieriids from the Lower Cambrian of North Greenland and their role in early protostome evolution. Philos Trans Biol Sci 347:305–358

Copley RR, Aloy P, Russell RB, Telford MJ (2004) Systematic searches for molecular synapomorphies in model metazoan genomes give some support for Ecdysozoa after accounting for the idiosyncrasies of Caenorhabditis elegans. Evol Dev 6:164–169

Cuvier G (1817) Le règne animal distribué selon son organisation, pour servir de base à l'hisoire naturelle des animaux et d'introduction à l'anatomie comparée. Deterville, Paris.

de Queiroz A, Gatesy J (2007) The supermatrix approach to systematics. Trends Ecol Evol 22: 34–41

de Rosa R, Grenier JK, Andreeva T, Cook CE, Adoutte A, Akam M, Carroll SB, Balavoine G (1999) Hox genes in brachiopods and priapulids and protostome evolution. Nature 399: 772–776

Delsuc F, Brinkmann H, Chourrout D, Philippe H (2006) Tunicates and not cephalochordates are the closest living relatives of vertebrates. Nature 439:965–968

Delsuc F, Brinkmann H, Philippe H (2005) Phylogenomics and the reconstruction of the tree of life. Nat Rev Genet 6:361–375

Deutsch JS (2008) Do acoels climb up the "Scale of Beings"?. Evol Dev 10:135–140

Domazet-Loso T, Brajkovic J, Tautz D (2007) A phylostratigraphy approach to uncover the genomic history of major adaptations in metazoan lineages. Trends Genet 23:533–539

Dunn CW, Hejnol A, Matus DQ, Pang K, Browne WE, Smith SA, Seaver E, Rouse GW, Obst M, Edgecombe GD, Sorensen MV, Haddock SH, Schmidt-Rhaesa A, Okusu A, Kristensen RM, Wheeler WC, Martindale MQ, Giribet G (2008) Broad phylogenomic sampling improves resolution of the animal tree of life. Nature 452:745-749

Eernisse DJ, Albert JS, Anderson FE (1992) Annelida and arthropoda are not sister taxa: a phylogenetic analysis of spiralian metazoan morphology. Syst Biol 41:305–330

Emig CC (1982) The biology of Phoronida. Adv Mar Biol 19:1–89

Felsenstein J (1978) Cases in which parsimony or compatibility methods will be positively misleading. Syst Zool 27:401–410

Field KG, Olsen GJ, Lane DJ, Giovannoni SJ, Ghiselin MT, Raff EC, Pace NR, Raff RA (1988) Molecular phylogeny of the animal kingdom. Science 239:748–753

Giribet G, Distel DL, Polz M, Sterrer W, Wheeler WC (2000) Triploblastic relationships with emphasis on the acoelomates and the position of Gnathostomulida, Cycliophora, Plathelminthes, and Chaetognatha: a combined approach of 18S rDNA sequences and morphology. Syst Biol 49:539–562

Gould SJ (1977) Ontogeny and phylogeny. Belknap/Harvard, Cambridge, MA.

Gould SJ (2002) The structure of evolutionary theory. Belknap/Harvard, Cambridge, MA.

Guindon S, Gascuel O (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol 52:696–704

Halanych K (2004) The New View of Animal Phylogeny. Annu Rev Ecol Evol Syst 35: 229–256

Halanych KM, Bacheller JD, Aguinaldo AM, Liva SM, Hillis DM, Lake JA (1995) Evidence from 18S ribosomal DNA that the lophophorates are protostome animals. Science 267: 1641–1643

Hausdorf B, Helmkampf M, Meyer A, Witek A, Herlyn H, Bruchhaus I, Hankeln T, Struck TH, Lieb B (2007) Spiralian phylogenomics supports the resurrection of Bryozoa comprising Ectoprocta and Entoprocta. Mol Biol Evol 24:2723–2729

Hennig W (1966) Phylogenetic systematics. University of Illinois Press, Urbana.

Hordijk W, Gascuel O (2005) Improving the efficiency of SPR moves in phylogenetic tree search methods based on maximum likelihood. Bioinformatics 21:4338–4347

Huelsenbeck JP, Ronquist F, Nielsen R, Bollback JP (2001) Bayesian inference of phylogeny and its impact on evolutionary biology. Science 294:2310–2314

Hyman LH (1940–1967) The invertebrates. McGraw-Hill, New York.

Hyman LH (1959) The invertebrates, Vol. 5. Smaller Coelomate groups. McGraw-Hill, New York

Irimia M, Maeso I, Penny D, Garcia-Fernandez J, Roy SW (2007) Rare coding sequence changes are consistent with Ecdysozoa, not Coelomata. Mol Biol Evol 24:1604–1607

Jeffery WR, Strickler AG, Yamamoto Y (2004) Migratory neural crest-like cells form body pigmentation in a urochordate embryo. Nature 431:696–699

Jeffroy O, Brinkmann H, Delsuc F, Philippe H (2006) Phylogenomics: the beginning of incongruence?. Trends Genet 22:225–231

Jenner RA (2000) Evolution of animal body plans: the role of metazoan phylogeny at the interface between pattern and process. Evol Dev 2:208–221

Jenner RA (2001) Bilaterian phylogeny and uncritical recycling of morphological data sets. Syst Biol 50:730–742

Jenner RA (2004) Libbie Henrietta Hyman (1888–1969): from developmental mechanics to the evolution of animal body plans. J Exp Zoolog B Mol Dev Evol 302:413–423

Kimura M (1983) The neutral theory of molecular evolution. Cambridge University Press, Cambridge.

Lake JA (1990) Origin of the Metazoa. Proc Natl Acad Sci U S A 87:763–766

Lartillot N, Brinkmann H, Philippe H (2007) Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model. BMC Evol Biol 7(Suppl 1):S4

Lartillot N, Philippe H (2004) A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. Mol Biol Evol 21:1095–1109

Lartillot N, Philippe H (2008) Improvement of molecular phylogenetic inference and the phylogeny of Bilateria. Philos Trans R Soc Lond B Biol Sci 363:1463–1472

Lemons D, McGinnis W (2006) Genomic evolution of Hox gene clusters. Science 313:1918–1922

Littlewood DT, Olson PD, Telford MJ, Herniou EA, Riutort M (2001) Elongation factor 1-alpha sequences alone do not assist in resolving the position of the acoela within the metazoa. Mol Biol Evol 18:437–442

Lynch M (2007) The origins of genome architecture. Sinauer, Sunderland.

Mallatt J, Winchell CJ (2002) Testing the new animal phylogeny: first use of combined large-subunit and small-subunit rRNA gene sequences to classify the protostomes. Mol Biol Evol 19:289–301

Marcus E (1958) On the evolution of the animal phyla. Quart Rev Biol 33:24–58

Marlétaz F, Gilles A, Caubit X, Perez Y, Dossat C, Samain S, Gyapay G, Wincker P, Le Parco Y (2008) Chaetognath transcriptome reveals ancestral and unique features among bilaterians. Genome Biol 9:R94

Marlétaz F, Martin E, Perez Y, Papillon D, Caubit X, Lowe CJ, Freeman B, Fasano L, Dossat C, Wincker P, Weissenbach J, Le Parco Y (2006) Chaetognath phylogenomics: a protostome with deuterostome-like development. Curr Biol 16:R

Martindale MQ (2005) The evolution of metazoan axial properties. Nat Rev Genet 6:917–927

Martindale MQ, Henry JQ (1999) Intracellular fate mapping in a basal metazoan, the ctenophore Mnemiopsis leidyi, reveals the origins of mesoderm and the existence of indeterminate cell lineages. Dev Biol 214:243–257

Maslakova SA, Martindale MQ, Norenburg JL (2004) Vestigial prototroch in a basal nemertean, Carinoma tremaphoros (Nemertea; Palaeonemertea). Evol Dev 6:219–226
Matus DQ, Copley RR, Dunn CW, Hejnol A, Eccleston H, Halanych KM, Martindale MQ, Telford MJ (2006) Broad taxon and gene sampling indicate that chaetognaths are protostomes. Curr Biol 16:R
Muller GB (2007) Evo-devo: extending the evolutionary synthesis. Nat Rev Genet 8:943–949
Nielsen C (2001) Animal Evolution: interelationships of the living phyla. Oxford University Press, New York.
Nielsen C (2008) Six major steps in animal evolution: are we derived sponge larvae?. Evol Dev 10:241–257
Papillon D, Perez Y, Caubit X, Le Parco Y (2004) Identification of chaetognaths as protostomes is supported by the analysis of their mitochondrial genome. Mol Biol Evol 21: 2122–2129
Papillon D, Perez Y, Caubit X, Le Parco Y (2006) Systematics of Chaetognatha under the light of molecular data, using duplicated ribosomal 18S DNA sequences. Mol Phylogenet Evol 38: 621–634
Passamaneck Y, Halanych KM (2006) Lophotrochozoan phylogeny assessed with LSU and SSU data: evidence of lophophorate polyphyly. Mol Phylogenet Evol 40:20–28
Peterson KJ, Eernisse DJ (2001) Animal phylogeny and the ancestry of bilaterians: inferences from morphology and 18S rDNA gene sequences. Evol Dev 3:170–205
Philip GK, Creevey CJ, McInerney JO (2005) The Opisthokonta and the Ecdysozoa may not be clades: stronger support for the grouping of plant and animal than for animal and fungi and stronger support for the Coelomata than Ecdysozoa. Mol Biol Evol 22:1175–1184
Philippe H, Brinkmann H, Martinez P, Riutort M, Baguna J (2007) Acoel flatworms are not platyhelminthes: evidence from phylogenomics. PLoS ONE 2:e717
Philippe H, Delsuc F, Brinkmann H, Lartillot N (2005a) Phylogenomics. Annu Rev Ecol Evol Syst 36:541–562
Philippe H, Lartillot N, Brinkmann H (2005b) Multigene analyses of bilaterian animals corroborate the monophyly of Ecdysozoa, Lophotrochozoa, and Protostomia. Mol Biol Evol 22: 1246–1253
Philippe H, Snell EA, Bapteste E, Lopez P, Holland PW, Casane D (2004) Phylogenomics of eukaryotes: impact of missing data on large alignments. Mol Biol Evol 21:1740–1752
Philippe H, Telford MJ (2006) Large-scale sequencing and the new animal phylogeny. Trends Ecol Evol 21:614–620
Putnam NH, Srivastava M, Hellsten U, Dirks B, Chapman J, Salamov A, Terry A, Shapiro H, Lindquist E, Kapitonov VV, Jurka J, Genikhovich G, Grigoriev IV, Lucas SM, Steele RE, Finnerty JR, Technau U, Martindale MQ, Rokhsar DS (2007) Sea anemone genome reveals ancestral eumetazoan gene repertoire and genomic organization. Science 317:86–94
Raible F, Tessmar-Raible K, Osoegawa K, Wincker P, Jubin C, Balavoine G, Ferrier D, Benes V, de Jong P, Weissenbach J, Bork P, Arendt D (2005) Vertebrate-type intron-rich genes in the marine annelid *Platynereis dumerilii*. Science 310:1325–1326
Remane A (1963) The enterocelic origin of the coelom. In: Dougherty EC (ed) The lower metazoa. University of California Press, Berkeley, CA, pp 78–90
Rogozin IB, Wolf YI, Carmel L, Koonin EV (2007) Ecdysozoan clade rejected by genome-wide analysis of rare amino acid replacements. Mol Biol Evol 24:1080–1090
Rokas A, Carroll SB (2005) More genes or more taxa? The relative contribution of gene number and taxon number to phylogenetic accuracy. Mol Biol Evol 22:1337–1344
Rokas A, Holland PW (2000a) Rare genomic changes as a tool for phylogenetics. Trends Ecol Evol 15:454–459
Rokas A, Holland PW (2000b) Rare genomic changes as a tool for phylogenetics. Trends Ecol Evol 15:454–459
Rokas A, Kruger D, Carroll SB (2005) Animal evolution and the molecular signature of radiations compressed in time. Science 310:1933–1938

Rokas A, Williams BL, King N, Carroll SB (2003) Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature 425:798–804
Roy SW, Gilbert W (2005) Resolution of a deep animal divergence by the pattern of intron conservation. Proc Natl Acad Sci U S A 102:4403–4408
Roy SW, Irimia M (2008) Rare genomic characters do not support Coelomata: intron loss/gain. Mol Biol Evol 25:620–623
Ruiz-Trillo I, Riutort M, Littlewood DT, Herniou EA, Baguna J (1999) Acoel flatworms: earliest extant bilaterian Metazoans, not members of Platyhelminthes. Science 283:1919–1923
Schubert M, Escriva H, Xavier-Neto J, Laudet V (2006) Amphioxus and tunicates as evolutionary model systems. Trends Ecol Evol 21:269–277
Scotland RW, Olmstead RG, Bennett JR (2003) Phylogeny reconstruction: the role of morphology. Syst Biol 52:539–548
Shubin N, Tabin C, Carroll S (1997) Fossils, genes and the evolution of animal limbs. Nature 388:639–648
Siewing R (1976) Probleme und neuere Erkenntnisse in der Großsystematik der Wirbellosen. Verh Dtsch Zool Ges 70:59–83
Stach T, Winter J, Bouquet JM, Chourrout D, Schnabel R (2008) Embryology of a planktonic tunicate reveals traces of sessility. Proc Natl Acad Sci U S A 105:7229–7234
Stamatakis A (2006) RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22:2688–2690
Struck TH, Fisse F (2008) Phylogenetic position of Nemertea derived from phylogenomic data. Mol Biol Evol 23:2058–2071
Swofford DL (1990) PAUP: Phylogenetic analysis using parsimony, Version 3.0. Illinois Natural History Survey, Champaign
Tautz D (2004) Segmentation. Dev Cell 7:301–312
Telford MJ, Holland PW (1993) The phylogenetic affinities of the chaetognaths: a molecular analysis. Mol Biol Evol 10:660–676
Valentine JW (1997) Cleavage patterns and the topology of the metazoan tree of life. Proc Natl Acad Sci U S A 94:8001–8005
Wagner GP (2007) The developmental genetics of homology. Nat Rev Genet 8:473–479
Wanninger A, Koop D, Bromham L, Noonan E, Degnan BM (2005) Nervous and muscle system development in Phascolion strombus (Sipuncula). Dev Genes Evol 215:509–518
Willmer P (1990) Invertebrates relationships: patterns in animal evolution. Cambrige University Press, Cambridge
Winchell CJ, Sullivan J, Cameron CB, Swalla BJ, Mallatt J (2002) Evaluating hypotheses of deuterostome phylogeny and chordate evolution with new LSU and SSU ribosomal DNA data. Mol Biol Evol 19:762–776
Wolf YI, Rogozin IB, Koonin EV (2004) Coelomata and not Ecdysozoa: evidence from genome-wide phylogenetic analysis. Genome Res 14:29–36
Zheng J, Rogozin IB, Koonin EV, Przytycka TM (2007) Support for the Coelomata clade of animals from a rigorous analysis of the pattern of intron conservation. Mol Biol Evol 24:2583–2592

Chapter 5
Metazoan Complexity

Florian Raible and Patrick R.H. Steinmetz

Abstract Evolution is often regarded as a process leading from simple ancestors to more complex descendants, a generalized view that has also impacted on different concepts of evolution. However, the study of new marine model systems, and the inclusion of new levels of analysis, challenge this paradigm, as they reveal that levels of complexity can diverge from the apparent organizational complexity of individual species. In this chapter, we analyze molecular genetic progress from different animal taxa, and how they help to determine the molecular changes associated with major evolutionary transitions, such as the transition to multicellularity or the origin of germ layers.

5.1 Approaches to Complexity

Haeckel's lithograph "Stammbaum des Menschen" ("Pedigree of Man") is a beautiful illustration of the concept of common evolutionary descent, and has widely been adapted for the representation of animal phylogeny. In this tree, animal evolution appears to progress steadily from "simple" life forms, such as "worms" ("vermes"), at the bottom, to more "elaborate" forms, with mammals forming the crown of the tree (Fig. 5.1). A similar trend can be found within the branches representing distinct animal phyla. For instance, "mud fish" are located close to the stem, whereas teleosts are found at the tips, consistent with the intuitive notion that teleosts are more "highly" evolved.

F. Raible (✉)
Max F. Perutz Laboratories, Campus Vienna Biocenter, University of Vienna, A-1030 Vienna, Austria
e-mail: florian.raible@mfpl.ac.at

P.R.H. Steinmetz (✉)
Department for Molecular Evolution and Development, University of Vienna, A-1090 Vienna, Austria
e-mail: patrick.steinmetz@univie.ac.at

J.M. Cock et al. (eds.), *Introduction to Marine Genomics*,
Advances in Marine Genomics 1, DOI 10.1007/978-90-481-8639-6_5,

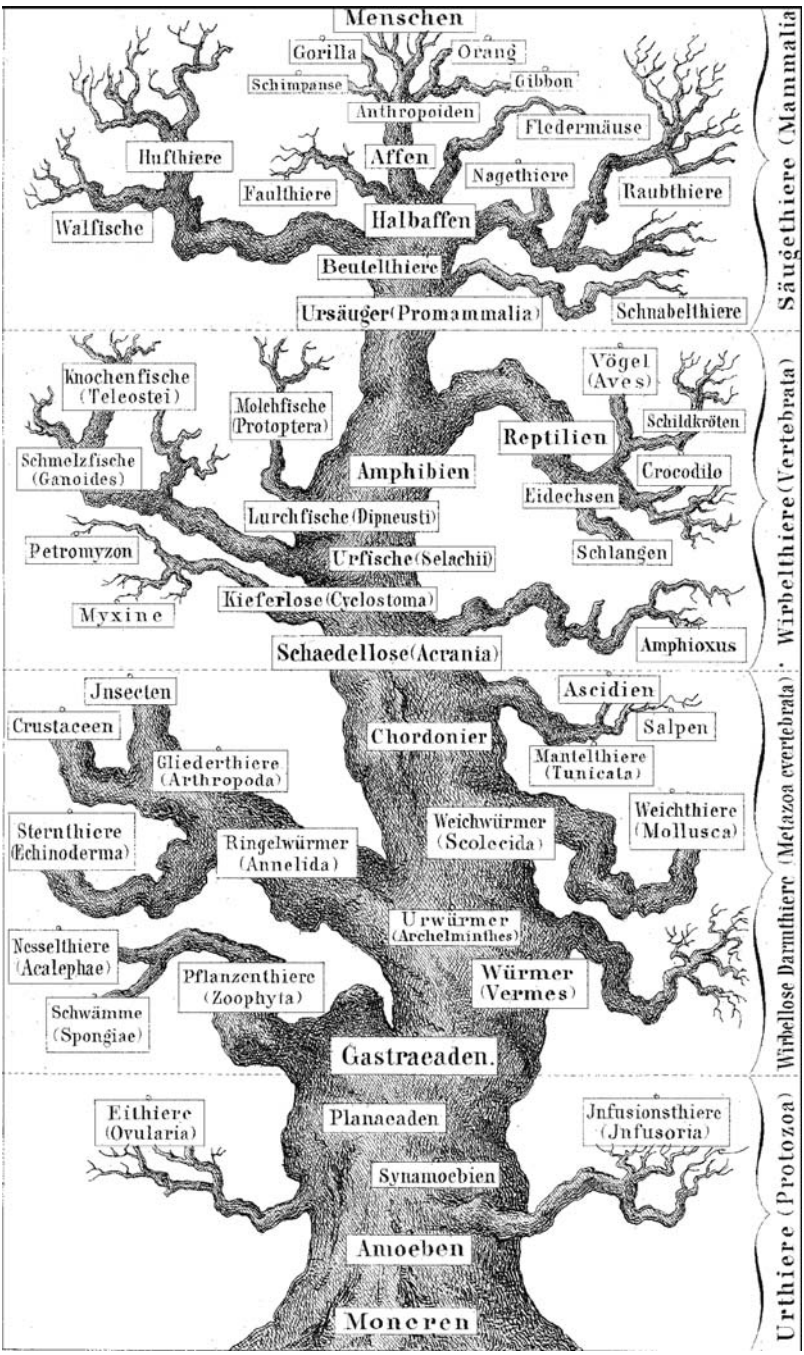

Fig. 5.1 Increasing complexity as a topological principle in Haeckel's "Stammbaum des Menschen" ("Pedigree of Man"). Animals with presumed simple organization are placed close to the bottom and stem of the tree, whereas more "elaborate" forms are found in the tips of the branches. Reproduced from (Haeckel 1903)

This notion of a steady gain of "complexity" in evolution has remained an influential template for the interpretation and evaluation of biological data on many different levels, ranging from genome characteristics to the number and characteristics of tissue types in an organism. The expectation that complexity steadily increases over time (and the often implicit assumption that primates represent the epitome of animal complexity) lets exceptions of this principle appear odd and counterintuitive. For instance, prior to the discovery of non-coding DNA, the term "c-value paradox" was used to describe the apparent lack of correlation between genome sizes and perceived complexity of animals (reviewed in Gregory 2005). The "primitive" lungfish *Protopterus aethiopicus* (Pedersen 1971) has a genome that is about 400 times larger than that of the "highly evolved" teleost genome of the puffer fish, *Tetraodon nigroviridis* (Jaillon et al. 2004), and about 40 times larger than the human genome.

After the c-value paradox was resolved by the discovery that these changes result mainly from the different amounts of non-coding DNA, the expectation was, in keeping with the notion of increasing complexity, that the amount of coding DNA (or the number of genes) would be higher in "complex" animals than in "simpler" ones. Cases like the duplication of the Hox cluster during chordate evolution, and the notion that this duplication was part of two rounds of whole-genome duplications (the "2R-hypothesis") lent support to the general notion that gene duplications were a major source of developmental innovation (reviewed in Taylor and Raes 2004). As we now know, however, the total excess of coding genes in the currently sequenced vertebrate genomes over those of the classical invertebrate models, *Caenorhabditis* and *Drosophila*, is surprisingly slim (Claverie 2001), and different vertebrates have roughly similar gene numbers. Moreover, important aspects of developmental patterning, such as the impact of the transcriptional regulator *pax6* on photosensitive structures (Halder et al. 1995), or the role of Hox genes in anteroposterior patterning (McGinnis et al. 1984), date back to early evolutionary times, providing evidence for an ancient core "developmental genetic toolkit" acting in animal development (see Cañestro et al. 2007). Considering that the connections between different components of this toolkit are of crucial importance for animal development, it is clear that a thorough analysis of animal complexity requires more than a simple quantification of genes or genetic functions.

On a more morphological level, animal complexity has been assessed by counting the number of morphologically distinct adult cell types (Bell 1997, Sempere et al. 2006, Valentine et al. 1994), sometimes complemented by the presence or absence of other morphological and embryonic characters (Aburomia et al. 2003, Heimberg et al. 2008). This method, however, has several pitfalls: as for some animals, more morphological data is available than for others, the complexity of well-analysed animals tends to be overestimated (Bonner 1988). In addition, morphological features fall short of describing other dimensions of complexity, such as the developmental and behavioural complexity of an animal, including life history strategies (larval or direct development) or the ontogeny of cell types (Bonner 1988, Valentine 2000, Valentine et al. 1994). Finally, morphological features are also difficult to quantify and weigh against each other, generating a need for more simple and quantitative measures of complexity.

Over the past few decades, molecular biology has provided many additional ways of assessing complexity. Beyond the global comparison of gene repertoires (counts of genes or gene families), the diversity of conserved gene families provides entry points into molecular features of different species. In an increasing number of species, such analyses can now be complemented by molecular techniques that address the expression dynamics and site of activity of individual genes, allowing researchers to characterise individual cell populations and tissues on the molecular level. Likewise, gene knockdown techniques are becoming available for more and more species, allowing more detailed analyses of the regulatory networks governing animal systems.

The fastest progress, however, has been made on the level of sequencing, and several whole-genome or transcriptome sequencing projects now provide a much better view of molecular evolution during animal diversification. In this chapter, we review recent progress in molecular analysis of different animal taxa and discuss what they reveal about the origin of complex features during animal evolution. A particular focus will be on the evolutionary changes that accompanied the origin and diversification of multicellular animals (Metazoa) from unicellular eukaryotic ancestors, as well as the differences in complexity found in different bilaterian groups.

5.2 Choanoflagellates: The Evolution of Multicellularity in Metazoa

Choanoflagellates form a phylum of flagellated, mostly unicellular organisms. They possess an apical flagellum surrounded by a bacteria- and detritus-catching collar of actin-rich microvilli. Their cell morphology and ability to form extracellular silicate spicules are very reminiscent of sponge choanocytes (Clark 1866, 1868), and recent molecular phylogeny studies have confirmed the phylogenetic position of choanoflagellates as a monophyletic sister group to all Metazoa (Carr et al. 2008, Clark 1868, Haeckel 1874, King et al. 2008, Shalchian-Tabrizi et al. 2008). Some extant choanoflagellate species can form colonies (e.g. *Proterospongia*) or differentiate into amoeboid cells and reproductive cysts (Bütschli 1883–1887, Leadbeater 1983, Siewing 1985). Hence, it is conceivable that the unicellular common ancestor of choanoflagellates and metazoans had the capacity of cell differentiation or colony formation and therefore exhibited primitive features of multicellularity (King et al. 2008, Lang et al. 2002). If true, extant colony-forming choanoflagellates might still resemble ancient, primitive multicellular life forms informative for the evolution of all animals on the metazoan line of evolution. Alternatively, colony formation and cell differentiation may have evolved independently in animals after the choanoflagellate-metazoan split. To distinguish between these two scenarios, searches were carried out for metazoan genes important in multicellular processes such as cell adhesion, cell–cell-communication and the division of labour in differentiated cell types in the genomes of the strictly unicellular choanoflagellate *Monosiga brevicollis* (using EST and genome data) and the colony-forming

Proterospongia-like species (using EST data) (Abedin and King 2008, King et al. 2003, 2008).

The most recent *Monosiga brevicollis* genome annotation (April 2008) listed 78 protein domains that are shared exclusively with metazoans but are absent from plants, fungi or slime molds (although some resemble bacterial protein domains) underscoring the close relatedness of choanoflagellates and metazoans (King et al. 2008). These domains are found in metazoan cell adhesion proteins (e.g. extracellular cadherin, sugar-binding C-type lectins, immunoglobulin, and integrin α domains) and extra-cellular matrix (ECM) components (fibronectins) including basement membrane elements (several collagen types and laminins). The respective functional domains therefore originated before the choanoflagellate-metazoan split, although many occur in choanoflagellate-specific combinations, as found for extracellular cadherin domains (ECDs)-containing proteins. However, some ECDs are conserved in Fat-type cadherins (conserved in sponges, cnidarians and bilaterians) and in a Hedgehog-related protein so far restricted to sponges and cnidarians (*hedgling*, see also below and Fig. 5.2) (Abedin and King 2008, King et al. 2008). Notably, "classical" metazoan-type cadherins with a highly conserved cadherin cytoplasmic domain have not been found in *M. brevicollis* (Abedin and King 2008). Also integrin β, a metazoan-specific cell adhesion receptor domain, and the ECM component laminin B(IV) are absent (King et al. 2008).

Whereas *M. brevicollis* possesses a surprisingly rich repertoire of metazoan cell adhesion and ECM domains, most of the intracellular signalling cascades associated with metazoan cell–cell communication are missing or highly divergent in choanoflagellates. Clear Wnt, TGF-β ligands or nuclear hormone receptor orthologues, present as large families in metazoans, are missing from the *M. brevicollis* genome, while other animal signalling pathways (JAK/STAT, Hedgehog, Delta/Notch) are incomplete. For some multidomain signalling pathway components (e.g. Notch or Hedgehog proteins), single domains are encoded in the *M. brevicollis* genome, but as for cell adhesion proteins, mostly not in metazoan-characteristic combinations. As for tyrosine kinase signal transduction pathways, *M. brevicollis* exhibits the largest number of tyrosine kinases and receptors, regulatory phosphatases and phospho-tyrosine-binding SH2 domain proteins (signal transducers) so far discovered in a single species. However, they appear to be largely divergent from metazoan members of the tyrosine kinase pathway (King et al. 2003, 2008, Manning et al. 2008). This is evident by the lack of clear orthologues, differences in regulation of tyrosine-kinase signalling and a large set of choanoflagellate tyrosine-kinase domain combinations not found in metazoan proteins (King et al. 2008, Manning et al. 2008, Pincus et al. 2008, Segawa et al. 2006). An exception is the conserved combination of intracellular tyrosine kinase domains and cytoplasmic SH2 domains with extracellular cadherin and EGF domains (in e.g. *hedgling*) suggesting that tyrosine-kinase signaling can relay extracellular signals (Abedin and King 2008, King et al. 2003). Functional assays have linked tyrosine kinase signalling to the regulation of the cell cycle by monitoring external food availability (King et al. 2003).

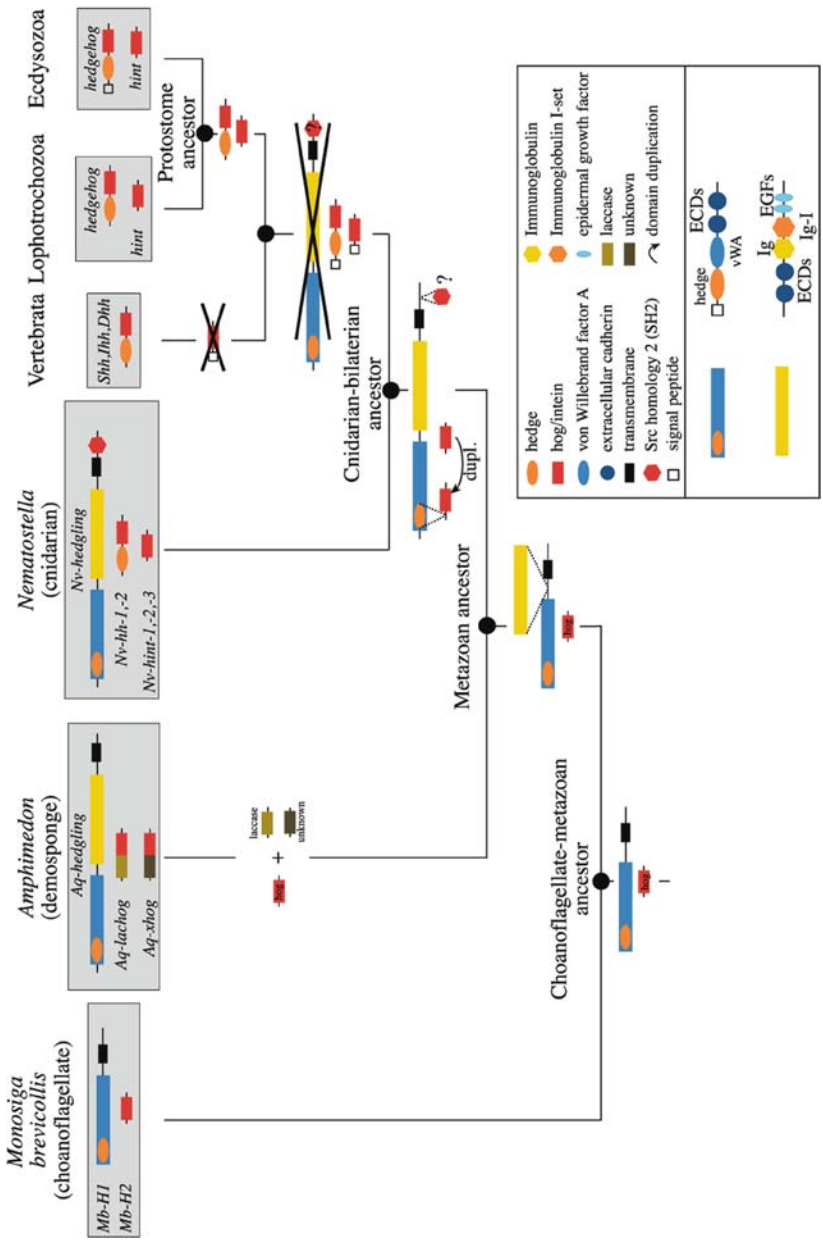

Fig. 5.2 Domain evolution. Scenario of Hedgehog domain evolution by domain shuffling, duplication and loss during the evolution of metazoans (based on Adamska et al. 2007b, King et al. 2008, Matus et al. 2008, Snell et al. 2006). Schematics at roots depict protein and domain arrangements present in the common ancestor. Schematics within terminal branches depict putative rearrangements not present in the ancestor (derived). Grey boxes highlight protein orthologues present in extant organisms

The importance of domain rearrangements during the evolution of metazoan proteins is well illustrated by the evolution of Hedgehog ligands (Fig. 5.2). Bilaterian Hedgehog proteins consist of two functional domains: An N-terminal diffusible receptor ligand domain, and a C-terminal autocatalytic domain (Bijlsma et al. 2004). Both domains occur in *M. brevicollis*, but are encoded by separate genes (King et al. 2008, Snell et al. 2006). Whereas the N-terminal Hedge domain forms part of a multidomain protein conserved only in sponges and cnidarians (named Hedgling), the C-terminal Hog/Intein domain occurs in a single domain protein in *M. brevicollis* (Adamska et al. 2007b, King et al. 2008, Matus et al. 2008). In *M. ovata*, not directly related to *M.* brevicollis, the C-terminal domain occurs on one protein (Hoglet) together with putative cellulose-binding domains (Carr et al. 2008, Snell et al. 2006). If separate Hedgehog domains represent an ancestral state, bilaterian Hedgehog proteins are the result of specific domain duplications and shuffling prior to the sponge/eumetazoan and the cnidarian/bilaterian split (Fig. 5.2). Accordingly, Hedgling, the older of the two Hedge-containing proteins in the cnidarian/bilaterian ancestor, was lost in early bilaterians while only the *bona fide* Hedgehog was maintained.

Whereas extracellular and signalling domains are indicative of a considerable interactive capacity of choanoflagellates, the repertoire of transcription factors reflects the complexity of regulatory networks and cell type differentiation. The *Monosiga* genome contains all major, ubiquitous classes of transcription factor motifs (e.g. zinc-finger, homeobox or helix-loop-helix proteins) (King et al. 2008). The few, previously metazoan-specific transcription factors found in *Monosiga* (e.g. p53, Myc) have rather general roles in cell cycle or transcriptional control (Nedelcu and Tan 2007). The existence of only two homeobox genes clustering specifically with Meis/Prep/TGIF genes (TALE superclass), but not with other TALE class genes (e.g. *Iroqois*) (King et al. 2008) indicates the loss of at least the Iroqois-related genes and non-TALE homeobox factors as these predate the emergence of choanoflagellates (Derelle et al. 2007, Mukherjee and Bürglin 2007). Also, *M. brevicollis* lacks metazoan-specific Ets, Hox, POU or T-box families (King et al. 2008).

In conclusion, the conservation of many metazoan-specific protein domains associated with cell adhesion and ECM-interacting proteins are indicative of a surprisingly high capacity of choanoflagellates to interact among themselves and with the environment, reflected in some choanoflagellate species by colony formation or settlement (Leadbeater 1983, Siewing 1985). However, the absence or fragmentary presence of metazoan signalling cascade proteins suggests a rather restricted ability to communicate between cells in a manner similar to metazoans. As shown for tyrosine-kinase proteins, the present signalling domains appear to function in adapting intracellular processes to changing environmental conditions rather than in cell–cell interactions. Although the majority of conserved cell adhesion and signalling domains are present, their unique combinations within multidomain proteins makes it difficult to deduce their ancestral functions in the choanoflagellate-metazoan ancestor. Extensive rearrangements of protein domains might have led to novel functions on the evolutionary lines towards metazoans and choanoflagellates. In the few cases where clear orthology between proteins of *M. brevicollis* or

the colony-forming *Proterospongia*-like species and metazoans (e.g. Fat-type cadherin proteins) could be established, functional analysis has helped to discriminate between a "primitive metazoan"-type function in cell–cell recognition and/or adhesion during colony formation, or a "unicellular"-type function prior to the evolution of multicellularity (e.g. predation and phagocytosis of bacteria, detection of environmental influences). In the former case, those domains evolved divergent functions in *Monosiga* after the loss of colony formation. In the latter case, the evolution of protein domains absent in choanoflagellates (e.g. integrin β) and the evolution of novel proteins by rearranging pre-existing domains (into e.g. "classical"-type cadherins) were crucial steps on the way to multicellularity.

As far as cell type diversity is concerned, the major absence of metazoan-specific transcription factor families correlates with the low potential for cell differentiation in choanoflagellates. However, the secondary loss of homeobox genes shows that in the course of choanoflagellate evolution, transcription factor complexity and cell type diversity could have undergone secondary reduction. Extending molecular analyses to a larger number of choanoflagellate species, and also to filasterean and ichtyosporean choanozoans (e.g. *Ministeria*, *Capsaspora* or *Sphaeroforma*) that form outgroups to the Choanoflagellate + Metazoa group, will clarify the pattern of gene gains and losses during the evolution of choanoflagellates and metazoans.

5.3 Sponges: The Evolution of Animal Development, Body Axis, Cell Types and Epithelia

Sponges are mostly marine, sessile and filter-feeding animals that are widely regarded as the most "simple" animal phylum. Their body plan consists only of two main epithelial layers: the single- or double-layered pinacoderm forming a protective outer and partially inner layer, and the choanoderm, built of ciliated collar cells (choanocytes), that produce a water flow entering through pores (ostia) and exiting via an larger opening (osculum). In between the pinaco- and the choanoderm lies the mesohyl, connective tissue filled with amoeboid cells and extracellular matrix. The choanocytes either cover most of the inner body layer or lie within a system of internal chambers. The apparent simplicity of sponges is also reflected on the level of cell types: demosponges are thought to have only a dozen histologically distinguishable adult cell types among which are skeleton-forming sclerocytes or contractile myocytes (Siewing 1985). Additional cell types may be larval-specific (e.g. primitive photoreceptor cells, see below) or further distinguishable by specific vesicle forms and contents. Sponges lack true muscle cells and neurons.

Traditionally, the phylum Porifera is subdivided into glass sponges (Hexactinellida), demosponges (Demospongiae) and calcareous sponges (Calcarea) all having evolved from a common ancestor (monophyly) (Philippe et al. 2009). Alternative models suggest that homoscleromorphs (traditionally demosponges), Calcaraea and/or Hexaxtinellida might be paraphyletic sponge groups more closely related to other metazoans than to the other sponges (Borchiellini et al. 2004, 2001, Haen et al. 2007, Sperling and Peterson 2007) (for a more extensive discussion, see Chapter 4). In any case, sponges are regarded as the oldest extant metazoan group

(see Chapter 4 and below for a discussion on the position of *Trichoplax*) and the first animals with embryonic and larval development. Comparison of cell adhesion and ECM-interacting proteins of different sponge groups with choanoflagellates and eumetazoans has helped to identify crucial events during the evolution of animal multicellularity. In addition, comparative study of developmentally relevant transcription factors and signalling molecules has provided a deeper understanding of the early evolution of embryonic and larval development including germ layer separation, axis specification and embryonic differentiation.

As described in the introduction, animal complexity is often correlated with cell type diversity. As cell types are specified by characteristic combinations of transcription factors (the "molecular fingerprint") (Arendt 2005), the expansion of transcription factor families might be a proxy for cell type diversification (Vogel and Chothia 2006). This hypothesis is being tested by determining the repertoire of transcription factor subclasses in extant sponges cnidarians and bilaterians to reconstruct their emergence during animal evolution. Already in the demosponge reference species *Amphimedon* all major metazoan transcription factor classes are present. Phylogenetic analyses have explored in detail the relationship of the *Amphimedon* bHLH, Fox, Sox, T-box, Paired-like, Antennapedia, NK, TALE, Six, POU and LIM homeodomain proteins with their correlates in other animals (Adell et al. 2003, Adell and Müller 2004, 2005, Jager et al. 2005, Larroux et al. 2007, 2006, 2008, Manuel and Le Parco 2000, Manuel et al. 2004, Simionato et al. 2007). This approach has proved useful to date the extent of gene families at the base of metazoan evolution. Some *Amphimedon* genes are equally related to several different bilaterian families (Fig. 5.3a, e.g. ARNT/bmal orthologs or NK2/4 orthologs) (Larroux et al. 2007, Peterson and Sperling 2007, Simionato et al. 2007), meaning that the sponge gene represents an ancestral gene that duplicated and founded novel families during eumetazoan evolution. Other *Amphimedon* genes clearly group with eumetazoan members of single gene families (Fig. 5.3b; e.g. *Amq3*/Myc,

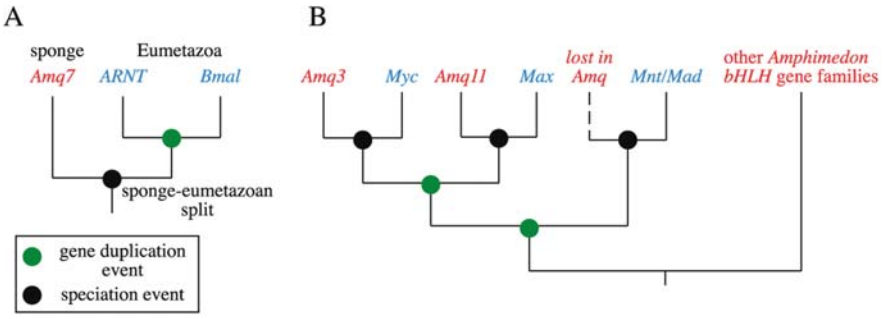

Fig. 5.3 Evolution of gene families. Evolution of gene families by gene duplication and speciation events (based on Simionato et al. 2007). (**a**) Evolution of the eumetazoan bHLH genes *ARNT* and *bmal* by gene duplication from a single ancestor present in the last sponge-eumetazoan ancestor. (**b**) Evolution of the bHLH genes *myc*, *max* and *mnt*/*mad* by gene duplication prior to the eumetazoan/sponge split implies that the absence of *mnt*/*mad* in *Amphimedon* is due to secondary loss. This is confirmed by the presence of a *mnt*/*mad* ortholog in the homoscleromorph sponge *Oscarella*. See text for further details. *Blue*: eumetazoan orthologs; *red*: *Amphimedon* orthologs

Amq11/*max* or the *Suberites brachyury*) (Larroux et al. 2008, Simionato et al. 2007), indicating that the respective gene families had already diversified in the poriferan-eumetazoan ancestor. In cases where these families group monophyletically with families that lack a clear *Amphimedon* orthologue, gene loss in the sponge can be assumed. Similarly, analysis of the NK homebox genes indicated loss of several NK gene families in demosponges (Peterson and Sperling 2007). Direct evidence for these gene losses comes from comparisons with other sponge species where the respective genes are present, examples including *brachyury* (present in *Suberites*, missing in *Amphimedon*) (Larroux et al. 2008) and *mnt*/*mad* (present in *Oscarella*, missing in *Amphimedon;* Fig. 5.3b) (Simionato et al. 2007). As sequence data is so far scarce for other sponge genomes, the real number of transcription factors in the poriferan-eumetazoan ancestor is probably higher than currently assumed.

The presence of representatives from almost all major eumetazoan transcription factor families is surprising, considering the relatively low number of adult cell types and the simplicity of the sponge body plan. This paradox might be resolved by considering that the life cycle of most sponges includes a ciliated larval stage followed by a complex metamorphosis (Leys and Ereskovsky 2006). The high variability of early embryonic development has led to currently opposing views on the presence and definition of germ layers and gastrulation in sponges (Ereskovsky and Dondua 2006, Leys 2004). However, in contrast to adults, many larvae are clearly bilayered and exhibit a single body axis (Ereskovsky and Dondua 2006, Leys and Ereskovsky 2006). Expression analysis suggests that molecular patterning along the axis in the *Amphimedon* larva is mediated by Wnt and TGF-β members (Adamska et al. 2007a). The axis is directly related to swimming direction and can be influenced by phototactic cues (Leys and Degnan 2001). Although lacking neurons and ciliary or rhabdomeric photoreceptors, some photoresponsive larvae possess pigmented cells with a long cilium that can simultaneously act as receptor and effector and could therefore represent the proto-photoreceptor cell before the specialisation into distinct pigment shading and photoreceptor cells (Arendt 2008, Leys and Degnan 2001). An almost complete set of bilaterian post-synaptic proteins in the *Amphimedon* genome also indicates that despite the lack of neurons, synapses or post-synaptic densities, a simple system of sensory stimuli transduction might be in place (Sakaraya 2007). Further functional, physiological and expression analysis will elucidate if the number of sponge cell types is higher than currently assumed. In fact, what is classified as one single cell type by morphological criteria might consist of several, functionally different cell types distinguishable only by their differential molecular composition.

5.4 The Placozoan *Trichoplax*: A Primitively Simple or Highly Reduced Metazoan?

Until recently, the phylum Placozoa consisted only of *Trichoplax adhaerens* and *Treptoplax reptans* (only described once) but molecular analysis have discovered additional cryptic (morphologically indistinguishable) species (Grell 1971b, Voigt

et al. 2004). So far, only asexual reproduction by fission or budding of spherical "swarmers" has been observed (Siewing 1985). The discovery of putative oocytes, cleaving stages up to the 64-cell stage (Grell 1971a, 1972, Grell and Ruthman 1991) and molecular signatures for recombination and sex (Signorovitch et al. 2005) are indications for the existence of sexual reproduction. Hence, the description of the *Trichoplax* life cycle is most likely incomplete and an intermediate parasitic stage in a so far unidentified host or a larval stage cannot be excluded (Miller and Ball 2005). *Trichoplax* exhibits a unique mode of algal feeding ("transepithelial cytophagy") through gaps of the upper, monocliliated epithelium and subsequent phagocytosis by inner fiber cells (Wenderoth 1986). In addition, the ventral, non-ciliated epithelium probably secretes digestive enzymes but is not able to perform phagocytosis.

Presenting four somatic cell types within three cell layers, a single, "top-bottom" axis and lacking a basement membrane, extracellular matrix, mouth, gut, and nervous system, the body plan of *Trichoplax* is often considered as simplest throughout metazoans (Syed and Schierwater 2002). The cell layers consist of an upper and lower epithelium separated by inner, contractile "fibre cells" probably controlling the amoeboid-like locomotion (Syed and Schierwater 2002). In the context of this simple axial organization, it is surprising to find a complete set of components required for functional Wnt or TGF-β signaling (conserved during bilaterian body axis patterning) in the *Trichoplax* genome (Srivastava et al. 2008).

Phylogenetic analysis based on a large dataset from the sequenced genomes of *Trichoplax* and other metazoans (Srivastava et al. 2008) favours the emergence of Placozoa after the poriferan split (Borchiellini et al. 2001, Collins 1998, da Silva et al. 2007) over a scenario based on mitochondrial sequences that groups Placozoa with sponges and cnidarians as a monophyletic sister phylum to bilaterians (Dellaporta et al. 2006, Haen et al. 2007, Signorovitch et al. 2007). At present, the position of placozoans as a sister group to all other metazoans is not supported by any molecular phylogeny. Therefore, the apparent morphological simplicity of placozoans evolved probably by secondary reduction. As an extracellular matrix and basement membranes are present in homoscleromorph sponges, these features were secondarily lost in placozoans.

The ANTP homeobox gene repertoire also indicates secondary losses in placozoans that might underlie morphological simplification. Genome sequencing (Schierwater et al. 2008, Srivastava et al. 2008) identified some ANTP homeobox genes in addition to the previously described members of the Hox/ParaHox (*Trox-2*) (Jakob et al. 2004), NK-like (*Dlx*, *Hmx*, *Not*) (Martinelli and Spring 2004, Monteiro et al. 2006) and Extended Hox (*Mnx)* (Monteiro et al. 2006) families. Altogether, however, the number of ANTP genes remains relatively low. Evidence for ANTP gene losses includes the absence of some NK genes (*msx, barH*), which are present in sponges (Schierwater et al. 2008), and the assignment of single members of large ANTP sublasses to distinct bilaterian/cnidarian gene families (e.g. *trox2* as a *cnox*/*gsx* ortholog within Hox/ParaHox subclass) (Monteiro et al. 2006, Schierwater et al. 2008). Either the distinct family is a founder family of the entire subclass (improbable in the case of *cnox*/*gsx*), or, as already proposed for sponges (see previous section), an extensive Hox/ParaHox genes loss occurred

during the evolution of Placozoa (Jakob et al. 2004, Monteiro et al. 2006, Peterson and Sperling 2007).

Many *Trichoplax* transcription factors (e.g. the T-box genes *Tbx2/3* and *brachyury*, or the paired box gene *TriPaxB*, a putative precursor gene of the sensory cell marker genes PaxA/B/C (Cnidaria), Pax2/5/8 and Pax4/6 (Bilateria)) are expressed at the outer border of the animal, similar to the expression of RF-amide, an abundant cnidarian neuropeptide (Martinelli and Spring 2003, Schuchert 1993). This region includes many cells with unfamiliar morphologies that may represent ancestral neural or multipotent cell types that have diversified into different cell types during the evolution of cnidarians and bilaterians (Jakob et al. 2004, Martinelli and Spring 2003). The existence of putative neuronal precursor cells is further supported by the presence of basic components for neurotransmitter synthesis, release, and uptake as well as for synapse formation, photoreception and the electric transmission of stimuli in the *Trichoplax* genome.

Although placozoans are in some aspects clearly reduced, the structure of the small *Trichoplax* genome (98 million base pairs) rather represents an ancestral state. The high level of conserved linkage (synteny) between large *Trichoplax* and vertebrate genomic regions, retention of ancient introns, and a high conservation of intron-exon-boundaries oppose a secondary genomic reduction as found in the *C. elegans*, *Drosophila*, or *Oikopleura* genomes. A better understanding of the *Trichoplax* life cycle and cell type morphology will help elucidate whether cryptic developmental stages or so far undetected cell type diversity can account for the relative complexity of the genome in terms of gene content and structure.

5.5 Cnidaria: A Simple Body with a Complex Genome

Cnidaria form a species-rich phylum of mainly marine animals, comprising the classes Anthozoa (e.g. sea anemones and corals), Cubozoa (box jellyfish), Scyphozoa (e.g. sea wasps) and Hydrozoa (e.g. *Hydra*) (Siewing 1985). Many anthozoans and hydrozoans have complicated life cycles involving a ciliated planula larva (Nielsen 2001). In addition, all groups except the anthozoans have a free-swimming medusa stage. Although the medusa is mainly considered to be a secondary innovation, arising after the emergence of Anthozoa, secondary loss in anthozoans is equally probable (Collins 2002). In contrast to their rather complex life cycles, cnidarians display a simple body plan with two germ layers: ectoderm and endoderm, separated by the acellular mesogloea consisting of extracellular matrix. Cnidarians have a single body opening derived from the blastopore that functions as both mouth and anus. At the opposite side of the planula resides a ciliary apical tuft with presumptive sensory functions. Although simple radial symmetry is widespread among cnidarians, many anthozoans possess a long-known second body axis (the “directive axis”) that runs orthogonal to the oral-aboral axis and defines bilateral symmetry (Stephenson 1928). It is morphologically apparent by the arrangement of muscles within endodermal folds (mesenteries) and by the

positioning of longer cilia at either end of the slit-like pharynx (Siewing 1985, Stephenson 1935). The homology of cnidarian and bilaterian axes is discussed in a later section.

Overall, the cnidarian nervous system is considered to be simple and diffuse (Bullock and Horridge 1965, Siewing 1985). However, a higher axonal concentration is found in the nerve ring of many medusae, and the rhopalia, a sensory structure containing lens eyes and statocysts in the medusae of Cubozoa and Scyphozoa (Nilsson et al. 2005, Piatigorsky and Kozmik 2004, Skogh et al. 2006). Molecular studies on neurogenesis and the diversity and specification of neural cell types in Cnidaria were mostly restricted to the freshwater polyp *Hydra* until recent work on *Nematostella* started to shed light on nervous system development in anthozoan larvae (Marlow et al. 2009). More specific analyses of neural structure development focused mainly on photosensory systems (Kozmik et al. 2003, Stierwald et al. 2004, Suga et al. 2008) and the patterning of the apical organ (Matus et al. 2007, Pang et al. 2004, Rentzsch et al. 2008). We believe that a better molecular understanding of the development of the nervous system and the diversity of neural cell types in cnidarians is the key to understanding the evolutionary innovations that led to the complex nervous systems of many bilaterians.

Interest in the use of cnidarians for non-bilaterian developmental and genomic studies is mainly due to the success of the freshwater hydrozoan *Hydra vulgaris* as a model organism for axial patterning and stem cell research. In addition, the last decade has seen the emergence of anthozoans (the brackish-water sea anemone *Nematostella vectensis*, and the marine coral *Acropora millipora*), cubozoans (*Tripedalia cystophora*) and hydrozoans (the marine *Hydractinia echinata*, *Clytia hemisphaerica*, *Podocoryne carnea*) for comparative genomic and developmental studies.

5.5.1 The Nematostella *Genome*

The sequencing of the *Nematostella* genome has confirmed previous assumptions based on *Nematostella* and *Acropora* EST analyses that the anthozoan genome is more complex than some bilaterian genomes (Kortschak et al. 2003, Miller and Ball 2008, Miller et al. 2005, Putnam et al. 2007, Technau et al. 2005). For example, *Nematostella* has more genes and a larger genome than insects and nematodes (Miller and Ball 2008). Furthermore, its genome encodes representatives of all bilaterian signalling pathways and of almost all transcription factor families (Larroux et al. 2008, Putnam et al. 2007, Ryan et al. 2006, Technau et al. 2005). In some cases (Wnt ligands and antagonists or BMP antagonists), *Nematostella* shares more orthologues with vertebrates than with insects or nematodes (Kusserow et al. 2005, Matus et al. 2006a, Rentzsch et al. 2006). On the other hand, several Fox genes appear secondarily lost in *Nematostella* (Larroux et al. 2008). Also, the cnidarian genome seems to be as complex as those of many bilaterians: in conserved genes, about 80% of human gene introns are conserved in *Nematostella* genes (considerably more than in *C. elegans* or *Drosophila*) (Putnam et al. 2007). The high

complexity of the *Nematostella* and bilaterian genomes in terms of gene content and architecture supports a morphologically complex cnidarian-bilaterian ancestor. However, considerable morphological differences of cnidarian and bilaterian body plans have made comparisons difficult so far. In the following sections, we highlight several examples where comparing conserved patterning systems shed light on the origin of putative bilaterian-specific features (e.g. the bilaterian body axes or the mesoderm) from a cnidarian-bilaterian ancestor.

5.5.2 Cnidarian BMP Patterning and the Evolution of the Bilaterian Dorso-Ventral Axis

The secreted bone morphogenetic proteins (BMPs) and their antagonists Chordin, Noggin, Gremlin and Follistatin have prominent and conserved roles in bilaterian dorso-ventral patterning (Grunz 2004). The neurogenic side of the embryo (ventral in protostomes and dorsal in vertebrates) expresses BMP antagonists, whereas the opposite side secretes BMP ligands. In contrast to bilaterians, a spatially antagonistic expression of anthozoan BMP agonists and antagonists along one axis is not discernable (Arendt and Nübler-Jung 1997, De Robertis and Sasai 1996). Whereas anthozoan BMP2/4 and BMP5/6/7/8 are expressed asymmetrically along both the directive axis (on one side of the blastopore and pharynx) and the oral-aboral axis (restricted to the oral side) in early stages, later expression is only restricted to one side of the directive axis (Hayward et al. 2002, Matus et al. 2006b, Rentzsch et al. 2006). Although both *Nematostella* Chordin and Gremlin proteins can antagonize *Nematostella* BMP2/4 in the heterologous zebrafish system, they are expressed inconsistently: while *gremlin* is the only described antagonist opposing *bmp2/4* expression along the directive axis, *chordin* expression, like that of *noggin*, largely overlaps *bmp2/4*. The radial expression of the *follistatin* antagonist expession further refutes a spatially clear agonist-antagonist situation (Matus et al. 2006a, b, Rentzsch et al. 2006). As most antagonists are also restricted to the oral side of the gastrula – just as BMPs are – a clear antagonism along the oral-aboral axis is also not observed. In general, the inconsistent expression with respect to any *Nematostella* axis suggests a spatially complex regulation of BMP signalling and questions the patterning of a single anthozoan axis by a clear, bilaterian-like BMP antagonism (Rentzsch et al. 2006). Rather, the recruitment of BMPs and their antagonists to pattern the dorso-ventral axis appears to have evolved after the bilaterian-cnidarian split. These results therefore imply no direct homology between the bilaterian dorso-ventral and any cnidarian body axis.

5.5.3 Cnidarian Hox Genes and the Evolution of the Antero-Posterior Axis

Besides the analysis of signalling pathways, the comparison of Hox gene expression may elucidate the relation of cnidarian and bilaterian axes. The Hox cluster is

a prominent example of a conserved antero-posterior patterning system in Bilateria, where the position of genes within the cluster reflects the relative expression site along the antero-posterior axis in the fly and vertebrate trunk (McGinnis and Krumlauf 1992). Bilaterian Hox genes are commonly grouped according to their relatedness into an "anterior" group (Hox1-2), a Hox3 group, a "central" group (Hox4-Hox8) and a "posterior" group (Hox9-13) (Garcia-Fernàndez 2005). Although cnidarians possess putative Hox gene homologs, their genomic structure and expression differ significantly from bilaterians:

- *Nematostella* possesses clear orthologs of "anterior" Hox genes, but no Hox3 ortholog. In earlier studies, some *Nematostella* Hox genes were assigned to "posterior" Hox genes (Hox9-13) (Finnerty and Martindale 1999, Finnerty et al. 2004), but more recent studies indicate that these genes have affinities to both "central" (Hox4-8) and "posterior" Hox genes without any clear orthology relationship (Chourrout et al. 2006, Ryan et al. 2006).
- The chromosomal linkage in *Nematostella* between a Hox gene and the non-Hox genes *Evx*, *Mnx* and *rough* is more or less conserved in bilaterians, but all other clustered Hox genes in the *Nematostella* genome appear to be the result of lineage-specific duplications (Chourrout et al. 2006, Kamm et al. 2006). Hence, a Hox cluster comparable to bilaterians is absent in *Nematostella*. Although there is local scrambling, the *Nematostella* genome shows a strong conservation of linkage groups with Bilateria (Putnam et al. 2007). It therefore seems unlikely that the absence of a Hox cluster is due to a secondary, lineage-specific disruption of an ancestral Hox cluster in Cnidaria.
- Most cnidarian Hox genes show endodermal expression (Finnerty et al. 2004, Ryan et al. 2007). In contrast, most bilaterian Hox genes are expressed in the ectoderm.
- Most *Nematostella* Hox genes show staggered expression – at most – along the directive axis, while only a few are differentially expressed along the oral-aboral axis (Finnerty et al. 2004, Ryan et al. 2007). Also, the expression of orthologous genes does not appear to be evolutionarily conserved in the planulae of the anthozoan *Nematostella* and the hydrozoans *Podocoryne* and *Eleutheria*; orthologues can be expressed at opposite poles of the oral-aboral axis (Finnerty et al. 2004, Kamm et al. 2006, Masuda-Nakagawa et al. 2000, Yanze et al. 2001). This refutes an evolutionarily conserved mode of Hox patterning along the cnidarian oral-aboral axis.

5.5.4 The Homology of Body Axes Between Cnidaria and Bilateria

These fundamental differences between bilaterian and cnidarian Hox genes do not rule out a general role for cnidarian Hox genes in axial patterning, but they do not support a direct homology of the bilaterian antero-posterior axis to the oral-aboral or the directive axis in cnidarians. Similarly, other bilaterian axial markers such as BMPs and their antagonists (see above), the dorsal-anterior marker *goosecoid*,

which is expressed in the *Nematostella* oral endoderm on both ends of the directive axis (Matus et al. 2006a), or orthologues of the anterior marker *otx,* which are expressed radially along the entire oral-aboral axis, allow for the same conclusion (de Jong et al. 2006, Mazza et al. 2007). The reason why no axis can be directly homologized between cnidarians and bilaterians might be the evolutionary innovation of the bilaterian trunk. In cnidarians, the initial axis running from the blastopore

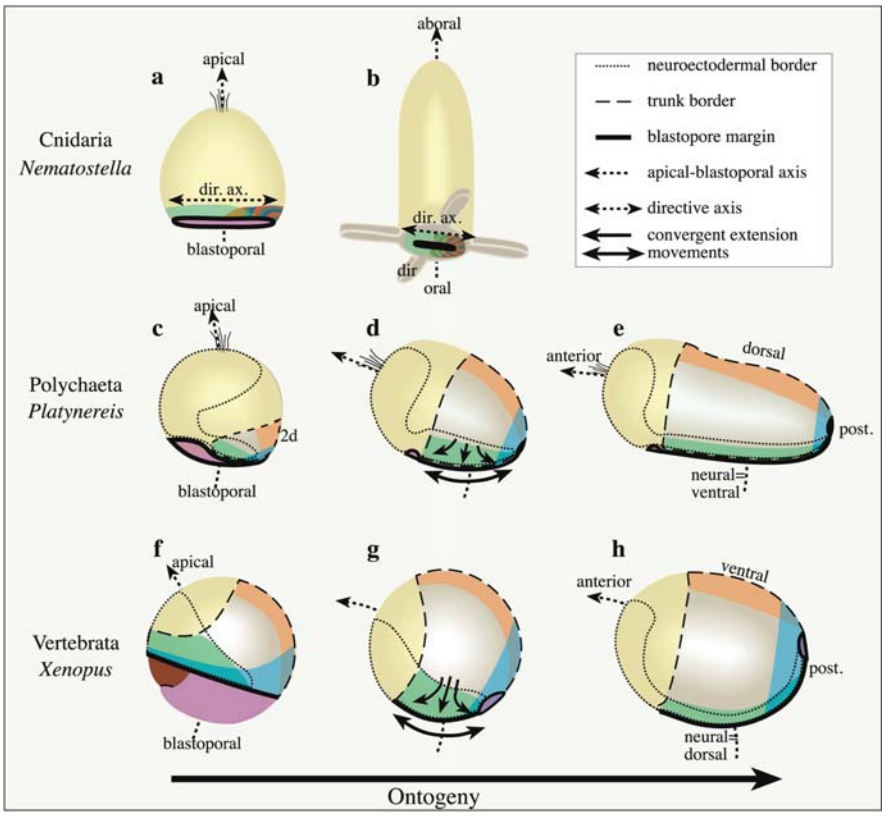

Fig. 5.4 Evolution of the bilaterian trunk and body axes. Ontogenetic comparison between *Nematostella* (**a**, **b**), *Platynereis* (**c**–**e**) and *Xenopus* (**f**–**h**) (Keller 1975) as representatives of Cnidaria, Protostomia and Deuterostomia to explain the evolution of the trunk and main body axes in Bilateria (based on Arendt 2004, Arendt and Nübler-Jung 1997, Denes et al. 2007, Shankland and Seaver 2000, Steinmetz et al. 2007). Selected species are considered relatively ancestral and prototypic but do not represent stem species of phyla. Note that in early bilaterian embryos (**c**, **f**), the blastopore margin (*thick black lines*) unifies ventral (*green*), posterior (*blue*) and dorsal (*orange*) fates that get separated by convergent extension movements (**d, e, g, h**). Also note the bending of the initial apical-blastoporal axis (*dotted arrow*) by the passive anterior tilting of the head (*yellow*) due to the proliferation and convergent extension of the 2d descendents that form and elongate the trunk. Orange and blue colouring of the blastopore rim in cnidarians represents the region on one end of the directive axis homologous to the trunk-forming region of Bilateria. *Purple*: endoderm and mesoderm. *Brown*: organizer region

to the opposite, apical pole, gives rise directly to the oral-aboral axis (Keller et al. 2000, Nielsen 2001). This is not the case in bilaterians, where the early apical-blastoporal (AB) axis is not directly comparable to the bilaterian antero-posterior (AP) or dorso-ventral (DV) axes. The single bilaterian AB axis often transforms into the AP and DV axis by gastrulation movements such as convergent extension (Fioroni 1992, Keller et al. 2000, Steinmetz et al. 2007). For example, the neuroectoderm adjacent to the early "organizer" regions at the blastopore margin of fishes or frogs gives rise to both anterior brain and posterior spinal chord (Hirose et al. 2004, Keller 1975, Woo and Fraser 1995). Also in protostome spiralians, the 2d blastomere is localized at one end of the early blastopore rim, and gives rise to the entire trunk ectoderm (Ackermann 2002, Nielsen 2004, Shankland and Seaver 2000). The fate of the blastopore, although considered to be conserved in bilaterians and cnidarians, differs and gives rise to the anterior mouth in deuterostomes, the posterior anus in protostomes, or both mouth and anus in "amphistome" animals, e.g. some annelids or nematodes (Arendt and Nübler-Jung 1997, Holland 2000, Nielsen 2001). Therefore, the apical-blastoporal axis of Bilateria is in fact as difficult to compare to the definite AP and DV axes as the apical-blastoporal (=aboral-oral) axis of Cnidaria (see Fig. 5.4).

In turn, this implies that it is the apical-blastoporal axes of cnidarians and bilaterians that might be truly homologous. Indeed, such a homology is supported by the conserved expression of *brachyury* (Technau 2001), *forkhead* (Fritzenwanker et al. 2004) and several Wnts (Kusserow et al. 2005) at the blastopore, the nuclear localization of β-catenin at the endoderm invagination site, and the capacity of the blastopore rim to induce a second apical-blastoporal axis (Kraus et al. 2007). The conflicting positions of the blastoporus at the cnidarian animal pole and the bilaterian vegetal pole is easily explained by the independent repositioning of the pronucleus, defining the animal-vegetal axis, in cnidarians or bilaterians (Lee et al. 2007, Martindale 2005). The hypothesis of axial evolution can further be tested by comparing the early cnidarian oral-aboral patterning with apical-blastoporal patterning in pre-gastrula stages of bilaterians.

5.5.5 Cnidarians and the Evolution of Mesoderm

Another example of how the ancestral components of newly acquired features can be detected by comparing cnidarian and bilaterian development is the origin of mesoderm, the third germ layer in bilaterians, from a bilayered cnidarian-bilaterian ancestor. Although a mesoderm proper is absent in cnidarians, they possess muscle cells, a main derivative of the mesoderm in Bilateria (Siewing 1985). Several general cnidarian muscle types can be discerned: the "myo-epithelial" type that is ubiquitous in cnidarians and combines contraction with additional sensory, secretory or digestive functions; and the more specialized muscle cell types in hydrozoan medusae and anthozoan polyps that often appear "striated" due to intracellular serial repetition of contractile units (Amerongen and Peteya 1980, Schuchert et al. 1993,

Siewing 1985). The comparison of cnidarian and bilaterian muscle structural genes will allow the debated evolutionary relationship between the bilaterian smooth and striated muscle types and the cnidarian muscle types to be clarified.

Cnidarians and bilaterians not only share a similar "mesodermal" derivative – muscle tissue – but also many transcription factors with specific roles during bilaterian mesodermal and muscle patterning. In the hydrozoan *Podocoryne carnaea*, the mesodermal specification genes *twist*, *mef2* (Spring et al. 2002) and *msx* (Galle et al. 2005) are also expressed in the "entocodon", an ectodermal cell mass proliferating between ecto- and endoderm during medusa formation in hydrozoan polyps. The entocodon develops into smooth and striated muscles and has been proposed to be a cnidarian homolog of the bilaterian mesoderm (Seipel and Schmid 2005). However, the entocodon is adult tissue that forms during non-sexual reproduction and therefore barely classifies as a third embryonic germ layer. Also, scyphozoan and cubozoan medusae develop muscles without entocodon (Burton 2007).

In *Nematostella*, several "mesodermal" transcription factors (*mef2*, *twist*, *gata, muscle-LIM*) are differentially expressed in the endoderm of the planula larva (Martindale et al. 2004). In anthozoans, most of the endoderm consists of myoepithelial cells, while more specialized muscles such as pharynx retractor muscles form together with germ cells in the "mesenteries", which are endodermal folds reaching into the body cavity (Siewing 1985). Functional analysis of conserved "mesodermal" transcription factors and the comparison with bilaterian orthologs is necessary to determine the extent of conservation between the gene regulatory networks governing bilaterian and cnidarian muscle and mesoderm development. As one hypothesis proposes that mesoderm evolved as a continuation of endodermal folding between ectoderm and endoderm, as can still be observed in bilaterians with enterocoelic mesoderm formation (e.g. *Amphioxus*, sea urchins, hemichordates, pogonophore annelids) (Arendt 2004, Remane 1950, Sedgwick 1884, Tautz 2004), it will be particularly interesting to compare the patterning and morphogenesis of the mesenteries between *Nematostella* and bilaterians. Alternatively, mesoderm might have evolved from endodermal cells that became mesenchymal as found in many extant spiralians (Technau and Scholz 2003).

5.5.6 "Cryptic" Complexity in Cnidarians?

In general, cnidarians appear morphologically simple at first sight. Therefore, the discovery that the *Nematostella* genome appears more complex than of insects or nematodes was surprising. However, a closer look reveals that the morphology of some cnidarians such as anthozoans might not be as simple after all: the presence of regulative development, a planula larva and two asymmetric body axes are all signs of complex development. Also, more cell types might remain to be described by molecular techniques that were so far not morphologically distinguishable. For example, the number of neuronal cell types might substantially increase as a result of investigations of the differential localisation of neurotransmitter receptor and

ligands. Also, cnidarians have independently expanded some gene families, such as light-perceiving opsins (Suga et al. 2008) or muscle contraction-regulating myosin light chains (unpublished) that might have led to an independent increase of cell types in some cnidarian lineages.

5.6 Ecdysozoans: Going Beyond the Established Systems

Ecdysozoans comprise a superphylum of molting animals that include, among others, nematodes, arthropods and priapulids (Aguinaldo et al. 1997) (also see Chapter 4). Two of the best-studied model systems of molecular biology – the nematode *Caenorhabditis elegans* and the fruitfly *Drosophila melanogaster* – belong to the ecdysozoans. Driven by the availability of powerful genetic tools, these two models have had an enormous impact on biological research, covering a broad range of topics that reaches from basic cellular principles to systems analyses. The significance of these studies for our understanding of biology remains undisputed. As outlined already above, however, more and more evidence challenges the unspoken assumptions that these findings are representative for the majority of invertebrates, or that they provide direct approximations of the processes prevailing in a simple animal predecessor. In contrast, these data suggest that some of the more simple features displayed by extant ecdysozoan models are the result of secondary simplification of ancestrally complex characters.

Notwithstanding the restricted retention of ancestral complexity, several ecdysozoan groups have evolved fascinatingly complex characters. For instance, the *Drosophila Down syndrome cell adhesion molecule (Dscam)* gene is an intron-rich gene that displays the highest number of splice variants identified in any species to date (Schmucker et al. 2000). In *Drosophila melanogaster*, Dscam isoforms appear to convey specific identity to migrating neurons, and mediate homophilic repulsions, an important prerequisite for the establishment of neural circuits. Dscam transcripts also display high diversity in related groups, including the crustacean *Daphnia*, suggesting that the cellular diversity mediated by *Dscam* alternative splicing could be a more basal arthropod feature (Brites et al. 2008). In contrast, vertebrate *Dscam* orthologs do not seem to undergo extensive alternative splicing, even though they also appear to be involved in homophilic interactions (reviewed in Hattori et al. 2008).

Examples like this illustrate that the evaluation of animal complexity, even when analysed on a quantitative molecular basis, needs to distinguish between the conservation of ancestral complexity and secondarily gained features of complexity. As it is still unclear how these two types of changes relate to one another, complex features of a given species or group will always have to be assessed for their evolutionary time of origin. It is also noteworthy in this context that both insects and nematodes are among the most diverse animal groups, with insects being the most species-rich group of all animals (Brusca and Brusca 2003). Despite several lost features of complexity, insects have thus evolved broad ranges of new morphological

complexity and variation. In the marine context, however, the diversity of insects is only moderate. Other ecdysozoan groups, including crustaceans and free-living nematodes, are more abundant in the sea, but suitable marine models and molecular data for these groups are still scarce.

One emerging marine model system among the arthropods is the amphipod *Parhyale hawaiiensis*, a developmental model species that is also subject to directed genome sequencing. In *Parhyale*, unlike the conventional arthropod model systems, early cleavages are total, unequal and invariant (Gerberding et al. 2002). Therefore, early patterning processes act in a fundamentally different context than in the insect model *Drosophila*, where early nuclear divisions generate a syncytium in which patterning molecules can diffuse. Future genetic analyses are thus necessary in order to reveal the extent to which the genetic machineries in arthropods are actually related. Such studies can rely on a growing repertoire of tools, including cell lineage tracing (Gerberding et al. 2002) and transgenesis (Pavlopoulos and Averof 2005). By allowing comparisons between distant arthropod taxa, these studies will also permit a clearer view of how complex the ancestral patterning machinery at the base of arthropods was. Likewise, studies in *Parhyale* reveal ancestral aspects of leg patterning in pancrustaceans, and thus provide data for the comparison – and ultimately the evolution – of the respective gene regulatory networks (Prpic and Telford 2008). Most of these studies can follow a candidate gene approach, yet a more unbiased view of the genetic constitution of amphipods will rely on genomic sequencing. Whereas the *Parhyale* genome is very large (3.6 Gbp), the smaller-sized genome (690 Mbp) of a related amphipod, *Jassa slatteryi*, has recently been proposed for genome sequencing and shall provide such unbiased insights into the amphipod genome.

5.7 Lophotrochozoans: An Evolutionary Branch Leading to New Perspectives

As outlined in Chapter 4, lophotrochozoans are an animal superphylum that combines, among others, molluscs and annelids (Halanych 2004, Halanych et al. 1995), and – along with the ecdysozoans – represents a large share of the protostomes. Although lophotrochozoans represent a major branch of eubilaterian evolution, they have for a long time remained relatively poorly covered by molecular descriptions, mostly owing to the success of ecdysozoan species (*Caenorhabditis elegans, Drosophila melanogaster*) as molecular genetic model systems. In recent years, expressed sequence tag- (EST), bacterial artificial chromosome- (BAC) and full genome-sequencing projects have started to close this gap by exploring the genomic repertoire of lophotrochozoans. Although none of the genome projects has so far been published, the emerging data already provide interesting facets of lophotrochozoan biology that are also of relevance for the understanding of animal complexity.

5.8 *Aplysia*: From Neural Circuits to Neurotranscriptomics

Molluscs comprise one of the largest phyla in the animal kingdom, second only to arthropods (Brusca and Brusca 2003). Many of the mollusc species are marine. Moreover, molluscs comprise the cephalopods, which are counted among the most highly evolved invertebrates. The sea hare, *Aplysia californica,* has become a classic model system for neurobiology, primarily due to the pioneering analyses on the setup and modulation of its simple neural circuitry, and the molecular mediators of learned behaviour (reviewed in Kandel 2001). The analysis of the *Aplysia* nervous system is greatly facilitated by the large and easy-to-recognize neurons of this species, some of which reach a diameter of 1 mm. A recent study made use of the size of these cells for RNA extractions to determine specific transcriptomes of the *Aplysia* nervous system. Specific libraries were generated from single types of neuron, for example the metacerebral cells (MCC) or even its neurites (Moroz et al. 2006). As a complementary approach to the ongoing *Aplysia* genome sequencing project, this EST sampling already revealed many interesting components. Indicative of the general mollusc gene repertoire, the search revealed additional genes missing from the representative ecdysozoan models, such as the zinc finger transcription factor *churchill* (also described in Kortschak et al. 2003), P2X receptor genes involved in pain sensing and forms of long-term synaptic plasticity, selenoprotein N homologs or major vault proteins with RNA binding capacity. Moreover, the sequences uncovered correlates of DNA methyl transferase 1, DNA methyl transferase associating protein, and the transcriptional repressor Methyl-CpG binding Domain Protein 2. The combination of these factors suggests that *Aplysia*, in contrast to *Caenorhabditis elegans* or insects, possesses a CpG methylation pathway, representing an independent route to gene regulatory complexity in this species.

This EST dataset can now be used to determine the cell-type-specific transcriptomes that will for instance allow the differences between two neuron types to be studied, or between neurons before and after stimulation. These data will not only be interesting for the particular neurobiological question, but will also have wider implications for the comparison of complexity: currently, it is unclear how easily new cell types can arise in evolution, and how many distinct molecular characteristics it takes to generate cell types of separate function. Moreover, a transcriptomic approach to define cell types could add a more quantitative and objective component to the analysis of cell type complexity, as this typically depends on more subjective measures. Empirical values for the molecular distinctions of individual cell types could in turn also provide new means to assess the evolution of cell types.

5.9 *Platynereis*: Ancestral Complexity of Cells and Genomic Features

Over the last decade, the marine annelid worm *Platynereis dumerilii* has begun to emerge as suitable model species for molecular evolutionary comparisons. Current molecular resources include high-quality sequences for more than 70,000 ESTs and

selected BAC sequences, with additional efforts directed towards the generation of a comprehensive EST dataset and a finished whole-genome sequence by the end of 2010. Moreover, *Platynereis* has a long history as an experimental model system, and gene delivery and gene interference tools are currently being established, adding to the available techniques for high-resolution gene expression studies (Jekely and Arendt 2007, Tessmar-Raible et al. 2005).

Based on the available EST and BAC sequences, a subset of the *Platynereis* transcriptome and genome has recently been used for a first systematic assessment of its gene structure and the evolution of its proteome. This study found a remarkable similarity between *Platynereis* and human genes, both with respect to their exon/intron organisation and concerning the speed with which the encoded proteins are evolving (Raible et al. 2005). These data provided for the first time evidence for the ancestrality of a large share of human introns, a notion that has since also been confirmed by the genomic analyses in the cnidarian *Nematostella* (Putnam et al. 2007). One illustrative example for this phenomenon and possible impacts on the evolution of regulatory complexity is provided by the analysis of the *pax6* gene. One of the key functions of the product of this gene lies in the specification of cells involved in the protostome and deuterostome photoreceptive systems (reviewed in Kozmik 2005).

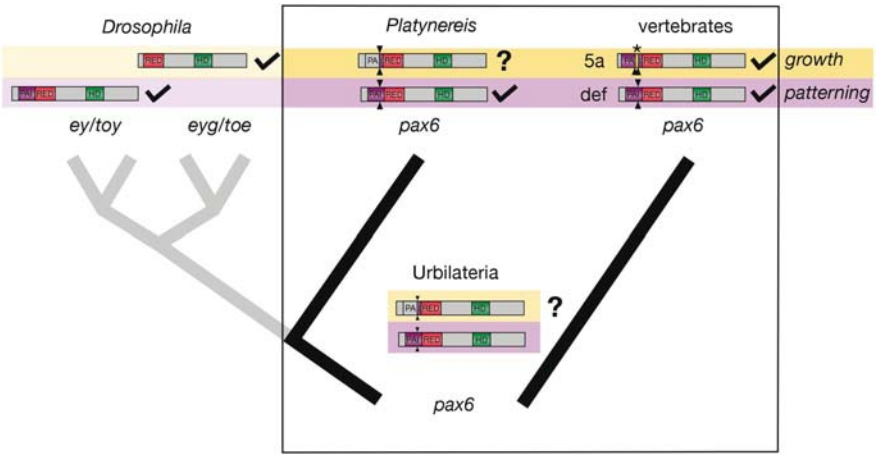

Fig. 5.5 Complex ancestral gene structures and the splicing potential of regulatory genes. Distinct *pax6* variants (*yellow* and *purple backround*, respectively) differ in the integrity of the N-terminal DNA binding domain and are preferentially associated with growth and patterning, respectively. Vertebrate variants are generated as splice isoforms, making use of the separation of the N-terminus into two exons, one of which encodes most of the N-terminal PAI domain that is structurally changed upon insertion of the additional 5a exon (*asterisk*). In contrast, *Drosophila* possesses distinct sets of orthologues lacking the respective intron site. Comparative analysis of the *Platynereis pax6* locus (*middle*) indicates that the intron (*black arrowheads*) within the N-terminal region (PAI) of the PAIRED domain is ancestral for Bilateria and was secondarily lost in the four *Drosophila pax6* orthologs (*left*). Therefore, one alternative scenario is that Urbilateria already generated a functional equivalent of Pax6(5a) (marked by "?") that became fixed as independent variants in the fly genome, but is subject to alternative splicing in both annelids and vertebrates

Moreover, mice deficient for this gene have demonstrated an essential role for *pax6* in the differentiation of glucagon-secreting alpha cells of the pancreas (St-Onge et al. 1997). As an example of the remarkable conservation of gene structures in Bilateria, the *Platynereis pax6* gene contains an intron in the PAIRED domain at the precise site where it is found in vertebrates (see Raible et al. 2005 and Fig. 5.5). This intron is instrumental in the generation of two functionally divergent splice variants in vertebrates, but is absent from the genomes of the ascidians *Ciona* and *Phallusia*, and has therefore traditionally been assumed to be a vertebrate innovation (Callaerts et al. 1997). The data from *Platynereis* suggest that the ancestral eubilaterian *pax6* locus already possessed the necessary introns for generating alternative splice variants at this site. There is not yet any experimental evidence for the existence of a specific splice variant in *Platynereis*. In keeping with this possibility, however, the two sets of orthologues observed in *Drosophila* are mainly distinct in their N-terminal portion, consistent with the fixation of ancestral splice variants in the fly genome (Fig. 5.5). Moreover, vertebrate Pax6(5a) can substitute for the Drosophila *eyg* gene (Dominguez et al. 2004), indicating a functional equivalence that has so far been regarded as the product of parallel evolution. Notably, a recent report describes the isolation of alternative splice variants in *Amphioxus* that are compatible with the notion of an ancestral splice event in the PAIRED domain predating the evolution of vertebrates (Short and Holland 2008).

5.10 Alternative Splicing: Modulating the Basic Layers of Genomic Complexity?

Besides the specific example outlined above, the repeated finding of complex gene structures at the base of animal evolution raises a more general question concerning the evolution of transcriptome complexity and its relationship with animal complexity. From early genome measurements, as well as from the first whole-genome projects, it became clear that neither genome size nor the number of protein-coding genes correlate well with the assumed morphological complexity of different animals: complex metazoans can have smaller genomes than protozoans (c-value paradox) and the total gene number in complex vertebrates is not fundamentally higher than in invertebrates (the n- or g-value paradox) (Claverie 2001, Hahn and Wray 2002). Alternative splicing provides a direct way to modulate the complexity of a proteome irrespective of changes in gene number (reviewed in Maniatis and Tasic 2002), and therefore is an attractive candidate for mediating cell type complexity. Whereas it is difficult to capture the full extent of alternative transcripts in any organism, first global analyses at least indicate that even between closely related species such as humans and chimpanzees, there are significant changes in the splicing of 6–8% of orthologous exons (Calarco et al. 2007). As this is a significant figure compared to the few documented cases of gene gain/loss between the two species (Chimpanzee Sequencing and Analysis 2005), this suggests that changes in alternative splicing could well contribute to changes in complexity between species, or

even contribute to speciation itself. Comparison of alternative splice variants over larger evolutionary time scales has so far been limited by the restricted coverage of splice variants in given model systems (Boue et al. 2003, Brett et al. 2002), and cases such as the aforementioned isoform diversity of Dscam in pancrustaceans rather seem to point towards mainly independent expansion of splice variants in different lineages. As the issue of coverage might soon be less relevant due to the technological progress in high-throughput sequencing, it will be interesting to see if there are systematic differences in the extent of alternative splicing in species of different organizational level.

5.11 Sea Urchins: Unexpected Functional Repertoires at the Base of Deuterostomes

Sea urchins are among the oldest experimental model systems used by marine biologists. In the second half of the twentieth century, experimental approaches have been taken to the biochemical and molecular levels, with seminal work focusing on the elucidation of transcriptional regulatory networks. The purple sea urchin, *Strongylocentrotus purpuratus*, has thereby become a major model system for the analysis of developmental gene-regulatory networks. The recent sequencing of the *Strongylocentrotus* genome (Sea Urchin Genome Sequencing et al. 2006) has boosted experimental research in the urchin, and also provided first insights into an echinoderm genome. Three aspects are of particular relevance in the context of this chapter: first, the sea urchin genome draft displayed an unprecedented number of genes that – judged by their domain composition – are likely to be associated with innate immunity, including 222 Toll-like receptors, more than 20 times the number found in the human genome (Hibino et al. 2006, Rast et al. 2006). Although direct functional assays have not yet been performed on these receptors, the numbers suggest that the family has undergone dramatic secondary expansion in the evolutionary lineage leading to sea urchins. Similarly, the sea urchin genome contains hundreds of fast-evolving G-protein coupled receptors (GPCRs) that, based on their mode of organisation and their expression patterns, are likely to be chemosensory receptors (Raible et al. 2006). By analogy to the immune-related genes, the dramatic expansion of the sensory GPCR family contradicts the old notion that the sea urchin would only possess a rudimentary sensory repertoire.

A third, fundamental aspect about the sea urchin genome is that it encodes both a bilaterally symmetric larval form and a fundamentally different structure, the radially symmetric postembryonic body whose cells replace most of the larval structures. Hence, the programs of two different body plans are encoded in the same genome. How this is achieved, is still not properly understood, as few studies have attempted to analyse the activity of genes in postembryonic development. What has been revealed by transcriptome profiling studies, however, is that nearly 80% of the transcription factors encoded in the sea urchin genome are already active during embryogenesis (Howard-Ashby et al. 2006), as are a similar proportion of genes

associated with signalling processes (Samanta et al. 2006). These data thus support the idea that dramatically different body plans can be generated by the differential use of the same components, and caution against a simple link between gene content and complexity of genomic regulatory programs.

5.12 Lancelets and the Chordate Prototype

The lancelets, also known as *Amphioxus* or *Branchiostoma*, are a small group of cephalochordates that have a long tradition as model systems. Due to their intermediate phylogenetic position between the non-chordate deuterostomes and the vertebrates, cephalochordates form a very informative group for evolutionary comparisons, and have traditionally served to trace back the origin of many vertebrate features (reviewed in Garcia-Fernàndez and Bentio-Gutiérrez 2009). Lancelets share important features with vertebrates, such as the dorsal nerve cord, as well as a notochord that lends stability to the swimming larvae, but they lack the vertebral column that protects the dorsal nerve cord in vertebrates. Moreover, lancelets possess a perforated pharynx (pharyngeal slits) and bilateral blocks of segmented muscle called myomeres that can be well compared with the somitic myotomes in vertebrates. In turn, limbs, neural crest cells, paired sensory organs and an elaborate anterior brain are missing in amphioxus, consistent with a later evolutionary emergence of these features.

On the molecular level, the recently published genome of *Branchiostoma floridae* (Holland et al. 2008, Putnam et al. 2008) continues a very interesting series of individual studies that shed light on molecular evolution along the deuterostome – chordate – vertebrate lineage: Findings like the discovery of a single Hox cluster in *Branchiostoma* (Garcia-Fernàndez and Holland 1994), in comparison to the four Hox clusters typical for vertebrates, already suggested that cephalochordates still reflect a "primitive" genomic condition before the occurrence of two rounds of whole-genome duplications ("2R hypothesis"). Likewise, the discovery of an intact Para-Hox cluster in *Branchiostoma* (Brooke et al. 1998), in contrast to degenerated Para-Hox clusters in other systems, provided evidence that the lancelet genome likely had preserved ancestral characteristics that secondarily changed in other lineages. This trend has recently been confirmed by the whole genome analysis of *Branchiostoma*. This analysis uncovered a remarkable degree of syntenic relationships between lancelet gene loci and their counterparts in the vertebrates. Among others, the comparison between the lancelet and human genome allows for the detection of macro-synteny between their chromosomes, suggesting a set of at least 17 linkage groups in the chordate ancestor (Putnam et al. 2008). The presence of four vertebrate relatives for each of these regions is additional support for the "2R hypothesis", and also emphasizes the status of the *Branchiostoma* as a good representative of the more primitive chordate condition.

Another remarkable finding in the *Branchiostoma* genome was the existence of more than 50 non-coding elements that were highly conserved with vertebrate genomes (Putnam et al. 2008). This number already excludes conserved features in

the UTRs of genes, and is speculated to comprise ultraconserved enhancer elements. More than 3,000 of such elements have already been identified by genome comparisons among vertebrate species and seem to primarily act during development (Pennacchio et al. 2006, Woolfe et al. 2005). The ability to assess the regulatory potential of these elements in heterologous systems such as the mouse (Holland et al. 2008) now opens exciting possibilities to test if they might represent key regulatory connections of pan-chordate relevance.

5.13 Ascidians: Changes and Constants in Developmental Programmes

The ascidian *Ciona intestinalis* is a very interesting species for the comparison of animal complexity. Phylogenetic analyses demonstrate that tunicates are the closest relatives of vertebrates (Delsuc et al. 2006, 2008), implying that they are suitable models for testing the evolution of vertebrate characters. Notably, whereas *Ciona* possesses a tadpole-type larva, similar to those of chordate vertebrates like the frog, the number of cells of this larva is only around 2,600, at least an order of magnitude less than what is found in vertebrates. In addition, the *Ciona* genome is only 5% of the size of the mouse genome. So, in comparison with vertebrate systems, *Ciona* seems to generate a similar level of larval organizational complexity using less non-coding DNA and cell types. Comparison of regulatory processes in *Ciona* and vertebrate model systems therefore holds a lot of potential for the understanding of regulatory complexity and developmental plasticity.

As early gene expression patterns in *Ciona* can be routinely mapped with cellular resolution (Tassy et al. 2006), and functional tools also exist for *Ciona*, it is not only possible to analyse and compare the expression of developmental marker genes in *Ciona*, but also to systematically assess the function of *Ciona* genes. This combination of high-resolution gene expression data and functional techniques makes *Ciona* a powerful tool for the analysis of developmental gene-regulatory networks in a simple marine chordate, information which can then be compared with other clades (Imai et al. 2006, Satou and Satoh 2006, Shoguchi et al. 2008). Importantly, such comparisons have revealed different extents of conservation among chordate gene-regulatory networks. Some components seem to act in similar ways in both ascidian and vertebrate embryogenesis. For instance, a fibroblast growth factor signal triggers the expression of the T-box transcription factor Brachyury and the formation of the notochord in both systems, hinting at a conserved regulatory system orchestrating the development of this chordate character (Imai et al. 2002). Likewise, effector genes encoding structural notochord components like type II collagen or proteoglycans are conserved between *Ciona* and vertebrates (Hotta et al. 2008). In contrast, upstream regulators of this process and their interconnections seem to differ more dramatically between the systems (Imai et al. 2006, Lemaire 2006). The apparent contrast between conserved and divergent aspects of gene regulatory networks raises the question how “hard-wired” early developmental programmes are,

and why certain levels of developmental programmes seem to tolerate more changes than others. As the ultimate functional "output" of these programmes, the swimming tadpole, appears to be remarkably stable, the answers to these questions probably also provide insights into the molecular correlates of morphological and cellular complexity.

5.14 Perspectives

As the examples covered in this chapter show, the study of new marine model systems adds new and interesting perspectives to our understanding of animal complexity. As a complement to traditional characteristics of complexity, such as the number of differentiated cell types, genomic approaches offer different measures of complexity. On a basic level, these are the number of protein domains, genes, introns and transcripts in a given organism that underlie its genetic networks. Such features are straightforward to compare, and have helped to date the existence of many domains, genes and introns back to more basal positions in diverse animal groups. These, as well as additional findings, have several conceptual implications.

Loss of Ancestral Complexity as an Evolutionary Principle: The study of new model systems recurrently finds evidence for features that were present in ancestral genomes, yet were secondarily lost along certain evolutionary lineages. This helps to readjust our view on animal evolution: apparently, loss of ancestral features is part of the normal evolutionary process, just as gene duplication and modification are plausible sources of new genomic complexity. The genes in question include regulatory genes such as transcription factors, but also extracellular signals, factors that – from functional analyses – are known to have significant impact on animal development. It is therefore an intriguing question as to the role the apparent loss of ancestral complexity has played in shaping the evolutionary process.

Differences in evolutionary speed: In addition to the existence of loss as a fundamental principle of evolution, we note that there are pronounced differences between animal groups with respect to the extent of losses they display. Examples of this are the slow evolutionary pace of annelids or cephalochordates, as opposed to the massive acceleration in ascidians. As these marine examples illustrate, the question of fast vs. slow evolution is not coupled with whether or not a group occurs in the sea. The new focus on marine species, however, helps to systematically fill important gaps in the phylogeny of animals that contributes to our understanding of what are ancestral, and what are secondarily acquired features.

Regulatory Networks: Much of the focus of modern genomic approaches has been on the comparison of simple features, such as the presence or absence of genes and introns. As discussed in the introduction, however, this level is only a first measure of complexity. How do the single entities work together to form regulatory networks, and how does complexity on the level of network components impact on the complexity of the resulting networks themselves? Questions of this type require a combination of functional tools like gene interference coupled with expression

analyses, e.g. by microarrays or high-throughput sequencing. As outlined above, such experiments now become feasible in a limited set of marine model species (like *Ciona, Nematostella*, sea urchin, *Amphioxus*, or *Platynereis*) and these experimental approaches should cast some light on the basic wiring, and thus complexity, of regulatory networks in diverse phyla.

Molecular correlates of classical features of complexity: Another fundamental question that still awaits future experimentation concerns the connection between the new, molecular measures of complexity and the traditional measures of morphological complexity and cell type diversity. For instance, what are the molecular correlates that enabled the evolution of multicellularity? Similarly, is the number of differentiated cell types correlated with the complexity of developmental networks, or is there an independent degree of "regulatory capacity" of an animal (e.g. depending on the individual mode of development). To date, our understanding of cell type differentiation is limited to a few cases such as the lymphoid lineage, or pancreatic beta cells. There is still much groundwork to do to identify the correlates of distinct cell types on the molecular level. In turn, such analyses could also determine whether apparently identical cells in fact represent molecularly distinct cell types.

Evolution acting on different regulatory layers: More and more evidence indicates that non-coding DNA (previously referred to as "junk-DNA") carries out important regulatory functions: such as providing binding sites for transcriptional regulators, being a substrate for epigenetic modifications, or producing non-coding transcripts with regulatory function. Each of these aspects is of relevance for the proper understanding of genomic and morphological complexity (Hahn and Wray 2002, Levine and Tjian 2003, Mattick 2007). For instance, modification of cis-regulatory elements is associated with changes in complex pigment patterns in the dipteran wing (reviewed in Prud'homme et al. 2007). Moreover, non-coding microRNAs have emerged as a previously unknown layer of post-transcriptional regulation. Given that many microRNAs are expressed in a tissue- or cell type-specific manner, they might be good indicators – or even determinants – of cellular complexity. Several studies have tried to correlate the absence or presence of microRNAs with the different extents of complexity in animals, with a trend towards reproducing the assumed patterns of complex vertebrate microRNA repertoires versus poorer repertoires of more basally branching animals (Grimson et al. 2008, Heimberg et al. 2008). Due the absence of comparable microRNA datasets from the different groups, it is, however, difficult to assess to what degree experimental coverage skews the outcome of such analyses.

New approaches to assess homology and morphological evolution: The revolution in sequencing technology, combined with the progress in the establishment of new molecular model systems that can make use of the new sequence data, also pave the way to re-approach many of the questions that were dealt with in pre-molecular times. One fundamental impact concerns the more precise description of homology: As outlined above, the availability of molecular markers helps to distinguish and compare tissues and cells on the molecular level, for instance to trace back mesodermal features in cnidarians, and to discriminate – with fresh data – between different scenarios concerning their relationship to tissues in other taxa. Moreover,

on the basis of such comparisons, and with the help of the molecular toolboxes used to dissect the developmental regulatory networks acting in these tissues, it also starts to become possible to identify the molecular correlates involved in the emergence and modification of body plans, and to study the molecular correlates of visible morphological features. Ultimately, these lines of research therefore aim to provide a link between the genome and the features it encodes and that traditionally have been used to discriminate between species.

Understanding animal diversity: The new marine model systems outlined in this chapter already significantly broaden our view of evolutionary principles, but still, they represent only a very small subset of the wealth of species living in the ocean. There is great potential for the use of the new genomic approaches to address questions related to the diversity of animal groups. On the one hand, sequence data available for reference models allows us to address how phenotypic differences (e.g. in form or colour) are encoded on the molecular level. On the other hand, there also appear to exist intriguing differences in the genetic networks even between closely related, morphologically similar species. In both fields, genomic techniques already provide a very useful basis for molecular analysis, and also hold a lot of potential for future studies. Such studies will hopefully help us to re-draw, at some point, a more accurate version of Haeckel's tree of life that reflects better the true origins of the different animals, along with the evolutionary paths that brought them into existence.

Acknowledgements The authors wish to thank Ferdinand Marlétaz and Benjamin Backfisch for critical reading of the manuscript, and Hanno Sandvik for help in locating the original lithograph reproduced in Fig. 5.1. Research in F.R.'s laboratory is supported by a start-up fund of the Max F. Perutz Laboratories.

References

Abedin M, King N (2008) The premetazoan ancestry of cadherins. Science 319: 946–948

Aburomia R et al (2003) Functional evolution in the ancestral lineage of vertebrates or when genomic complexity was wagging its morphological tail. J Struct Funct Genomics 3: 45–52

Ackermann C (2002) Markierung der Zellinien im Embryo von *Platynereis*. In: Fachbereich biologie, ed. Mainz: Johannes Gutenberg-Universität

Adamska M et al (2007a) Wnt and TGF-β expression in the sponge *Amphimedon queenslandica* and the origin of metazoan embryonic patterning. PLOS One 2: e1031

Adamska M et al (2007b) The evolutionary origin of hedgehog proteins. Curr Biol 17: R836–R837

Adell T et al (2003) Isolation and characterization of two T-box genes from sponges, the phylogenetically oldest metazoan taxon. Dev Genes Evol 213: 421–434

Adell T, Müller WEG. (2004) Isolation and characterization of five Fox (Forkhead) genes from the sponge *Suberites domuncula*. Gene 334: 35–46

Adell T, Müller WEG. (2005) Expression pattern of the Brachyury and Tbx2 homologues from the sponge *Suberites domuncula*. Biol Cell 97: 641–650

Aguinaldo AM et al (1997) Evidence for a clade of nematodes, arthropods and other moulting animals. Nature 387: 489–493

Amerongen HM, Peteya DJ (1980) Ultrastructural study of two kinds of muscle in sea anemones: the existence of fast and slow muscles. J Morphol 166: 145–154

Arendt D (2004) Comparative aspects of gastrulation. In: Stern C (ed) Gastrulation, edn. Cold Spring Harbor Laboratory Press, Cold SPring Harbor, New York
Arendt D (2005) Genes and homology in nervous system evolution: comparing gene functions, expression patterns, and cell type molecular fingerprints. Theory Biosci 124: 185–197
Arendt D (2008) The evolution of cell types in animals: emerging principles from molecular studies. Nat Rev Genet 9: 868–882
Arendt D, Nübler-Jung K (1997) Dorsal or ventral: similarities in fate maps and gastrulation patterns in annelids, arthropods and chordates. Mech Dev 61: 7–21
Bell G (1997) Size and complexity among multicellular organisms. Biol J Linnean Soc 60: 345–363
Bijlsma MF et al (2004) Hedgehog: an unusual signal transducer. Bioessays 26: 387–394
Bonner JT. (1988) The evolution of complexity. Princeton University Press, Princeton, NJ
Borchiellini C et al (2004) Molecular phylogeny of demospongiae: implications for classification and scenarios of character evolution. Mol Phylogenet Evol 32: 823–837
Borchiellini C et al (2001) Sponge paraphyly and the origin of Metazoa. J Evol Biol 14: 171–179
Boue S et al (2003) Alternative splicing and evolution. Bioessays 25: 1031–1034
Brett D et al (2002) Alternative splicing and genome complexity. Nat Genet 30: 29–30
Brites D et al (2008) The Dscam homologue of the crustacean *Daphnia* is diversified by alternative splicing like in insects. Mol Biol Evol 25: 1429–1439
Brooke NM et al (1998) The ParaHox gene cluster is an evolutionary sister of the Hox gene cluster. Nature 392: 920–922
Brusca RC, Brusca GJ. (2003) Invertebrates. Sinauer Associates. Sunderland, Massachusetts. http://www.sinauer.com/detail.php?id=0973
Bullock TH, Horridge GA (1965) Structure and function in the nervous system of invertebrates. San Francisco: Freeman
Burton PM (2007) Inisghts from diploblasts; the evolution of mesoderm and muscle. J Exp Zool (Mol Dev Evol) 308B: 1–10
Bütschli O. (1883–1887) Klassen und Ordnungen des Thier-Reichs. Winter, C. F., Leipzig
Calarco JA et al (2007) Global analysis of alternative splicing differences between humans and chimpanzees. Genes Dev 21: 2963–2975
Callaerts P et al (1997) PAX-6 in development and evolution. Ann Rev Neurosci 20: 483–532
Cañestro C et al (2007) Evolutionary developmental biology and genomics. Nat Rev Genet 8: 932–942
Carr M et al (2008) Molecular phylogeny of choanoflagellates, the sister group to Metazoa. PNAS 105: 16641–16646
Chimpanzee Sequencing and Analysis C (2005) Initial sequence of the chimpanzee genome and comparison with the human genome. Nature 437: 69–87
Chourrout D et al (2006) Minimal ProtoHox cluster inferred from bilaterian and cnidarian Hox complements. Nature 442: 684–687
Clark H (1866) Note on the infusoria flagellate and the spongiae ciliatae. Am J Sci 1: 113–114
Clark H (1868) On the Spongiae ciliatae as *Infusoria flagellata*, or observations on the structure, animality and relationship of *Leucosolenia botryoides* Bowerbank. Ann Mag Nat Hist 4: 133–142, 188–215, 250–264
Claverie JM (2001) Gene number. What if there are only 30,000 human genes? Science 291: 1255–1257
Collins AG (1998) Evaluating multiple alternative hypotheses for the origin of Bilateria: an analysis of 18S rRNA molecular evidence. Proc Natl Acad Sci USA 95: 15458–15463
Collins AG (2002) Phylogeny of Medusozoa and the evolution of cnidarian life cycles. J Evol Biol 15: 418–432
da Silva FB et al (2007) Phylogenetic position of Placozoa based on large subunit (LSU) and small subunit (SSU) rRNA genes. Genet Mol Biol 30: 127–132
de Jong DM et al (2006) Components of both major axial patterning systems of the Bilateria are differentially expressed along the primary axis of a 'radiate' animal, the anthozoan cnidarian *Acropora millepora*. Dev Biol 298: 632–643

De Robertis EM, Sasai Y (1996) A common groundplan for dorsoventral patterning in Bilateria. Nature 380: 37–40
Dellaporta SL et al (2006) Mitochondrial genome of *Trichoplax adhaerens* supports placozoa as the basal lower metazoan phylum. Proc Natl Acad Sci USA 103: 8751–8756
Delsuc F et al (2006) Tunicates and not cephalochordates are the closest living relatives of vertebrates. Nature 439: 965–968
Delsuc F et al (2008) Additional molecular support for the new chordate phylogeny. Genesis 46: 592–604
Denes AS et al (2007) Molecular architecture of annelid nerve cord supports common origin of nervous system centralization in Bilateria. Cell 129: 277–288
Derelle R et al (2007) Homeodomain proteins belong to the ancestral molecular toolkit of eukaryotes. Evol Dev 9: 212–219
Dominguez M et al (2004) Growth and specification of the eye are controlled independently by Eyegone and Eyeless in *Drosophila melanogaster*. Nat Genet 36: 31–39
Ereskovsky AV, Dondua AK (2006) The problem of germ layers in sponges (Porifera) and some issues concerning early metazoan evolution. Zoologischer Anzeiger 245: 65–76
Finnerty JR, Martindale MQ (1999) Ancient origins of axial patterning genes: Hox genes and ParaHox genes in the Cnidaria. Evol Dev 1: 16–23
Finnerty JR et al (2004) Origins of bilateral symmetry: Hox and dpp expression in a sea anemone. Science 304: 1335–1337
Fioroni P (1992) Allgemeine und vergleichende Embryologie. Springer, Berlin, Heidelberg, New York
Fritzenwanker JH et al (2004) Analysis of *forkhead* and *snail* expression reveals epithelial-mesenchymal transitions during embryonic and larval development of *Nematostella vectensis*. Dev Biol 275: 389–402
Galle S et al (2005) The homeobox gene Msx in development and transdifferentiation of jellyfish striated muscle. Int J Dev Biol 49: 961–967
Garcia-Fernàndez J (2005) The genesis and evolution of homeobox gene clusters. Nat Rev Genet 6: 881–892
Garcia-Fernàndez J, Bentio-Gutiérrez E (2009) It's a long way from amphioxus: descendants of the earliest chordate. Bioessays 31: 665–675
Garcia-Fernàndez J, Holland PWH (1994) Archetypal organization of the amphioxus hox gene-cluster. Nature 370: 563–566
Gerberding M et al (2002) Cell lineage analysis of the amphipod crustacean *Parhyale hawaiensis* reveals an early restriction of cell fates. Development 129: 5789–5801
Gregory TR (2005) Genome size evolution in animals. In: Gregory TR (ed) The evolution of the genome, 1st edn. Elsevier, San Diego
Grell KG (1971a) Embryonalentwicklung bei *Trichoplax adherens* F.E. Schulze. Naturwiss 58: 507
Grell KG (1971b) *Trichoplax adherens*: F.E. Schulze und die Entstehung der Metazoen. Naturwiss Rundschau 24: 160–161
Grell KG (1972) Eibildung und Furchung von *Trichoplax adherens* F.E. Schulze (Placozoa). Z Morph Tiere 73: 297–314
Grell KG, Ruthman A (1991) Placozoa, Porifera, Cnidaria and Ctenophora. In Harrisson FW, Westfall JA (eds) Microscopic anatomy of invertebrates. Wiley-Liss, New York
Grimson A et al (2008) Early origins and evolution of microRNAs and Piwi-interacting RNAs in animals. Nature 455: 1193–1197
Grunz H (2004) The vertebrate organizer. Springer, Berlin Heidelberg
Haeckel E (1874) Die Gastraea-Theorie, die phylogenetische Classification des Thierreiches und die Homologie der Keimblätter. Jena Z. Naturwiss 8: 1–55
Haeckel E (1903) Anthropogenie oder Entwickelungsgeschichte des Menschen. Keimes- und Stammes-Geschichte. Wilhelm Engelmann, Leipzig

Haen KM et al (2007) Glass sponges and bilaterian animals share derived mitochondrial genomic features: a common ancestry or parallel evolution? Mol Biol Evol 24: 1518–1527
Hahn MW, Wray GA (2002) The g-value paradox. Evol Dev 4: 73–75
Halanych KM (2004) The new view of animal phylogeny. Ann Rev Ecol Evol Sys 35: 229–256
Halanych KM et al (1995) Evidence from 18S ribosomal DNA that the lophophorates are protostome animals. Science 267: 1641–1643
Halder G et al (1995) Induction of ectopic eyes by targeted expression of the *eyeless* gene in *Drosophila*. Science 267: 1788–1792
Hattori D et al (2008) Dscam-mediated cell recognition regulates neural circuit formation. Annu Rev Cell Dev Biol 24: 597–620
Hayward DC et al (2002) Localized expression of a dpp/BMP2/4 ortholog in a coral embryo. Proc Natl Acad Sci USA 99: 8106–8111
Heimberg AM et al (2008) MicroRNAs and the advent of vertebrate morphological complexity. Proc Natl Acad Sci USA 105: 2946–2950
Hibino T et al (2006) The immune gene repertoire encoded in the purple sea urchin genome. Dev Biol 300: 349–365
Hirose Y et al (2004) Single cell lineage and regionalization of cell populations during Medaka neurulation. Development 131: 2553–2563
Holland LZ (2000) Body-plan evolution in the Bilateria: early antero-posterior patterning and the deuterostome-protostome dichotomy. Curr Opin Genet Dev 10: 434–442
Holland LZ et al (2008) The amphioxus genome illuminates vertebrate origins and cephalochordate biology. Genome Res 18: 1100–1111
Hotta K et al (2008) Brachyury-downstream gene sets in a chordate, *Ciona intestinalis*: integrating notochord specification, morphogenesis and chordate evolution. Evol Dev 10: 37–51
Howard-Ashby M et al (2006) High regulatory gene use in sea urchin embryogenesis: Implications for bilaterian development and evolution. Dev Biol 300: 27–34
Imai KS et al (2006) Regulatory blueprint for a chordate embryo. Science 312: 1183–1187
Imai KS et al (2002) Early embryonic expression of FGF4/6/9 gene and its role in the induction of mesenchyme and notochord in *Ciona savignyi* embryos. Development 129: 1729–1738
Jager M et al (2005) Expansion of the SOX gene family predated the emergence of the Bilateria. Mol Phylogenet Evol 39: 468–477
Jaillon O et al (2004) Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-karyotype. Nature 431: 946–957
Jakob W et al (2004) The Trox-2 Hox/ParaHox gene of *Trichoplax* (Placozoa) marks an epithelial boundary. Dev Genes Evol 214: 170–175
Jekely G, Arendt D (2007) Cellular resolution expression profiling using confocal detection of NBT/BCIP precipitate by reflection microscopy. Biotechniques 42: 751–755
Kamm K et al (2006) Axial patterning and diversification in the cnidaria predate the Hox system. Curr Biol 16: 920–926
Kandel ER (2001) The molecular biology of memory storage: a dialogue between genes and synapses. Science 294: 1030–1038
Keller R et al (2000) Mechanisms of convergence and extension by cell intercalation. Philos Trans R Soc Lond B Biol Sci 355: 897–922
Keller RE (1975) Vital dye mapping of the gastrula and neurula of *Xenopus laevis*. I. Prospective areas and morphogenetic movements of the superficial layer. Dev Biol 42: 222–241
King N et al (2003) Evolution of key cell signaling and adhesion protein families predates animal origins. Science 301: 361–363
King N et al (2008) The genome of the choanoflagellate *Monosiga brevicollis* and the origin of metazoans. Nature 451: 783–788
Kortschak RD et al (2003) EST analysis of the cnidarian *Acropora millepora* reveals extensive gene loss and rapid sequence divergence in the model invertebrates. Curr Biol 13: 2190–2195
Kozmik Z (2005) Pax genes in eye development and evolution. Curr Opin Genet Dev 15: 430–438

Kozmik Z et al (2003) Role of Pax genes in eye evolution: a cnidarian *PaxB* gene uniting Pax2 and Pax6 functions. Dev Cell 5: 773–785
Kraus Y et al (2007) The blastoporal organiser of a sea anemone. Curr Biol 17: R874–R876
Kusserow A et al (2005) Unexpected complexity of the Wnt gene family in a sea anemone. Nature 433: 156–160
Lang BF et al (2002) The closest unicellular relatives of animals. Curr Biol 12: 1773–1778
Larroux C et al (2007) The NK homeobox gene cluster predates the origin of Hox genes. Curr Biol 17: 706–710
Larroux C et al (2006) Developmental expression of transcription factor genes in a demosponge: insights into the origin of metazoan multicellularity. Evol Dev 8: 150–173
Larroux C et al (2008) Genesis and expansions of metazoan transcription factor classes. Mol Biol Evol 25: 980–996
Leadbeater BSC (1983) Life-history and ultrastructure of a new marine species of *Proterospongia* (Choanoflagellida). J Mar Biol Assoc UK 63: 135–160
Lee PN et al (2007) Asymmetric developmental potential along the animal-vegetal axis in the anthozoan cnidarian, *Nematostella vectensis*, is mediated by Dishevelled. Dev Biol 310: 169–186
Lemaire P (2006) Developmental biology. How many ways to make a chordate? Science 312: 1145–1156
Levine M, Tjian R (2003) Transcription regulation and animal diversity. Nature 424: 147–151
Leys SP (2004) Gastrulation in sponges. In Stern CD (ed) Gastrulation. Cold Spring Harbor Laboratory Press, Cold Sping Harbor, New York
Leys SP, Degnan BM (2001) Cytological basis of photoresponsive behavior in a sponge larva. Biol Bull 201: 323–338
Leys SP, Ereskovsky AV (2006) Embryogenesis and larval differentiation in sponges. Can J Zool 84: 262–287
Maniatis T, Tasic B (2002) Alternative pre-mRNA splicing and proteome expansion in metazoans. Nature 418: 236–243
Manning G et al (2008) The protist, *Monosiga brevicollis*, has a tyrosine kinase signaling network more elaborate and diverse than found in any known metazoan. PNAS 105: 9674–9679
Manuel M, Le Parco Y (2000) Homeobox gene diversification in the calcareous sponge, *Sycon raphanus*. Mol Phylogenet Evol 17: 97–107
Manuel M et al (2004) Comparative analysis of Brachyury T-domains, with the characterization of two new sponge sequences, from a hexactinellid and a calcisponge. Gene 340: 291–301
Marlow HQ et al (2009) Anatomy and development of the nervous system of *Nematostella vectensis*, an anthozoan cnidarian. Dev Neurobiol 69: 235–254
Martindale MQ (2005) The evolution of metazoan axial properties. Nat Rev Genet 6: 917–927
Martindale MQ et al (2004) Investigating the origins of triploblasty: 'mesodermal' gene expression in a diploblastic animal, the sea anemone *Nematostella vectensis* (phylum, Cnidaria; class, Anthozoa). Development 131: 2463–2474
Martinelli C, Spring J (2003) Distinct expression patterns of the two T-box homologues Brachyury and Tbx2/3 in the placozoan *Trichoplax adhaerens*. Dev Genes Evol 213: 492–499
Martinelli C, Spring J (2004) Expression pattern of the homeobox gene Not in the basal metazoan *Trichoplax adhaerens*. Gene Expr Patterns 4: 443–447
Masuda-Nakagawa LM et al (2000) The HOX-like gene Cnox2-Pc is expressed at the anterior region in all life cycle stages of the jellyfish *Podocoryne carnea*. Dev Genes Evol 210: 151–156
Mattick JS (2007) A new paradigm for developmental biology. J Exp Biol 210: 1526–1547
Matus DQ et al (2008) The Hedgehog gene family of the cnidarian, *Nematostella vectensis*, and implications for understanding metazoan Hedgehog pathway evolution. Dev Biol 313: 501–518
Matus DQ et al (2006a) Molecular evidence for deep evolutionary roots of bilaterality in animal development. Proc Natl Acad Sci USA 103: 11195–11200
Matus DQ et al (2006b) Dorso/ventral genes are asymmetrically expressed and involved in germ-layer demarcation during cnidarian gastrulation. Curr Biol 16: 499–505

Matus DQ et al (2007) FGF signaling in gastrulation and neural development in *Nematostella vectensis*, an anthozoan cnidarian. Dev Genes Evol 217: 137–148
Mazza ME et al (2007) Genomic organization, gene structure, and developmental expression of three clustered otx genes in the sea anemone *Nematostella vectensis*. J Exp Zoolog B Mol Dev Evol 308: 494–506
McGinnis W et al (1984) A homologous protein-coding sequence in *Drosophila* homeotic genes and its conservation in other metazoans. Cell 37: 403–408
McGinnis W and Krumlauf R. (1992) Homeobox genes and axial patterning. Cell 68: 283–302
Miller DJ, Ball EE (2005) Animal evolution: the enigmatic phylum placozoa revisited. Curr Biol 15: R26–R28
Miller DJ, and Ball EE (2008) Cryptic complexity captured: the *Nematostella* genome reveals its secrets. Trends Genet 24: 1–4
Miller DJ et al (2005) Cnidarians and ancestral genetic complexity in the animal kingdom. Trends Genet 21: 536–539
Monteiro AS et al (2006) A low diversity of ANTP class homeobox genes in Placozoa. Evol Dev 8: 174–182
Moroz LL et al (2006) Neuronal transcriptome of aplysia: neuronal compartments and circuitry. Cell 127: 1453–1467
Mukherjee K, Bürglin TR (2007) Comprehensive analysis of animal TALE homeobox genes: new conserved motifs and cases of accelerated evolution. J Mol Evol 65: 137–153
Nedelcu AM, Tan C (2007) Early diversification and complex evolutionary history of the p53 tumor suppressor gene family. Dev Genes Evol 217: 801–806
Nielsen C (2001) Animal Evolution. Interrelationships of the Living Phyla. Oxford University press, Oxford
Nielsen C (2004) Trochophora Larvae: Cell-Lineages, Ciliary Bands, and Body Regions. 1. Annelida and Mollusca. J Exp Zool (Mol Dev Evol) 302B: 35–68
Nilsson DE et al (2005) Advanced optics in a jellyfish eye. Nature 435: 201–205
Pang K et al (2004) The ancestral role of COE genes may have been in chemoreception: evidence from the development of the sea anemone, *Nematostella vectensis* (Phylum Cnidaria; Class Anthozoa). Dev Genes Evol 214: 134–138
Pavlopoulos A, Averof M (2005) Establishing genetic transformation for comparative developmental studies in the crustacean *Parhyale hawaiensis*. Proc Natl Acad Sci USA 102: 7888–7893
Pedersen RA (1971) DNA content, ribosomal gene multiplicity, and cell size in fish. J. Exp. Zool. 177: 65–78
Pennacchio LA et al (2006) In vivo enhancer analysis of human conserved non-coding sequences. Nature 444: 499–502
Peterson KJ, Sperling EA (2007) Poriferan ANTP genes: primitively simple or secondarily reduced? Evol Dev 9: 405–408
Philippe H et al (2009) Phylogenomics revives traditional views on deep animal relationships. Curr Biol 19: 706–712
Piatigorsky J, Kozmik Z (2004) Cubozoan jellyfish: an Evo/Devo model for eyes and other sensory systems. Int J Dev Biol 48: 719–729
Pincus D et al (2008) Evolution of the phospho-tyrosine signaling machinery in premetazoan lineages. PNAS 105: 9680–9684
Prpic NM, Telford MJ (2008) Expression of homothorax and extradenticle mRNA in the legs of the crustacean *Parhyale hawaiensis*: evidence for a reversal of gene expression regulation in the pancrustacean lineage. Dev Genes Evol 218: 333–339
Prud'homme B et al (2007) Emerging principles of regulatory evolution. Proc Natl Acad Sci USA 104: 8605–8612
Putnam NH et al (2008) The amphioxus genome and the evolution of the chordate karyotype. Nature 453: 1064–1071

Putnam NH et al (2007) Sea anemone genome reveals ancestral eumetazoan gene repertoire and genomic organization. Science 317: 86–94
Raible F et al (2006) Opsins and clusters of sensory G-protein-coupled receptors in the sea urchin genome. Dev Biol 300: 461–475
Raible F et al (2005) Vertebrate-type intron-rich genes in the marine annelid *Platynereis dumerilii*. Science 310: 1325–1326
Rast JP et al (2006) Genomic insights into the immune system of the sea urchin. Science 314: 952–956
Remane A (1950) Die Entstehung der Metamerie der Wirbellosen. Vh Dt Zool Ges Mainz: 16–23
Rentzsch F et al (2006) Asymmetric expression of the BMP antagonists *chordin* and *gremlin* in the sea anemone *Nematostella vectensis*: Implications for the evolution of axial patterning. Dev Biol 296: 375–387
Rentzsch F et al (2008) FGF signalling controls formation of the apical sensory organ in the cnidarian *Nematostella vectensis*. Development 315:1761–1769
Ryan JF et al (2006) The cnidarian-bilaterian ancestor possessed at least 56 homeoboxes: evidence from the starlet sea anemone, *Nematostella vectensis*. Genome Biol 7: R64
Ryan JF et al (2007) Pre-bilaterian origins of the Hox cluster and the Hox code: evidence from the sea anemone, *Nematostella vectensis*. PLOS One 2: e153
Sakaraya O et al (2007) A post-synaptic scaffold at the origin of the animal kingdom. PLoS One 2(6): e506
Samanta MP et al (2006) The transcriptome of the sea urchin embryo. Science 314: 960–962
Satou Y, Satoh N (2006) Gene regulatory networks for the development and evolution of the chordate heart. Genes Dev 20: 2634–2638
Schierwater B et al (2008) The early ANTP gene repertoire: Insights from the placozoan genome. PLOS One 3: e2457
Schmucker D et al (2000) Drosophila Dscam is an axon guidance receptor exhibiting extraordinary molecular diversity. Cell 101: 671–684
Schuchert P (1993) *Trichoplax adhaerens* (Phylum Placozoa) has cells that react with antibodies against the neuropetide RFamide. Acta Zoologica (Stockholm). 74: 115–117
Schuchert P et al (1993) Life stage specific expression of a myosin heavy chain in the hydrozoan *Podocoryne carnea*. Differentiation 54: 11–18
Sea Urchin Genome Sequencing C et al (2006) The genome of the sea urchin *Strongylocentrotus purpuratus*. Science 314: 941–952
Sedgwick A (1884) On the origin of metameric segmentation and some other morphological questions. Q J Microsc Sci 24: 43–82
Segawa Y et al (2006) Functional development of Src tyrosine kinases during evolution from a unicellular ancestor to multicellular animals. Proc Natl Acad Sci USA 103: 12021–12026
Seipel K, Schmid V (2005) Evolution of striated muscle: Jellyfish and the origin of triploblasty. Dev Biol 282: 14–26
Sempere LF et al (2006) The phylogenetic distribution of metazoan microRNAs: insights into evolutionary complexity and constraint. J Exp Zoolog B Mol Dev Evol 306: 575–588
Shalchian-Tabrizi K et al (2008) Multigene phylogeny of choanozoa and the origin of animals. PLOS One 3: e2098
Shankland M, Seaver EC (2000) Evolution of the bilaterian body plan: what have we learned from annelids? Proc Natl Acad Sci USA 97: 4434–4437
Shoguchi E et al (2008) Genome-wide network of regulatory genes for construction of a chordate embryo. Dev Biol 316: 498–509
Short S, Holland LZ (2008) The evolution of alternative splicing in the Pax family: the view from the Basal chordate amphioxus. J Mol Evol 66: 605–620
Siewing R (1985) Lehrbuch der Zoologie. Systematik. Gustav Fischer Verlag, Stuttgart, New York
Signorovitch AY et al (2007) Comparative genomics of large mitochondria in placozoans. PLoS Genet 3: e13
Signorovitch AY et al (2005) Molecular signatures for sex in the Placozoa. Proc Natl Acad Sci USA 102: 15518–15522

Simionato E et al (2007) Origin and diversification of the basic helix-loop-helix gene family in metazoans: insights from comparative genomics. BMC Evol Biol 7: 33
Skogh C et al (2006) Bilaterally symmetrical rhopalial nervous system of the box jellyfish *Tripedalia cystophora*. J Morphol 267: 1391–1405
Snell EA et al (2006) An unusual choanoflagellate protein released by Hedgehog autocatalytic processing. Proc R Soc B 273: 401–407
Sperling EA, Peterson KJ. (2007) Poriferan paraphyly and its implication for precambrian paleobiology. In Vickers-Rich P, Komarower P (eds) The rise and fall of the ediacaran biota. Geological Society, London
Spring J et al (2002) Conservation of Brachyury, Mef2, and Snail in the myogenic lineage of jellyfish: a connection to the mesoderm of bilateria. Dev Biol 244: 372–384
Srivastava M et al (2008) The *Trichoplax* genome and the nature of placozoans. Nature 454: 955–960
St-Onge L et al (1997) Pax6 is required for differentiation of glucagon-producing alpha-cells in mouse pancreas. Nature 387: 406–409
Steinmetz PR et al (2007) Polychaete trunk neuroectoderm converges and extends by mediolateral cell intercalation. Proc Natl Acad Sci USA 104: 2727–2732
Stephenson TA (1928) The British Sea Anemones. Dulau & Co, London
Stephenson TA (1935) The British Sea Anemones. Dulau & Co, London
Stierwald M et al (2004) The Sine oculis/Six class family of homeobox genes in jellyfish with and without eyes: development and eye regeneration. Dev Biol 274: 70–81
Suga H et al (2008) Evolution and functional diversity of jellyfish opsins. Curr Biol 18: 51–55
Syed T, Schierwater B (2002) *Trichoplax adherens*: discovered as a missing link, forgotten as a hydrozoan, re-discovered as a key to metazoan evolution. Vie Milieu 52: 177–187
Tassy O et al (2006) A quantitative approach to the study of cell shapes and interactions during early chordate embryogenesis. Curr Biol 16: 345–358
Tautz D (2004) Segmentation. Dev Cell 7: 301–312
Taylor JS, Raes J (2004) Duplication and divergence: The evolution of new genes and old ideas. Ann Rev Genet 38: 615–643
Technau U (2001) *Brachyury*, the blastopore and the evolution of the mesoderm. BioEssays 23: 788–794
Technau U et al (2005) Maintenance of ancestral complexity and non-metazoan genes in two basal cnidarians. Trends Genet 21: 633–639
Technau U, Scholz CB (2003) Origin and evolution of endoderm and mesoderm. Int J Dev Biol 47: 47
Tessmar-Raible K et al (2005) Fluorescent two color whole-mount in situ hybridization in *Platynereis dumerilii* (Polychaeta, Annelida), an emerging marine molecular model for evolution and development. BioTechniques 39:460–464
Valentine JW (2000) Two genomic paths to the evolution of complexity in bodyplans. Paleobiology 26: 513–519
Valentine JW et al (1994) Morphological complexity increase in metazoans. Paleobiology 20: 131–142
Vogel C, Chothia C (2006) Protein family expansions and biological complexity. PLOS Comput Biol 2: e48
Voigt O et al (2004) Placozoa – no longer a phylum of one. Curr Biol 14: R944–R945
Wenderoth H (1986) Transepithelial cytophagy by *Trichoplax adherens* F.E. Schulze (Placozoa) feeding on yeast. Zeitschrift für Naturforschung. Section C, Biosciences 41: 343–347
Woo K, Fraser S (1995) Order and coherence in the fate map of the zebrafish nervous system. Development 121: 2595–2609
Woolfe A et al (2005) Highly conserved non-coding sequences are associated with vertebrate development. PLoS Biol 3: e7
Yanze N et al (2001) Conservation of Hox/ParaHox-related genes in the early development of a cnidarian. Dev Biol 236: 89–98

Chapter 6
Genomics of Marine Algae

Susana M. Coelho, Svenja Heesch, Nigel Grimsley, Hervé Moreau, and J. Mark Cock

Abstract The algae are an extremely diverse group of organisms from several different perspectives; including their phylogeny, their basic biology and biochemistry, the range of complexity they exhibit and their adaptation to a large number of different habitats. As a result, algal research touches on a broad spectrum of questions ranging from the importance of algae as key species in marine ecosystems to algae as a source of novel biomolecules and bioprocesses. Two key developments have accelerated research in this field over the past few years. The first is the application of high-throughput sequencing approaches both to whole genome sequencing and to metagenomic exploration of marine environments. The second is the emergence of model organisms in several of the less well studied algal lineages. These model organisms are important because they permit efficient application of genomic approaches to these groups. This chapter will describe how genomic research is providing new insights into many aspects of algal biology and will attempt to outline where this research is likely to lead in the future.

Abbreviations

DD	(2E,4E/Z)-decadienal
kbp	kilobase pairs
Mbp	megabase pairs
NO	nitric oxide
PCR	polymerase chain reaction
pg	picogramme
rDNA	ribosomal DNA

S.M. Coelho (✉)
UPMC Univ. Paris 06, The Marine Plants and Biomolecules Laboratory, UMR 7139, 29682, Roscoff Cedex, France
e-mail: coelho@sb-roscoff.fr

J.M. Cock et al. (eds.), *Introduction to Marine Genomics*,
Advances in Marine Genomics 1, DOI 10.1007/978-90-481-8639-6_6,

6.1 What Are Algae?

The term "algae" does not have a precise taxonomic meaning and different authors have used this term to refer collectively to different groups of photosynthetic organisms. It is, therefore, important that we begin this chapter by defining what we mean by this term. Here we define algae as all photosynthetic eukaryotes other than the embryophytes (or land plants; Keeling 2004). This definition does not include photosynthetic prokaryotes. Nonetheless, it still covers a very broad range of taxa because it includes all the groups of organisms from across the eukaryotic tree that have acquired, by one means or another, a chloroplast. Note that, despite being referred to as land plants, the embryophytes include some species that have returned to the marine environment, such as seagrasses. We do not define these organisms as algae, although they are discussed briefly later in the chapter.

The logic behind treating such a diverse collection of organisms as a group is based on the origin of their plastids. The plastids of all eukaryotic algae are all thought to be derived, either directly or indirectly, from a single primary endosymbiotic event that occurred in a common ancestor of the green, red and glaucophyte algae (see below).

6.2 Why Algae Are Interesting

Algal research is motivated by a number of different aspects of the biology, ecology and evolutionary history of this group of organisms. Many unicellular marine algae are constituents of the phytoplankton and these organisms make important contributions to global geochemical cycles. Phytoplankton algae are also key primary producers in marine food chains. Along with these essentially positive roles, algae can also have a dark side, for example when toxins are produced by algal blooms. In coastal ecosystems, red, green and brown macroalgae are usually the most obvious components, sometimes forming dense forests that cover an impressive area of the seabed. These organisms represent a major component of coastal ecosystems, being classified as ecosystem engineers (Coleman and Williams 2002). For instance, large kelp forests not only act as sites of primary production, but also contribute significantly to secondary production by marine food webs and provide habitats to numerous animal and algal species (Duggins et al. 1989). The phylogenetic diversity of the algae is also a major point of interest. Many algae occupy key phylogenetic positions in the tree of life and can provide important insights into many of the more ancient events that occurred during the evolution of the eukaryotes. This phylogenetic diversity is also of interest with respect to the general biology of these organisms. As a result of their very divergent evolutionary histories the different major groups of algae often exhibit innovations in terms of their metabolisms, cell biology, morphology, life histories, etc. These diverse characteristics are of interest both from a fundamental point of view and as a potential source of novel molecules and procedures for a wide range of biotechnological applications.

In recent years, the application of genomic approaches has led to significant advances in many areas of algal research. In many instances the application of

genomic approaches has led to surprising and exciting new insights into algal biology. The following sections will look at several domains of algal research in detail and will discuss the impact that genomic approaches have had in these areas.

6.3 Endosymbiosis and the Origins of the Algae

At first view it is surprising that algae are so widely dispersed through the eukaryotic tree of life; photosynthetic organisms are found in four of the six eukaryotic "supergroups" (in the plantae, the chromalveolates, the excavates and the rhizaria, but not in the opistokonts and the amoebozoa; Fig. 6.1). Considerable progress has

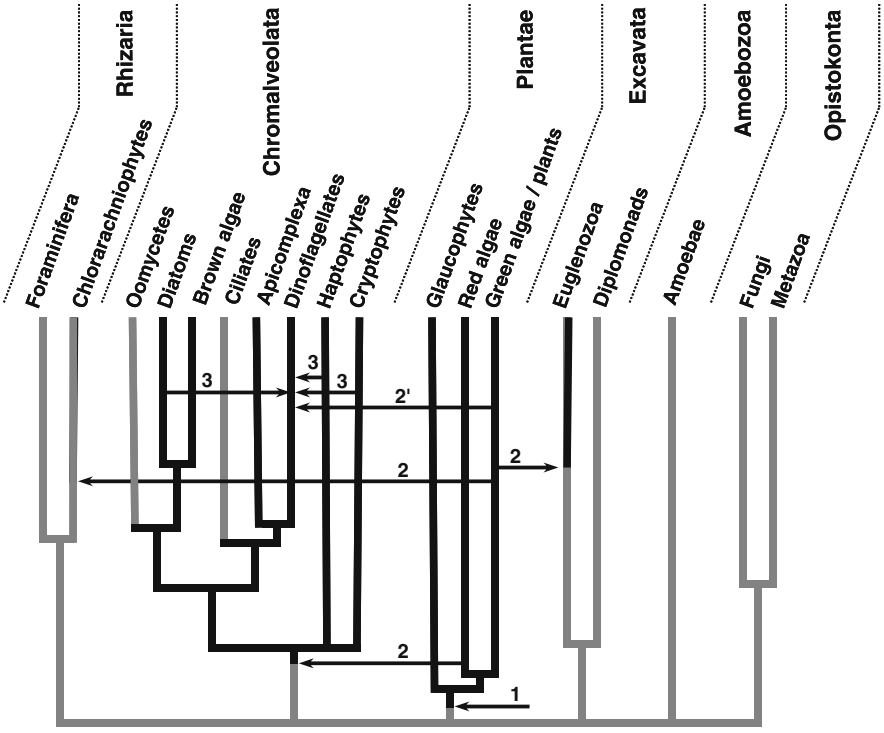

Fig. 6.1 Acquisition of plastids through endosymbiosis during the evolution of the eukaryotes. *Thick bars* indicate the evolutionary relationships between the groups shown; these are in *grey* for lineages without plastids and in *black* for lineages with plastids. *Horizontal arrows* indicate the endosymbiotic events that led to plastid acquisition. 1. Primary endosymbiosis involving the capture of a cyanobacterium by a heterotrophic ancestor of the Plantae group, 2. secondary endosymbiosis involving the enslavement of either a red or a green alga, 3. tertiary endosymbiosis involving the replacement of a plastid from a secondary endosymbiosis with another plastid also derived from a secondary endosymbiosis, 2′. serial secondary endosymbiosis in which the process of secondary endosymbiosis has occurred twice in the same lineage, the second event leading to the elimination of the plastid from the first event. Note that the relationships between the major groups of eukaryotes are poorly supported in many cases. There is accumulating evidence, for example, that the Rhizaria fall within the Chromalveolate group (Yoon et al. 2008, Baldauf 2008)

been made in recent years in understanding how this came about and genomic-scale analyses have played an important part in this story (Reyes-Prieto et al. 2007). It is generally thought that all plastids are derived from a single primary endosymbiosis involving the capture of a coccoid cyanobacterium by a heterotrophic eukaryote about 1.6 billion years ago (Yoon et al. 2006; it should be mentioned here, however, that there is recent evidence for another independent, and more recent, primary endosymbiotic event in the filose thecamoeba *Paulinella chromatophora*, Nowack et al. 2008). The enslavement of the cyanobacterium involved a gradual reduction in the size of the bacterial genome over evolutionary time, and transfer of the endosymbiont's genes to the host nucleus. This process necessitated the evolution of a protein targeting system allowing the proteins encoded by these genes to be transported into the plastid across the two surrounding membranes.

The ancient primary endosymbiotic event gave rise to three different groups within the group Plantae: the glaucophyte, the red and the green lineages. All the algae in the other supergroups (the chromalveolates, the excavates and the rhizarians) are derived from either secondary or tertiary endosymbioses involving a eukaryotic cell capturing a photosynthetic eukaryote from either the red or the green lineages, which then evolved, as in the primary endosymbiosis, to become a plastid (Keeling 2004, Yoon et al. 2006).

The chromalveolate hypothesis postulates that haptophytes, cryptophytes, heterokonts (or stramenopiles) and alveolates form a supergroup and that the capture of a red alga occurred in a common ancestor of these lineages leading to the plastids in present day groups such as diatoms, brown algae, some dinoflagellates, coccolithophores, etc.. The apicoplast of apicomplexans would also have been derived from such an event. If, as this hypothesis postulates, a secondary plastid was acquired very early in a common ancestor, then it would have had to have been lost from groups like the ciliates and the oomycetes which do not possess plastids. There is some evidence for this, as genes that could potentially have been derived from such a secondary endosymbiotic have been identified in both ciliates and oomycetes (Tyler et al. 2006, Reyes-Prieto et al. 2008). It should be noted, however, that there is still considerable debate about the chromalveolate hypothesis and it is not possible to rule out alternative hypotheses in which individual lineages of the chromalveolate group independently acquired their secondary plastids. Moreover, the chromalveolate taxon is not strongly supported by phylogenies based on nuclear genes so it is also important to bear in mind the possibility that this group is an artefact, for which the strongest support may actually be based on similar but independent secondary endosymbiosis events (Parfrey et al. 2006).

Convincing evidence for secondary endosymbiosis can be found in cryptophytes and chlorarachniophytes, which have retained a remnant of the nucleus of the endosymbiont, called a nucleomorph (Archibald 2007). Their plastids are surrounded by four membranes (the two outer membranes presumably originally corresponded to the host vesicle membrane and cell membrane of the endosymbiont) and the nucleomorph is found between these two membranes and the two inner membranes. The presence of 4 membranes is a common feature of

secondary plastids, except in euglenids and most dinoflagellates, which appear to have lost one of these membranes. The presence of these additional membranes makes transporting proteins into the plastid very complicated, and organisms whose plastids are derived from a secondary endosymbioses possess complex transport systems that recognise proteins with bipartite target peptides (Lang et al. 1998).

As if the story of secondary endosymbiosis were not sufficiently complicated, some dinoflagellates appear to have undergone even more complex events involving either serial secondary endosymbiosis, in which the red-alga-derived plastid was replaced with a green-alga-derived plastid, or they have been involved in tertiary endosymbioses, capturing a secondary endosymbiotic haptophyte, cryptophyte or heterokont alga (Keeling 2004). The chlorarachniophytes and euglenophytes obtained their plastids in independent secondary endosymbiotic events involving the capture of green algae (Fig. 6.1).

Horizontal (or lateral) gene transfer is a process in which genes from one organism are integrated into the genome of a second organism (which may be very distantly related to the first) and subsequently inherited with the rest of the genetic material of the cell (Keeling and Palmer 2008). The gene transfers associated with endosymbiotic events are spectacular examples of horizontal gene transfer because of the large numbers of genes that are transferred from the endosymbiont to the host nucleus. However, they are not the only horizontal gene transfers that occur in eukaryotes and it has become increasingly clear in recent years that many eukaryotes, particularly protists, acquire genes from the organisms with which they interact at a significant rate (e.g. Nosenko and Bhattacharya 2007, Bowler et al. 2008).

In the last few years genomic data has had a significant impact on our understanding of the evolution of the eukaryotes and the role that endosymbiosis has played in this process. The availability of extensive sequence data for many key organisms in the eukaryotic tree has allowed the selection of the most relevant gene sequences for phylogenetic analyses and their combination in multigene sets, significantly improving the resolution of these analyses. Genome data also provides detailed information about the process of endosymbiosis, providing information about the partners involved in a particular endosymbiotic event and also about the process of enslavement, particularly the transfer of endosymbiont genes to the host nucleus. Some of the key genome projects in this domain are listed in Table 6.1 and described below.

6.4 Algae and Marine Ecosystems

The following sections will look at how genomic approaches are being used to explore the biology of algae, particularly with regard to their functions in a broad range of marine ecosystems.

Table 6.1 Algal genome sequencing projects

Species	Strain	Phylogenetic group	Marine	Genome size (Mbp)	Status or Reference	URL
Bathycoccus prasinos	Bban7	Plantae, Prasinophyceae (green alga)	Yes		Pending	http://www.cns.fr/externe/English/corps_anglais.html
Ostreococcus tauri	OTH95	Plantae, Prasinophyceae (green alga)	Yes	12.6	Derelle et al. (2006)	http://bioinformatics.psb.ugent.be/genomes/
Ostreococcus "lucimarinus"	CC9901	Plantae, Prasinophyceae (green alga)	Yes	13.2	Palenik et al. (2007)	http://www.jgi.doe.gov/genome-projects/
Ostreococcus sp.	RCC809	Plantae, Prasinophyceae (green alga)	Yes		Available	http://www.jgi.doe.gov/genome-projects/
Micromonas pusilla	RCC827	Plantae, Prasinophyceae (green alga)	Yes	15	Worden et al. (2009)	http://www.jgi.doe.gov/genome-projects/
Micromonas pusilla	CCMP1545	Plantae, Prasinophyceae (green alga)	Yes	15	Worden et al. (2009)	http://www.jgi.doe.gov/genome-projects/
Dunaliella salina	CCAP 19/18	Plantae, Chlorophyceae (green alga)	No	130	Sequencing	http://www.jgi.doe.gov/genome-projects/
Chlorella vulgaris	C-169	Plantae, Trebouxiophyceae (green alga)	No	40	Annotation	http://www.jgi.doe.gov/genome-projects/
Chlorella sp.	NC64A	Plantae, Chlorellaceae (green alga)	No	46.2	Available	http://genome.jgi-psf.org/ChlNC64A_1/ChlNC64A_1.home.html

Table 6.1 (continued)

Species	Strain	Phylogenetic group	Marine	Genome size (Mbp)	Status or Reference	URL
Chlamydomonas reinhardtii	CC-503 cw92 mt+	Plantae, Chlorophyceae (green alga)	No	120	Merchant et al. (2007)	http://genome.jgi-psf.org/Chlre3/Chlre3.home.html
Volvox carteri		Plantae, Chlorophyceae (green alga)	No	140	Available	http://genome.jgi-psf.org/Volca1/Volca1.home.html
Zostera marina		Plantae, Zosteraceae (sea-grass)	Yes		Pending	http://www.jgi.doe.gov/genome-projects/
Cyanophora paradoxa		Plantae, Glaucophyta	No		Assembly	http://www.biology.uiowa.edu/cyanophora/cyanophora_home.htm
Cyanidioschyzon merolae	10D	Rhodophyta, Cyanidiaceae (red alga)	No	16.5	Matsuzaki et al. (2004)	http://merolae.biol.s.u-tokyo.ac.jp/
Galdieria sulphuraria		Rhodophyta, Cyanidiaceae (red alga)	No		Ongoing	http://genomics.msu.edu/galdieria
Chondrus crispus		Rhodophyta, Florideophyceae (red alga)	Yes	150	Ongoing	http://www.genoscope.cns.fr/spip/spip.php?lang=en
Porphyra umbilicalis		Rhodophyta, Bangiophyceae (red alga)	Yes	300–400?	Ongoing	http://www.jgi.doe.gov/genome-projects/
Thalassiosira pseudonana	CCMP1335	Heterokonta, Bacillariophyceae (diatom)	Yes	34.5	Armbrust et al. (2004)	http://www.jgi.doe.gov/genome-projects/

Table 6.1 (continued)

Species	Strain	Phylogenetic group	Marine	Genome size (Mbp)	Status or Reference	URL
Phaeodactylum tricornutum	CCAP1055/1	Heterokonta, Bacillariophyceae (diatom)	Yes	20	Bowler et al. (2008)	http://www.jgi.doe.gov/genome-projects/
Pseudo-nitzschia multiseries	CLN-47	Heterokonta, Bacillariophyceae (diatom)	Yes	250	Sequencing	http://www.jgi.doe.gov/genome-projects/
Fragilariopsis cylindrus	CCMP 1102	Heterokonta, Bacillariophyceae (diatom)	Yes	35	Sequencing	http://www.jgi.doe.gov/genome-projects/
Aureococcus anophagefferens	CCMP1984	Heterokonta, Pelagophyceae	Yes	32	Available	http://genome.jgi-psf.org/Auran1/Auran1.home.html/
Ochromonas	CCMP1393	Heterokonta, Chrysophyceae	No		Ongoing	http://www.jgi.doe.gov/sequencing/why/nanoflagellates.html
Ectocarpus siliculosus	Ec 32	Heterokonta, Phaeophyceae (brown alga)	Yes	200	Completed	http://www.genoscope.cns.fr/spip/Ectocarpus-siliculosus.html
Phaeocystis globosa		Haptophyta, Prymnesiophyceae	Yes		Pending	http://www.jgi.doe.gov/genome-projects/
Phaeocystis antarctica		Haptophyta, Prymnesiophyceae	Yes		Pending	http://www.jgi.doe.gov/genome-projects
Emiliania huxleyi	CCMP1516	Haptophyta, Prymnesiophyceae	Yes	220	Available	http://www.jgi.doe.gov/genome-projects/
Guillardia theta	CCMP2712	Cryptophyta, Cryptophyceae	Yes		Assembly	http://www.jgi.doe.gov/sequencing/why/50026.html/
Bigelowiella natans	CCMP2755	Rhizaria, Chlorarachniophyceae	Yes		Assembly	http://www.jgi.doe.gov/sequencing/why/50026.html

6.4.1 Diversification of the Phytoplankton During the Evolution of the Earth

The term plankton refers to the free-floating organisms found in diverse marine and fresh-water environments. The photosynthetic organisms in these ecosystems, referred to collectively as the phytoplankton, represent only a small percentage of the total global primary producer biomass (0.2%) and yet they make a very significant contribution to the planet's primary production (estimated at 45% of the global primary production; Field et al. 1998). Many eukaryotic algae are found in the phytoplankton but the most abundant phytoplankton organisms are cyanobacteria. From an evolutionary point of view cyanobacteria are also the oldest component of the phytoplankton. These organisms were responsible for the "invention" of oxygenic photosynthesis and hence for the transition from the ancient anaerobic world to the aerobic world of the modern Earth. As described above, the endosymbiotic enslavement of these cyanobacteria resulted in the emergence of photosynthetic eukaryotes and this led to an increase in the complexity of the phytoplankton. Fossil records indicate that red algae have existed for at least 1.2 billion years ago and there is some evidence that eukaryotic phytoplankton may have been present as long ago as 1.6–1.8 billion years (in Falkowski et al. 2004). Green algae are likely to have been prominent in the marine phytoplankton during the mid-Paleozoic area but the abundance and diversity of this group started to decline during the Triassic and they were replaced to a large extent by the three dominant algal groups of modern phytoplankton ecosystems, the dinoflagellates, the coccolithophores and the diatoms (Falkowski et al. 2004). All three of these latter groups are derived from secondary endosymbioses, in most cases involving the capture of a red alga. This suggests that the possession of a red-alga-derived plastid may have conferred an advantage, perhaps related to the requirements for trace elements (Falkowski et al. 2004). In more recent geological time, there have also been fluctuations in the relative abundances of dinoflagellates, coccolithophores and diatoms and these are thought to be related to niche preference, with dinoflagellates and coccolithophores being better adapted to stable environments, whilst diatoms tend to thrive in rapidly changing environments where nutrients are supplied with high pulse frequencies.

6.4.2 Algae Are Important Components of the Phytoplankton

It has been known for many years that phytoplankton communities are complex in terms of the phylogenetic range of the organisms present. This has been seen as a paradox, conflicting with models in which these organisms were thought to be competing for limited resources in fairly uniform ecological niches (Hutchinson 1961). Several hypotheses have been put forward to explain this situation, including some propositions by Hutchinson himself, but efforts to resolve this paradox are handicapped by the fact that the ecosystems themselves remain very poorly characterised.

Over the last decade, the development of molecular and, later, genomic techniques to investigate the composition of plankton communities and to study individual phytoplankton species has had an important impact in this domain. These techniques have provided an alternative means to study planktonic organisms, many of which cannot currently be isolated in culture. Sequencing of rDNA cloned directly from environmental samples has proved to be a powerful tool to study the phylogenetic diversity of eukaryotic planktonic species. This "environmental cloning" approach was originally pioneered for prokaryotic planktonic organisms (Giovannoni et al. 1990) and it has been used much more extensively for this group than for the eukaryotes. Nonetheless, the studies that have been carried out on planktonic eukaryotes have already revealed a remarkably high level of diversity and have led to the discovery of several new eukaryotic groups, particularly from amongst the heterokonts and the alveolates (reviewed in Moreira and López-García 2002). Moreover, these studies have only sampled part of the diversity present in these ecosystems and deeper sequencing is expected to reveal many more novel sequences in the future. One of the current challenges is to link these environmental sequences to identifiable organisms in order to gain some understanding of the ecological roles of these newly discovered species (Massana et al. 2002). For this it will be important both to develop improved culture techniques and to improve in vivo detection methods based on techniques such as fluorescence in situ hybridisation (FISH). An important question, for example, for the new heterokonts and alveolates mentioned above is whether they are autotrophs or heterotrophs.

PCR amplification and cloning of rDNA sequences from environmental samples clearly allows access to a broad range of organisms. Other similar genomic approaches are being developed, such as the use of oligonucleotide microarrays to detect organisms and assess biodiversity in environmental samples (Medlin et al. 2006, Metfies and Medlin 2008). However, these methods are not without biases of their own. In particular, because they are based on PCR, the similarity between a DNA sequence and the degenerate primers used will influence whether that sequence is detected. In recent years, random, high-throughput sequencing of cloned DNA fragments has been developed as an alternative approach to explore the genetic complexity of a number of different ecosystems. Chapter 1 provides a broad overview of how this approach is being used to explore marine systems. Here we will concentrate on a specific example to illustrate how this approach can be used to characterise planktonic ecosystems and its potential for the study of the algal component of the phytoplankton.

6.4.3 Exploration of Planktonic Ecosystems Using High-Throughput Sequencing

Craig Venter and colleagues (Venter et al. 2004, Rusch et al. 2007) used a high-throughput sequencing approach to analyse samples of micro-organisms filtered (0.8 μm) from seawater that had been collected from several sites in the Sargasso

sea and, in a later study, from a transect that extended along the eastern border of the North Atlantic through the Panama canal to the South Pacific (the Sorcerer II Global Ocean Survey). In total, more than 9.3 million sequencing reads were carried out on 2–6 kbp DNA fragments cloned from these samples, generating more than 6.3 billion base pairs of sequence. This approach provides a wealth of information about the diversity of organisms present in a particular environment. For example, the Sargasso Sea samples alone were estimated to contain samples from 1,800 species including 148 previously unknown bacterial phylotypes. The interest of high-throughput sequencing is not limited to the investigation of species diversity, however. The sequence data also furnishes information at the gene level, leading to the discovery of new genes and providing an overview of the ensemble of genes present in a community. This global collection of genes, often described as the metagenome, can provide important insights, for example, into the different functional pathways that are represented and relative importance of each pathway in a particular environment.

The studies carried out by Craig Venter and colleagues focused on the bacterial components of the ecosystems studied. To do this they employed filters which both eliminated smaller entities, such as viruses, and larger organisms, such as large eukaryotic cells (Venter et al. 2004, Rusch et al. 2007). The samples did however contain a significant proportion of picoeukaryotes, which have approximately the same size as bacteria (2.8% of the sequences in the Sorcerer II study). These sequences represent species that are broadly dispersed throughout the eukaryotic tree of life (Piganeau et al. 2008). The studies have therefore provided a considerable amount of sequence data and information about planktonic algae. Perhaps more importantly, however, these two analyses clearly show the potential of high-throughput sequencing as a means to explore the diversity of marine algae and the ecosystems in which they live. Moreover, samples at other filter sizes were collected during these surveys so it will be possible to carry out a more complete analysis of the algal component of these ecosystems at a later date.

As with all novel approaches, high-throughput environmental sequencing also creates new problems to be solved. In particular, the assembly of sequence data from ecosystems with a high level of biodiversity such as the plankton can pose a serious challenge because, despite the large number of sequences generated, the depth of sequencing is much less than when the sequencing is focused on a single genome. In this respect, more directed, organism-based approaches can provide information that is highly complementary to that generated by environmental sequencing. For example, two complete genomes have been published for prasinophytes of the genus *Ostreococcus* since the data from the Sargasso Sea survey was made available. These genome sequences have been used to recover related sequences from the Sargasso Sea dataset and it has been possible to show that at least two species of *Ostreococcus* were represented in these samples (Piganeau and Moreau 2007). The two types of information are highly complementary, the genome sequences allowing the recovery of species-specific data from the environmental dataset and the environmental dataset providing additional information about both the ecology of the organism that has been sequenced and about the molecular evolution of its genome, based on sequence comparisons.

6.4.4 Diversity and Dynamics of Planktonic Ecosystems

High-throughput sequencing has been applied to only a limited number of samples so far and there is still an enormous potential for identifying new organisms and understanding new ecosystems. High-throughput sequencing has confirmed that individual planktonic ecosystems are remarkably diverse, but we also know that there are significant differences between planktonic ecosystems in different areas of the marine environment.

The oligotrophic environments of the open ocean differ markedly from mesotrophic coastal areas (Guillou et al. 1999, Massana et al. 2004, Romari and Vaulot 2004) and the communities of organisms found in arctic waters differ significantly from those of warmer waters (Lovejoy et al. 2006). One environment that is attracting a lot of attention is sea-ice (Mock and Thomas 2005). The organisms that are able to live in the extreme conditions of cold and highly variable salinity, pH and light conditions found in this ecosystem are of interest for a number of reasons ranging from biotechnological applications to understanding the origin of life on Earth. The coming years will almost certainly see a more widespread application of high-throughput sequencing to these diverse marine ecosystems.

Another important consideration is the dynamics of planktonic ecosystems. The density and the types of organisms in these ecosystems can be highly variable, being affected both by physical and chemical constraints (such as light, temperature and nutrient availability) and by interactions with other organisms. Predation by other phytoplankton such as nanoflagellate protists or cellular lysis caused by viral infection are two major forces impacting on phytoplankton populations. Predation may cause nutrients such as carbon to be sequestered into sinking particles or accumulated in larger animals, whereas viral lysis is likely to release more nutrients back into the environment (Suttle 2005). Collectively, the planet's oceans are thought to contain about 4×10^{30} virus particles, probably including pathogens of most, if not all, other marine organisms. These predators are therefore likely to play a critical role in marine carbon cycling but little is known about their basic biology. Marine microalgae are infected by a wide range of viruses (Nagasaki 2008) and genomic approaches have started to be used to study some of these, such as the Coccolithovirus EhV-86, which infects *Emiliania huxleyi* (Wilson et al. 2005), and OtV5, which infects the prasinophyte *Ostreococcus tauri* (Derelle et al. 2008). Moreover, the Joint Genome Institute (California, USA) has recently initiated sequencing of the genomes of heterotrophic nanoflagellates that predate microalgae and some uncultivated viruses, so more information about these important organisms can be expected in the near future.

Algal blooms are particularly impressive examples of the dynamic nature of planktonic ecosystems. These blooms, which involve rapid multiplication of one or a small number of phytoplankton species, can have a significant impact on human activity, particularly when the blooming alga produces toxins. Many toxic algal blooms are caused by rapid growth of dinoflagellates in response to favourable environmental conditions such as high temperatures, high nutrient concentrations and a stagnant water column during the summer months. These so-called "red

tides", which can contain up to several million cells per litre of seawater (Guiry and Guiry 2008), can have a huge impact on the environment through the bioaccumulation of their toxins in the food chain, affecting fish, birds and mammals. High toxin concentrations in fish and filter-feeding molluscs consumed by humans may cause gastrointestinal disorders, permanent neurological damage or even result in death (Faust and Gulledge 2002). *Alexandrium tamarense* and *A. catenella* produce saxitoxins leading to paralytic shellfish poisoning (PSP), while the brevetoxin of *Karenia brevis* is responsible for neurolytic shellfish poisoning (NSP) (Faust and Gulledge 2002). The fish mortality associated with blooms of *Heterocapsa triquetra*, on the other hand, appears to be due to oxygen depletion of the water when cells decompose, rather than toxin production (Guiry and Guiry 2008). The cytotoxins of *Amphidinium gibbosum* have been studied for their potent antitumor activities (Bauer et al. 1995). Additional toxin-producing groups include the heterokonts and the haptophytes. Some species of the diatom *Pseudo-nitzschia* (e.g. *P. pseudodelicatissima*), for example, produce domoic acid, a neurotoxin that may cause amnesic shellfish poisoning (AMS) in humans when consuming contaminated molluscs. The pelagophyte *Aureococcus anophagefferens* is thought to secrete a toxin into its extracellular polysaccharide sheath. The haptophyte *Prymnesium parvum* produces prymnesin toxins, which are highly toxic to fish (La Claire 2006). Genomic approaches such as EST sequencing are being used to study the organisms that cause toxic algal blooms, with a major aim being to understand how and why toxins are produced (http://genome.imb-jena.de/ESTTAL/cgi-bin/Index.pl).

6.4.5 Organism-Based Approaches for Exploring the Biology of Planktonic Algae

The approaches discussed above are providing increasingly detailed information about the organisms that are present in the plankton. By describing the ensemble of genes present in the ecosystem (the metagenome), high-throughput sequencing is also providing insights into the metabolic and cellular process that are going on in these ecosystems. However, to really understand how these communities work it is also necessary to have detailed information about the biology of the individual species that they are composed of. Classical biological techniques, in particular isolation and study of individual strains of phytoplankton in culture, have, and still are, making extremely important contributions to understanding phytoplankton biology (Vaulot et al. 2008). In recent years, however, these techniques have been complemented by powerful new genomic approaches based on whole genome sequencing and the establishment of model organisms.

Initially, genome sequencing was only applied to a very limited number of algae. These were selected primarily based on their small genome sizes, although additional factors such as ecological relevance, phylogenetic position and availability of axenic cultures and other biological characteristics, were also taken into account. However, as sequencing technologies have improved, the number of algal genome

projects has been increasing rapidly, and quite a large number of projects have been recently completed or are nearing completion (Table 6.1, Fig. 6.2). To date, eight algal genome sequences have been published. These are all from unicellular algae and include six marine species, the diatoms *Thalassiosira pseudonana* (Armbrust et al. 2004) and *Phaeodactylum tricornutum* (Bowler et al. 2008), two green prasinophyte algae *Ostreococcus tauri* (Derelle et al. 2006) and *O. lucimarinus nomen nudum* (Palenik et al. 2007) and two *Micromonas* isolates (Worden et al. 2009), plus two freshwater species, the red alga *Cyanidioschyzon merolae* (Matsuzaki et al. 2004) and the green alga *Chlamydomonas reinhardtii* (Merchant et al. 2007). Apart from *C. reinhardtii,* which has quite a large genome (120 Mbp), these algae all have small genomes, ranging from about 12 Mbp for *Ostreococcus* to 34 Mbp for *Thalassiosira*. The following sections will look in detail at the application of genomic approaches to specific marine microalgae, with a particular emphasis on the diatom *P. tricornutum* and the prasinophyte *O. tauri*, which are being developed as model organisms.

6.4.5.1 Diatom Genomics

The centric diatom *T. pseudonana* was the first marine microalga to be sequenced (Armbrust et al. 2004). This phytoplankter is ecologically important and is distributed throughout the world's oceans. Diatoms are heterokonts, and are therefore only very distantly related to both animals and green plants. As a result, the genome was found, rather surprisingly for a photosynthetic organism, to share characteristics with both of the latter groups. This was true both at the whole genome level and in terms of specific metabolic processes. For example, *Thalassiosira* was found to possess a complete urea cycle, a typical feature of animals.

The *Thalassiosira* genome sequence is now being exploited to investigate some of the more exotic features of diatom biology. Many diatoms are able to build intricately patterned silica cell walls. The fabrication of these walls is of great interest both as a biological process and because of potential nanotechnological applications. Recently, a whole genome tiling array approach was used to identify genes

Fig. 6.2 Some examples of marine algae for which genome sequencing projects have been carried out or are currently in process. (**a**) *Ostreococcus* sp., (**b**) *Batycoccus* sp. (photograph courtesy of Marie-Josèphe Dinet), (**c**) *Micromonas*sp. (from Guillou et al. 2004), (**d**) *Emiliania huxleyi* (photograph courtesy of Jeremy R. Young, The Natural History Museum, London, UK), (**e**) *Guillardia theta* (photograph courtesy of Geoff McFadden, University of Melbourne, Australia), (**f**) *Fragilariopsis cylindrus* (photograph courtesy of Gerhard Dieckmann Alfred Wegener Institute for Polar and Marine Research, Bremerhaven, Germany), (**g**) *Thalassiosira pseudonana* (photograph courtesy of Virginia Armbrust, University of Washington, USA), (**h**) *Phaeodactylum tricornutum* (photograph courtesy of Alessandra De Martino and Chris Bowler, Ecole National Supérieur, Paris, France), (**i**) *Pseudo-nitzschia multiseries* (photograph courtesy of the Joint Genome Institute, USA), (**j**) detail of *Ectocarpus siliculosus* thallus showing release of meiospores from a uniocular sporangium. (**k**) *Chondrus crispus* plantlet (Photograph courtesy of Jonas Collén, Station Biologique de Roscoff)

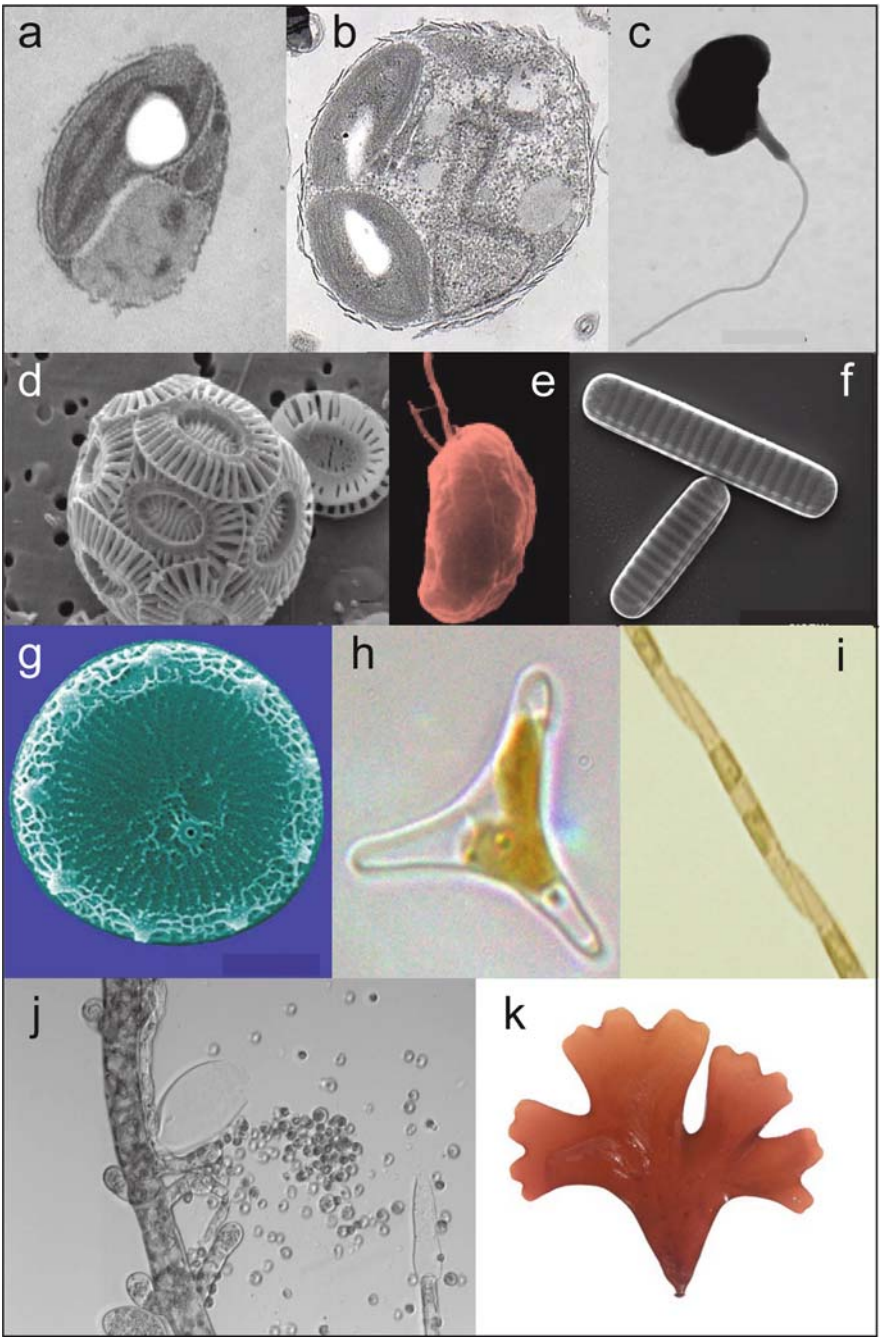

Fig. 6.2 (continued)

with a possible role in the formation of these cell walls. This approach involves the construction of a microarray bearing oligonucleotides corresponding to contiguous sequences along the length of each chromosome so that transcription of any part of the genome can be detected by hybridising with labelled cDNA. The *Thalassiosira* tiling array oligonucleotides corresponded to 36 nucleotide-long regions separated by 10 base pair gaps (Mock et al. 2008). This experiment identified a set of 75 genes that were induced under conditions of limiting silica. Another set of genes was induced when both silica and iron were limiting, showing possible interactions between these two metabolisms or, alternatively, that iron is a required cofactor for silica metabolism. Interestingly, these experiments also identified 3,000 new genes, which had not been annotated previously, including non-coding and antisense RNAs.

Diatoms fall into two major morphological groups, centric and pennate (*Thalassiosira* being a centric diatom), and recently a second genome sequence has been completed for the pennate diatom, *P. tricornutum* (Bowler et al. 2008). One of the most striking outcomes of the analysis of this second diatom genome was the identification of hundreds of genes that appear to have been acquired from bacteria by horizontal gene transfer. These genes represent at least 5% of the genome, suggesting that the extent of horizontal gene transfer has been an order of magnitude higher in the diatoms than in other free-living eukaryotes that were the subject of previous studies.

The availability of genome data for two diatom species has permitted comparative analyses of several aspects of diatom metabolism including nitrogen and carbon metabolism and carotenoid biosynthesis (Allen et al. 2006, Coesel et al. 2008, Kroth et al. 2008). Several observations suggest that the growth of diatoms in their marine environments is not limited by the available CO_2. The genomes of both *T. pseudonana* and *P. tricornutum* potentially encode all of the enzymes necessary for C_4-photosynthesis (Kroth et al. 2008) raising the intriguing possibility that, despite its high energetic cost, C_4-photosynthesis could give a critical ecological advantage in CO_2-limiting conditions, such as in phytoplankton blooms. Moreover, recent experimental work indicates that a wide range of diatoms may use C_4-based CO_2-concentrating mechanisms (McGinn and Morel 2008).

One strong argument for sequencing the *Phaeodactylum* genome was the amenability of this organism to laboratory experimentation. It can be transformed (Zaslavskaia et al. 2000) and a number of molecular tools have been developed to analyse gene expression and gene function (Siaut et al. 2007). A recent study of aldehyde-based cell-to-cell signalling illustrates how such experimental approaches can provide insights into processes that may play key roles in diatom ecology in the oceans. Diatoms are known to produce reactive aldehydes as defence molecules (Ianora et al. 2004). Addition of the aldehyde (2E,4E/Z)-decadienal (DD) to laboratory cultures of *Phaeodactylum* induced intracellular calcium transients and the generation of nitric oxide (NO), inducing cell death. Diatoms treated with a low concentration of the aldehyde, however, developed a resistance to the toxic effects of the higher dose. These observations indicate the existence of a system that allows

communication between diatoms via a secreted molecule that may have important implications for the population dynamics of algal blooms.

Phaeodactylum is also being used as an experimental system to investigate the import of proteins into diatom plastids. As mentioned above, this is a complex process because of the four concentric membranes surrounding the plastid. Targeting of nucleus-encoded proteins to the plastid has been shown to involve a bipartite target sequence (Lang et al. 1998). Large-scale sequence data has allowed the comparison of many plastid-targeted proteins leading to the identification of a conserved motif within the bipartite target sequence (Kilian and Kroth 2005). This conserved motif has been investigated experimentally by expressing wild type and mutant fusion proteins in *Phaeodactylum* cells.

Phaeodactylum also has potential for biotechnological applications. An important advance in this area was the demonstration that transformation of this diatom with a gene encoding a glucose transporter allowed it to grow on a carbon source in the dark (Zaslavskaia et al. 2001). This modification allows the use of simpler culture conditions than those necessary for light dependent growth, particularly for large-scale cultures.

The combination of genomic data and molecular tools for post-genomic analysis of gene function is likely to provide many other insights into diatom biology in the coming years. In parallel, additional diatom genome-sequencing projects are underway that will provide a deeper understanding of the role of diatoms in particular environments. Examples of the latter include projects to sequence the genomes both of a toxic bloom diatom, *Pseudo-nitzschia multieries,* and of a sea ice diatom *Fragilariopsis cylindrus* (Table 6.1). From a comparative point of view, it is also important to mention that other heterokont genome sequences are available. These include two genomes from the non-photosynthetic oomycete group (Tyler et al. 2006) and the genome of the brown macroalga *Ectocarpus siliculosus* (see below). The genome of another heterokont microalgae, the pelagophyte *Aureococcus anophagefferens,* has also been completed recently (Table 6.1).

6.4.5.2 Prasinophyte Genomics

The prasinophytes are an ancient lineage of the Plantae and include abundant members of the marine phytoplankton. The first prasinophyte genome to be sequenced was that of *O. tauri* (Derelle et al. 2006), followed rapidly by *Ostreococcus lucimarinus* (Palenik et al. 2007). This genus is remarkable for several reasons; *Ostreococcus* spp. are the smallest known free-living eukaryotes with a diameter of less than 0.8 μm and each cell contains only one mitochondria and one chloroplast. Their genomes are also extremely small (12.56 Mbp for *O. tauri*) and highly compact. For example, intergenic sequences in *O. tauri* are only 196 bp long on average. The compaction of *Ostreococcus* genomes appears to have occurred under significantly different conditions to those that have led to the extremely reduced genomes of parasites like the microsporidian *Encephalitozoon cuniculi,* in as far as the former have retained a very complete set of genes but have reduced their genome

size essentially by eliminating duplicate genes and reducing the amount of non-coding (intergenic and intronic) DNA in their genomes (Keeling 2007). Genome streamlining in parasitic organisms, on the other hand, tends to involve the elimination of many genes that are non-essential because their function is carried out by the host. One mysterious feature in both *Ostreococcus* species is the presence of two chromosomes that have a markedly different composition to the other chromosomes that make up the genome, and contain the majority of the transposable elements. The origin and the function of these chromosomes are still poorly understood. Interestingly, these two chromosomes exhibit a very low level of synteny between the two *Ostreococcus* genomes despite a generally high level of synteny when the other chromosomes are compared (Palenik et al. 2007). Based on this observation and the differences in structure compared to the rest of the genome, it has been suggested that one of these chromosomes, chromosome 2, may be related to speciation (Palenik et al. 2007).

O. tauri and *O. lucimarinus* provide an unusual example of cryptic species. Despite being indistinguishable morphologically when analysed using electron microscopy and having 99.8% identical 18S rDNA sequences, orthologous genes from the two species only share about 70% amino-acid identity (Palenik et al. 2007). At the genome level, therefore, they are clearly separate species and could even be classified as separate genera. A third *Ostreococcus* strain whose genome is currently being analysed exhibits the same phenomenon. Hence the diversity of phytoplankton algae may be considerably greater than that judged simply on the basis of cellular morphology.

The *O. tauri* and *O. lucimarinus* genome sequences have provided many insights into prasinophyte biology. Genes encoding all of the enzymes required for C4-photosynthesis are present in the *Ostreococcus* genomes indicating that, like the diatoms, these algae may use this pathway as a CO_2 concentration mechanism. This may, therefore, be a strategy that has been adapted by planktonic microalgae from diverse phylogenetic origins. Similarly, the two *Ostreococcus* genomes encode an unusually high number of selenoenzymes (Palenik et al. 2007), and this seems to be a general feature of marine microalgae, including diatoms (Lobanov et al. 2007). The increased catalytic activity of selenoproteins compared to equivalent enzymes that lack selenium might provide a selective advantage in the marine environments where these organisms are found. It may also be that, for some unknown reason, aquatic habitats favour the use of selenoproteins, compared to terrestrial habitats (Lobanov et al. 2007).

Two additional prasinophyte genomes, corresponding to two isolates of *Micromonas,* have recently been described (Worden et al. 2009). These isolates are morphologically identical and have been considered to be two members of the species *Micromonas pusilla* but genome sequencing has shown that only about 90% of the genes identified are present in both genomes. The *Micromonas* genomes are larger (20.9 and 21.9 Mbp) than those of the *Ostreococcus* species, and gene families are in general more extensive. One particularly interesting feature was the discovery in the one of the genomes of abundant intronic repeat sequences (introners) that extended nearly to the donor and acceptor sites of the introns.

O. tauri is also being used as a model organism in the laboratory, allowing experimental investigation of prasinophyte biology in much the same way as *P. tricornutum* is being used to study diatoms. *O. tauri* is attractive as a model organism for several reasons. One of the most important of these is the simplicity of the *O. tauri* genome, in particular the fact that most genes are unique and not part of redundant gene families. In this respect, *O. tauri* differs from more classical, land plant models such as *Arabidopsis* and, as a result, is attracting interest as a system to investigate cellular processes of general relevance to the green lineage. Two other advantageous features are the short intergenic regions, which allow promoter regions to be easily isolated and studied, and the fact that cell populations are haploid. This latter feature is a potential advantage for genetic approaches, although at present it is not possible to go through the sexual cycle in the laboratory. Several tools have been developed for *O. tauri* including clonal isolation of colonies on plates, genetic transformation with reporter gene constructs, gene expression monitoring under defined growth conditions and genome-scale microarray analysis. These tools are currently being applied to understanding the relationship between the mitotic cell cycle and the control of circadian rhythms in this alga (Moulager et al. 2007).

As with the diatoms, therefore, there is extensive genome sequence information available for the prasinophytes and a powerful model organism that with allow post-genomic analysis of gene function. Additional genome data will be available shortly for a broad range of prasinophytes (Table 6.1) that have been isolated from diverse environments (Rodríguez et al. 2005, Slapeta et al. 2006) and comparisons of these genomes will provide new hypotheses for future investigations both in the field and in the laboratory.

6.4.5.3 Other Microalgal Genome Projects

Coccolithophores are unicellular, haptophyte algae that tend to be found in nutrient-poor regions of the oceans. They can form large blooms that are seen as patches of turquoise due to the reflection of light from the calcium carbonate coccoliths that protect the outsides of the cells. These abundant organisms play an important role in biogeochemical cycles, particularly in trapping carbon via the sinking of cellular debris to ocean floor sediments following death. The genome of the coccolithophore *Emiliania huxleyi* has been sequenced (Table 6.1) and is currently being analysed with the aim of improving our understanding of many features of the biology of these ecologically important organisms. Genome sequencing projects are also planned for two species of haptophyte algae from the genus *Phaeocystis* (Table 6.1) and EST sequences are also available for *Prymnesium parvum* (La Claire 2006).

Several recently initiated genome projects target key organisms from the various groups that have been involved in primary or secondary plastid endosymbiosis events (Fig. 6.1), and a wealth of new information about these events will soon be available. The glaucophytes are not marine algae but the current genome project for one member of this group, *Cyanophora paradoxa* (http://www.biology.uiowa.edu/cyanophora/cyanophora_home.htm), will constitute an important part

of this initiative because of the single primary endosymbiosis in this group that gave rise to the plastids of all the algae in the eukaryotic tree. Cryptophytes (or cryptomonads) are unicellular algae that are found in both freshwater and marine environments. This group is of particular interest because their plastids, which are derived from a secondary endosymbosis, retain a remnant of the endosymbiont nucleus, the nucleomorph. A genome sequencing project is underway for one member of this group, *Guillardia theta,* an alga that is found in coastal regions (Table 6.1). Chlorarachniophytes are marine amoeboflagellates belonging to the supergroup Cercozoa, which, like the cryptophytes, possess a nucleomorph-bearing plastid derived from a secondary endosymbiosis. A genome sequencing project has also been initiated for a member of this group, *Bigelowiella natans* (Table 6.1). Considerable EST data is already available for *Bigelowiella natans* and, interestingly, analysis of this data has found evidence for numerous lateral gene transfers both from bacteria and from other eukaryotic lineages such as streptophytes, heterokonts and red algae (Archibald et al. 2003). As far as we are aware there is no genome project currently planned for a photosynthetic member of the Euglenozoa, although EST sequences are available for the freshwater alga *Euglena gracilis* (Durnford and Gray 2006).

6.4.5.4 Dinoflagellates

Although only about half of dinoflagellate species contain plastids and are thus capable of photosynthesis, collectively these species are important primary producers in the marine environment. Most dinoflagellate plastids are thought to have originated from a secondary endosymbiosis involving a red algal cell. Dinoflagellate plastids show several unusual features, such as the presence of the accessory pigment peridinin and the presence of three membranes surrounding the plastid (Nassoury et al. 2003, Patron et al. 2005). Moreover, the plastid genome is highly reduced, containing only a small number of genes, each on a separate minicircular chromosome (Zhang et al. 1999). All other genes necessary for plastid function have been transferred to the nucleus, including a nuclear-encoded proteobacterial Form II RuBisCO (Morse et al. 1995, Bachvaroff et al. 2004, Hackett et al. 2004). As mentioned above, some dinoflagellates have lost these red algal derived plastids and appear to have replaced them with plastids captured from other photosynthetic eukaryotes such as haptophytes or diatoms via tertiary endosymbioses (Inagaki et al. 2000, Bhattacharya et al. 2004). The dinoflagellate nucleus also shows some unusual features. It contains extremely large amounts of DNA (3–250 pg, or the equivalent of 3,000–215,000 Mbp per cell) organised into hundreds of chromosomes (for example there are 143 in *Alexandrium tamarense*; Hackett et al. 2005). The high concentration of DNA in the nucleus, which exceeds the concentration of basic DNA-binding proteins by about 10-fold, results in its being condensed in a liquid crystal state, and attachment to the nuclear envelope leads to the unusual nuclear morphology known as a “dinokaryon” (Spector 1984, Gautier et al. 1986, LaJeunesse et al. 2005, Hackett et al. 2005). The large size of dinoflagellate genomes has significantly hindered the application of genomic approaches, and no genome sequencing

projects exist for this group at the present time. However, a considerable amount of information can be obtained by sequencing cDNA clones and individual genes isolated from these genomes. ESTs data is currently available for several dinoflagellates, particularly species that produce toxic metabolites, such as *Alexandrium tamarense*, *A. catenella*, *Karenia brevis*, *Amphidinium carterae*, *Heterocapsa triquetra* and *Lingulodinium polyedrum* (Genbank Sequence Database, Tanikawa et al. 2004, Hackett et al. 2005, Bachvaroff and Place 2008).

6.4.6 Macroalgal Genomics

In contrast to pelagic plankton communities, which contain primarily unicellular algae, coastal ecosystems are often dominated by macroalgae of the brown, red and green lineages. These three types of seaweed play key roles in coastal ecosystems, not only as primary producers but also by providing habitats for a wide variety of marine life. Seaweeds are also important from an evolutionary perspective. Brown, red and green macroalgae represent diverse branches of the eukaryotic phylogenetic tree and each group has independently evolved complex multicellularity. This independent evolution of complex multicellularity in the three seaweed groups is particularly important as it has placed them in a very select club within the eukaryotes, shared with only three other lineages, the metazoans, land plants and fungi. Marine macroalgae therefore represent an important resource for understanding how complex developmental systems have evolved.

Seaweeds have also attracted interest for many other reasons. Fertilisation is external in brown algae, involving fusion of gametes that have been released into the surrounding seawater. This is a major advantage for the study of early developmental events compared, for example, to angiosperms where fertilisation occurs within several layers of maternal tissue (Brownlee et al. 2001). Macroalgae also exhibit a remarkably wide range of different life cycles, several of which are very complex, making them ideal models to study how variant life cycles evolve and why they are stable over evolutionary time (Coelho et al. 2007). Finally, seaweeds are also a major source of food and industrial biomolecules such as alginates, agars and carrageenans. Seaweeds represent an important resource with a wide range of uses in the food, cosmetic, and fertiliser industries and an estimated annual global value of about 4.5 billion Euros (McHugh 2003). They are also attracting increasing attention as a novel source of active biomolecules. One example of this is IODUS 40, a formulation derived from cell wall extracts of *Laminaria digitata* that stimulates the natural defence responses of crop plants (Klarzynski et al. 2000).

The following sections will look at how genomic approaches are being used to investigate the biology of macroalgae for each of the three phylogenetic groups.

6.4.6.1 Brown Macroalgae

The most evolved and morphologically complex members of the brown algae (or Phaeophyceae) are found in the orders Laminariales and Fucales. These include the

fucoid seaweeds of the intertidal zone and the large kelps that form forests in the subtidal zone. Historically, research on brown algae has been focused, to a large degree, on these two groups. This is not only because of their ecological importance but also either because they have important industrial applications (for the kelps; McHugh 2003, Bartsch et al. 2008) or because they have been developed as models for fundamental research (for the Fucales, see below). As far as genomic analysis of these two orders of seaweeds is concerned, a collection of nearly 3,000 ESTs has been generated for *Laminaria digitata*, including sequences from genes expressed during the sporophyte and gametophyte generations of the life cycle (Crépineau et al. 2000) and in sporophyte-derived protoplasts (Roeder et al. 2005). These sequences have allowed access to genes involved in a number of different processes including carbon-concentrating mechanisms, cell wall biosynthesis, halogen metabolism and stress responses. In contrast, EST sequences have only very recently been established for fucoid seaweeds (Gareth Pearson, CCMAR Faro, personal communication) despite the fact that this group has been used extensively to characterise events during early embryogenesis (Berger et al. 1994, Bouget et al. 1998, Corellou et al. 2000, Goddard et al. 2000, Corellou et al. 2001, Coelho et al. 2002) and are well characterised in terms of their ecology (Serrao et al. 1996, Coyer et al. 2007, Muhlin et al. 2008).

Although EST approaches are providing insights into the biology the Laminariales and Fucales groups, at the present time these organisms are not appropriate models for more extensive genomic approaches such as genome sequencing or gene function analysis. There are two main reasons for this. Firstly, both groups consist of organisms with large genomes (e.g. 650 Mbp for *Laminaria digitata* and 1095 Mbp for *Fucus serratus*; Le Gall et al. 1993, Peters et al. 2004). Secondly, they are large organisms with long life cycles, and this limits the scope for laboratory experimentation. In response to these limitations, a search was recently carried out for a model organism from within the brown algae that would be better adapted for genomic analysis. This search led to the selection of the filamentous brown alga *Ectocarpus siliculosus* (Peters et al. 2004).

Ectocarpus siliculosus is a member of the Ectocarpales, which has recently been shown to be among the most evolved of the brown algal orders, closely related to the kelps. *Ectocarpus* has been studied in the laboratory since the nineteenth century (see Charrier et al. 2008 and references therein). Early research included the description of the species followed by studies of its reproductive biology and life history. Other aspects that have been investigated include ultrastructure, photosynthesis and carbon uptake, pheromone production, gamete recognition and interactions with pathogens, in particular with the virus EsV-1 which integrates into the genome of this alga following infection.

The choice to develop *Ectocarpus* as a model organism for the brown algae was based on a number of features that make it well adapted for the application of both genomic and genetic approaches. One particularly important factor was the size of its genome, which at 200 Mbp is significantly smaller than those of kelps and fucoid brown algae. It also has several features that make it well adapted to laboratory work, including its small size, the fact that the life cycle can be completed in Petri dishes

under laboratory conditions, its high fertility and rapid growth (the life cycle can be completed in 2 months) and the ease with which genetic crosses can be carried out (Peters et al. 2004). The aim with this organism, therefore, has been to develop it as a model system that will allow analyses to go beyond a simple inventory of the genes present in the genome. *Ectocarpus* was selected because it presents the possibility of using genetic approaches to analyse gene function.

Sequencing of the *Ectocarpus* genome was completed in 2007 and a number of additional genomic tools are now available including whole genome tiling array data and 91,000 cDNA sequences corresponding to several stages of the life cycle. In terms of laboratory techniques, classical genetic approaches such as mutagenesis, screening for mutant lines, crosses and complementation analysis can be carried out routinely. Several cell biology tools including in vivo and in vitro imaging, microinjection and protoplast regeneration are also available. Additional tools currently under development include a genetic map, a genetic transformation protocol, RNAi-based gene knockdowns and positional cloning of mutated loci.

With the above list of tools in hand, *Ectocarpus* is now being used as a model to investigate a broad range of topics related to brown algal biology. Recent work aimed at understanding how the life cycle is regulated provides a good example (Coelho et al. 2007, Peters et al. 2008). *Ectocarpus* has a haploid-diploid life cycle involving alternation between sporophyte and gametophyte generations (Müller 1967). In an effort to understand how the switch between the two generations is controlled, screens have been carried out for mutants in which this process is perturbed. In one mutant, *immediate upright* (*imm*), the sporophyte is partially converted into a gametophyte, exhibiting a pattern of early development that closely resembles that of the gametophyte but, nonetheless, producing spores rather than gametes. Analysis of gene expression in this mutant using microarray analysis and quantitative PCR, showed that a large number of genes that are normally expressed during the gametophyte generation are expressed in the sporophyte of this mutant. Future work involving the analysis of additional mutant lines, genome-scale analyses of changes in gene regulation in these mutants and identification of mutated genes by positional cloning is expected to provide further insights into the regulatory mechanisms that control the life cycle in the coming years.

Similar approaches have been initiated in an effort to understand other aspects of *Ectocarpus* biology including, for example, sex determination, morphogenesis, cell wall biosynthesis and responses to biotic and abiotic stresses. Much of the information obtained will be of general relevance to the brown algae as a group, with potential applications including the identification of novel biomolecules and the exploitation of genetic data in future seaweed breeding programs.

6.4.6.2 Red Macroalgae

Red macroalgae are found in a wide range of shoreline habitats ranging from the extreme high shore to the lower limit of the photic zone. Several species are exploited for food and industrial applications. About 2.8 million tonnes of red algae are harvested annually, with a value of approximately 2,000 million US dollars (FAO

2003). The species harvested are principally from the genera *Porphyra*, for consumption as nori, and *Eucheuma* and *Kappaphycus,* for carrageenan production. The red algae are an ancient eukaryotic group; estimated to have originated about 1,500 Mya (Yoon et al. 2004). Fossils of red algae exhibiting evidence of multicellular development and sexuality have been dated at approximately 1,200 Mya, indicating that this group was the first to evolve complex multicellularity (Butterfield 2000). As with the brown algae, this ancient evolutionary history is correlated with many unusual features such as their complex life cycles and novel metabolic processes; the latter particularly in terms of the production of oxilipins, cell polysaccharides and halogenated compounds (Siegel and Siegel 1973, Manley 2002, Bouarab et al. 1999, Coelho et al. 2007). As discussed above, red algae also played a key role in the evolutionary history of photosynthesis, notably via the endosymbiotic processes that led to the evolution of secondary chloroplasts.

Taken together these features provide strong arguments for using genomic approaches to explore red algal genomes. EST sequences are available for several species of red macroalga, including *Porphyra yezoensis* (20,000 ESTs, Nikaido et al. 2000, Asamizu et al. 2003), *Chondrus crispus* (4,056 ESTs, Collén et al. 2006b), *Gracilaria tenuistipitata* (3,000 ESTs, Pi Nyvall, personnal communication) and *Griffithsia okiensis* (1,104 ESTs, Lee et al. 2007), and cDNA micro- or macroarrays have been used to monitor gene expression in both *Chondrus crispus* (Collén et al. 2006a) and *Porphyra yezoensis* (Kitade et al. 2008). However, currently the only complete red algal genome sequence available is that of *C. merolae,* a unicellular organism from hot acid springs that has a highly reduced and unusual genome. A second project is in progress, for *Galdieria sulphuraria,* (http://genomics.msu.edu/galdieria/about.html) but this is also a non-marine, unicellular, extremophile red alga. Both of these organisms have very small, highly derived genomes, which are of limited relevance to understanding many red algal features. A genome sequence for a more "typical" red alga would be extremely useful both to further our understanding of red algal biology and as a reference for tracing the origins of genes acquired via endosymbiotic events. In response to this need, genome projects have recently been initiated for two red macroalgae, the florideophyte *Chondrus crispus* (at Genoscope, France) and the bangiophyte *Porphyra umbilicalis* (at the Joint Genome Institute, USA).

These two genome projects are likely to be highly compatible. Both seaweeds are ecologically important in specific habitats. They can both be handled relatively easily in the laboratory and each has been the subject of extensive laboratory studies. *Porphyra* spp. are perhaps the best adapted to laboratory work, previous studies have reported the isolation of mutant strains (Ohme and Miura 1988, Mitman and van der Meer 1994, Yan et al. 2000), identification of genetic markers (Park et al. 2007), the preparation and regeneration of protoplasts (Waaland et al. 1990), whole mount in situ hybridisation (Shimizu et al. 2004) and progress towards the development of genetic transformation (Cheney et al. 2001, He et al. 2001, Lin et al. 2001). Indeed, a related species, *Porphyra yezoensis,* has been proposed as a candidate model macroalgae (Kitade et al. 2004, Waaland et al. 2004). *C. crispus* has the advantage of possessing a smaller genome (150 Mbp, Peters et al. 2004) than

P. umbilicalis (genomes of *Porphyra* spp. tend to be around 400 Mbp, Kapraun 2005) and of being more representative of the "typical" red algae (the majority of red algae are florideophytes). Both the *C. crispus* and the *P. umbilicalis* genomes will be of interest for applied research, *C. crispus* for understanding carrageenan biosynthesis and *P. umbilicalis* because of the importance of *Porphyra* spp. (nori) in the multibillion dollar nori industry. It is also important to note that the common ancestor of *C. crispus* and *P. umbilicalis* is though to date back to about 1,400 Mya, so these two species represent very different evolutionary groups (Yoon et al. 2004).

6.4.6.3 Green Macroalgae

Genomic approaches have not been applied to the same extent to green marine macroalgae as they have for the reds and browns. This is probably because of the less obvious potential for industrial applications. However, green macroalgae such as *Ulva* are important for their roles in fouling and because of their tendency to bloom under high nutrient conditions, for example when pollution by fertilisers is combined with the warm temperatures and high light fluxes during the summer months in temperate regions. Among the green marine macroalgae, *Ulva* has been studied most extensively and a collection of ESTs has been established (Bryhni 1974, Fjeld and Løvle 1976, Reddy et al. 1992). At present, however, there are no plans to sequence the genome of a member of this group.

Sea-grasses are angiosperms and, therefore, are not defined as algae. It is important to mention these organisms here though because, like the macroalgae, they are ecologically import photosynthetic organisms in many coastal regions. Sea-grasses can form extensive "meadows" on sandy or muddy shorelines, providing a habitat for many other organisms. These important ecosystems are threatened by human activity in many parts of the world. A genome sequencing project has been initiated for one species of sea-grass, the eelgrass *Zostera marina* (Table 6.1).

6.5 Future Research in Algal Genomics

Genomic approaches are currently being used to address a very broad range of questions related to algal biology. These include questions addressing fundamental aspects of the evolution of the eukaryotes, questions concerning the key roles played by algae in various open-ocean and coastal ecosystems and in-depth studies of specific aspects of algal biology, the latter often with novel biotechnological applications. Future research will extend our knowledge in all of these areas. Sequence data is still very fragmentary and many of the major eukaryotic groups are still very poorly explored at the molecular level. Future work will fill in many of these gaps, adding robustness to the eukaryote phylogeny. However, it is likely that major eukaryotic groups still remain to be discovered (Not et al. 2007) and, consequently, that currently recognised supergroups of the eukaryotic tree may need to be reorganised. New information is also expected about the various endosymbiotic events that played such an important role during evolution.

Genomic approaches have only just started to be applied to marine algal ecosystems. The handful of genome sequences that are available for planktonic microalgae, for example, have provided important insights into how these organisms function in their environment but these studies have only scratched the surface of the complexity of these ecosystems. Future studies will combine environmental sequencing data with genome information for individual strains to provide a better understanding both of the biodiversity present and of the characteristics of component species. These approaches will be combined with methods that allow the analysis of gene expression such as cDNA sequencing or microarray analysis providing, not only a census of the genes present in an ecosystem but also an overview of which genes are active under particular conditions. Similar studies are planned for macroalgae in coastal environments. The larger size of macroalgal genomes limits the number of genome sequencing projects at present but the rapid progress that is being made with these technologies is expected to modify this situation in the coming years. Together these future studies are expected to provide more detailed descriptions of the roles of algae in marine biosytems, particularly in terms of their key influence on biogeochemical cycles and their responses to climate change and other anthropogenic impacts.

One of the consequences of the broad phylogenetic distribution of the algae across several very ancient eukaryotic groups is that they represent an enormous potential for the discovery of novel biological processes including metabolic pathways, signalling pathways, cellular processes, developmental regulation, etc. One of the surprising results of algal sequencing projects has often been the very high proportion of genes that have no matches in the public databases. For example, this was the case for about half the genes in the genome of the diatom *T. pseudonana* (Armbrust et al. 2004). A major challenge for the future will be to understand the cellular functions of these "orphan" genes and the development of a number of model organisms for both the micro and the macroalgae is an important step towards this objective.

References

Allen AE, Vardi A, Bowler C (2006) An ecological and evolutionary context for integrated nitrogen metabolism and related signaling pathways in marine diatoms. Curr Opin Plant Biol 9: 264–273

Archibald JM (2007) Nucleomorph genomes: structure, function, origin and evolution. Bioessays 29:392–402

Archibald JM, Rogers MB, Toop M, Ishida K, Keeling PJ (2003) Lateral gene transfer and the evolution of plastid-targeted proteins in the secondary plastid-containing alga *Bigelowiella natans*. Proc Natl Acad Sci U S A 100:7678–7683

Armbrust EV, Berges JA, Bowler C, Green BR, Martinez D, Putnam NH, Zhou S, Allen AE, Apt KE, Bechner M, Brzezinski MA, Chaal BK, Chiovitti A, Davis AK, Demarest MS, Detter JC, Glavina T, Goodstein D, Hadi MZ, Hellsten U, Hildebrand M, Jenkins BD, Jurka J, Kapitonov VV, Kröger N, Lau WW, Lane TW, Larimer FW, Lippmeier JC, Lucas S, Medina M, Montsant A, Obornik M, Parker MS, Palenik B, Pazour GJ, Richardson PM, Rynearson TA, Saito MA, Schwartz DC, Thamatrakoln K, Valentin K, Vardi A, Wilkerson FP, Rokhsar DS (2004) The

genome of the diatom *Thalassiosira pseudonana*: Ecology, evolution, and metabolism. Science 306:79–86

Asamizu E, Nakajima M, Kitade Y, Saga N, Nakamura Y, Tabata S (2003) Comparison of RNA expression profiles between two generations of *Porphyra yezoensis* (Rhodophyta), based on expressed sequence tag frequency analysis. J Phycol 39:923–330

Bachvaroff TR, Concepcion GT, Rogers CR, Herman EM, Delwiche CF (2004) Dinoflagellate expressed sequence tags data indicate massive transfer of chloroplast genes to the nuclear genome. Protist 155:65–78

Bachvaroff TR, Place AR (2008) From stop to start: tandem gene arrangement, copy number and trans-splicing sites in the dinoflagellate *Amphidinium carterae*. PLoS ONE 3:e2929

Baldauf SL (2008) An overview of the phylogeny and diversity of eukaryotes. J Syst Evol 46: 263–273

Bartsch I, Wiencke C, Bischof K, Buchholz CM, Buck BH, Eggert A, Feuerpfeil P, Hanelt D, Jacobsen S, Karez R, Karsten U, Molis M, Roleda M, Schubert H, Schumann R, Valentin K, Weinberger F, Wiese J (2008) The genus *Laminaria* sensu lato: recent insights and developments. Eur J Phycol 43:1–86

Bauer I, Maranda L,Young KA, Shimizu Y, Fairchild C, Cornell L, MacBeth J, Huang S (1995) Isolation and structure of caribenolide I, a highly potent antitumor macrolide from a culture free-swimming Caribbean dinoflagellate, *Amphidinium* sp. S1-36-5. J Org Chem 60: 1084–1086

Berger F, Taylor A, Brownlee C (1994) Cell fate determination by the cell wall in early *Fucus* development. Science 263:1421–1423

Bhattacharya D, Yoon HS, Hackett JD (2004) Photosynthetic eukaryotes unite: endosymbiosis connects the dots. Bioessays 26:50–60

Bouarab K, Potin P, Correa J, Kloareg B (1999) Sulfated oligosaccharides mediate the interaction between a marine red alga and its green algal pathogenic endophyte. Plant Cell 11(9): 1635–1650

Bouget FY, Berger F, Brownlee C (1998) Position dependent control of cell fate in the *Fucus* embryo: role of intercellular communication. Development 125:1999–2008

Bowler C, Allen AE, Badger JH, Grimwood J, Jabbari K, Kuo A, Maheswari U, Martens C, Maumus F, Otillar RP, Rayko E, Salamov A, Vandepoele K, Beszteri B, Gruber A, Heijde M, Katinka M, Mock T, Valentin K, Verret F, Berges JA, Brownlee C, Cadoret JP, Chiovitti A, Choi CJ, Coesel S, De Martino A, Detter JC, Durkin C, Falciatore A, Fournet J, Haruta M, Huysman MJ, Jenkins BD, Jiroutova K, Jorgensen RE, Joubert Y, Kaplan A, Kröger N, Kroth PG, La Roche J, Lindquist E, Lommer M, Martin-Jézéquel V, Lopez PJ, Lucas S, Mangogna M, McGinnis K, Medlin LK, Montsant A, Secq MP, Napoli C, Obornik M, Parker MS, Petit JL, Porcel BM, Poulsen N, Robison M, Rychlewski L, Rynearson TA, Schmutz J, Shapiro H, Siaut M, Stanley M, Sussman MR, Taylor AR, Vardi A, von Dassow P, Vyverman W, Willis A, Wyrwicz LS, Rokhsar DS, Weissenbach J, Armbrust EV, Green BR, Van de Peer Y, Grigoriev IV (2008) The Phaeodactylum genome reveals the evolutionary history of diatom genomes. Nature 456:239–244

Brownlee C, Bouget FY, Corellou F (2001) Choosing sides: establishment of polarity in zygotes of fucoid algae. Semin Cell Dev Biol 12:345–351

Bryhni E (1974) Control of morphogenesis in the multicellular alga *Ulva mutabilis.* Defect in cell wall production. Dev Biol 37:273–277

Butterfield NJ (2000) *Bangiomorpha pubescens* n. gen., n. sp.; implications for the evolution of sex, multicellularity, and the Mesoproterozoic/Neoproterozoic radiation of eukaryotes. Paleobiol 26:386–404

Charrier B, Coelho SM, Le Bail A, Tonon T, Michel G, Potin P, Kloareg B, Boyen C, Peters AF, Cock JM (2008) Development and physiology of the brown alga *Ectocarpus siliculosus*: two centuries of research. New Phytol 177:319–332

Cheney D, Metz B, Stiller J (2001) *Agrobacterium*-mediated genetic transformation in the macroscopic marine red alga *Porphyra yezoensis*. J Phycol 37:11

Coelho S, Peters AF, Charrier B, Roze D, Destombe C, Valero M, Cock JM (2007) Complex life cycles of multicellular eukaryotes: new approaches based on the use of model organisms. Gene 406:152–170

Coelho SM, Taylor AR, Ryan KP, Sousa-Pinto I, Brown MT, Brownlee C (2002) Spatiotemporal patterning of reactive oxygen production and Ca(2+) wave propagation in *Fucus* rhizoid cells. Plant Cell 14:2369–2381

Coesel S, Oborník M, Varela J, Falciatore A, Bowler C (2008) Evolutionary origins and functions of the carotenoid biosynthetic pathway in marine diatoms. PLoS ONE 3:e2896

Coleman FC, Williams SL (2002) Overexploiting marine ecosystem engineers: potential consequences for biodiversity. Trends Ecol Evol 17:40–44

Collén J, Hervé C, Guisle-Marsollier I, Leger J, Boyen C (2006a) Expression profiling of *Chondrus crispus* (Rhodophyceae) after exposure to methyl jasmonate. J Exp Bot 57:3869–3881

Collén J, Roeder V, Rousvoal S, Collin O, Kloareg B, Boyen C (2006b) An expressed sequence tag analysis of thallus and regenerating protoplasts of *Chondrus crispus* (Gigartinales, Rhodophyceae). J Phycol 42:104–112

Corellou F, Bisgrove SR, Kropf DL, Meijer L, Kloareg B, Bouget FY (2000) A S/M DNA replication checkpoint prevents nuclear and cytoplasmic events of cell division including centrosomal axis alignment and inhibits activation of cyclin-dependent kinase-like proteins in fucoid zygotes. Development 127:1651–1660

Corellou F, Brownlee C, Kloareg B, Bouget FY (2001) Cell cycle-dependent control of polarised development by a cyclin-dependent kinase-like protein in the *Fucus* zygote. Development 128:4383–4392

Coyer JA, Hoarau G, Stam WT, Olsen JL (2007) Hybridization and introgression in a mixed population of the intertidal seaweeds *Fucus evanescens* and *F. serratus*. J Evol Biol 20:2322–2333

Crépineau F, Roscoe T, Kaas R, Kloareg B, Boyen C (2000) Characterisation of complementary DNAs from the expressed sequence tag analysis of life cycle stages of *Laminaria digitata* (Phaeophyceae). Plant Mol Biol 43:503–513

Derelle E, Ferraz C, Escande ML, Eychenié S, Cooke R, Piganeau G, Desdevises Y, Bellec L, Moreau H, Grimsley N (2008) Life-cycle and genome of OtV5, a large DNA virus of the pelagic marine unicellular green alga *Ostreococcus tauri*. PLoS ONE 3:e2250

Derelle E, Ferraz C, Rombauts S, Rouze P, Worden AZ, Robbens S, Partensky F, Degroeve S, Echeynie S, Cooke R, Saeys Y, Wuyts J, Jabbari K, Bowler C, Panaud O, Piegu B, Ball SG, Ral JP, Bouget FY, Piganeau G, De Baets B, Picard A, Delseny M, Demaille J, Van de Peer Y, Moreau H (2006) Genome analysis of the smallest free-living eukaryote *Ostreococcus tauri* unveils many unique features. Proc Natl Acad Sci U S A 103:11647–11652

Duggins DO, Simenstad CA, Estes JA (1989) Magnification of secondary production by kelp detritus in coastal marine ecosystems. Science 245:170–173

Durnford DG, Gray MW (2006) Analysis of Euglena gracilis plastid-targeted proteins reveals different classes of transit sequences. Eukaryot Cell 5:2079–2091

FAO (2003) A guide to the seaweed industry. FAO Fisheries technical paper 441

Falkowski PG, Katz ME, Knoll AH, Quigg A, Raven JA, Schofield O, Taylor FJ (2004) The evolution of modern eukaryotic phytoplankton. Science 305:354–360

Faust MA, Gulledge RA (2002) Identifying harmful marine dinoflagellates. Contributions from the United States National Herbarium 42:1–144

Field CB, Behrenfeld MJ, Randerson JT, Falkowski P (1998) Primary production of the biosphere: integrating terrestrial and oceanic components. Science 281:237–240

Fjeld A, Løvle A (1976) Genetics of multicellular algae. In: Lewin, RA (ed) The genetics of algae, Blackwell Scientific Publications, Oxford, pp 219–235

Gautier A, Michel-Salamin L, Tosi-Couture E, McDowall AW, Dubochet J (1986) Electron microscopy of the chromosomes of dinoflagellates in situ: confirmation of Bouligand's liquid crystal hypothesis. J Ultrastruc Mol Struct Res 97:10–30

Giovannoni SJ, Britschgi TB, Moyer CL, Field KG (1990) Genetic diversity in Sargasso Sea bacterioplankton. Nature 345:60–63

Goddard H, Manison NF, Tomos D, Brownlee C (2000) Elemental propagation of calcium signals in response-specific patterns determined by environmental stimulus strength. Proc Natl Acad Sci USA 97:1932–1937

Guillou L, Eikrem W, Chrétiennot-Dinet MJ, Le Gall F, Massana R, Romari K, Pedrós-Alió C, Vaulot D (2004) Diversity of picoplanktonic prasinophytes assessed by direct nuclear SSU rDNA sequencing of environmental samples and novel isolates retrieved from oceanic and coastal marine ecosystems. Protist 155:193–214

Guillou L, Moon-Van Der Staay SY, Claustre H, Partensky F, Vaulot D (1999) Diversity and abundance of Bolidophyceae (Heterokonta) in two oceanic regions. Appl Environ Microbiol 65:4528–4536

Guiry MD, Guiry GM (2008) AlgaeBase. Worldwide electronic publication, National University of Ireland, Galway. http://www.algaebase.org.

Hackett JD, Scheetz TE, Yoon HS, Soares MB, Bonaldo MF, Casavant TL, Bhattacharya D (2005) Insights into a dinoflagellate genome through expressed sequence tag analysis. BMC Genomics 6:80

Hackett JD, Yoon HS, Soares MB, Bonaldo MF, Casavant TL, Scheetz TE, Nosenko T, Bhattacharya D (2004) Migration of the plastid genome to the nucleus in a peridinin dinoflagellate. Curr Biol 14:213–218

He P, Yao Q, Chen Q, Guo M, Xiong A, Wu W, Ma J (2001) Transferring and expression of glucose oxidase gene – gluc in *Porphyra yezoensis*. J Phycol 37(suppl):22

Hutchinson GE (1961) The paradox of the plankton. Amer Nat 95:137–145

Ianora A, Miralto A, Poulet SA, Carotenuto Y, Buttino I, Romano G, Casotti R, Pohnert G, Wichard T, Colucci-D'Amato L, Terrazzano G, Smetacek V (2004) Aldehyde suppression of copepod recruitment in blooms of a ubiquitous planktonic diatom. Nature 429:403–407

Inagaki Y, Dacks JB, Doolittle WF, Watanabe KI, Ohama T (2000) Evolutionary relationship between dinoflagellates bearing obligate diatom endosymbionts: insight into tertiary endosymbiosis. Int J Syst Evol Microbiol 50:2075–2081

Kapraun DF (2005) Nuclear DNA content estimates in multicellular green, red and brown algae: phylogenetic considerations. Ann Bot 95:7–44

Keeling PJ (2004) Diversity and evolutionary history of plastids and their hosts. Amer J Bot 91:1481–1493

Keeling PJ (2007) *Ostreococcus tauri*: seeing through the genes to the genome. Trends Genet 23:151–154

Keeling PJ, Palmer JD (2008) Horizontal gene transfer in eukaryotic evolution. Nat Rev Genet 9:605–628

Kilian O, Kroth PG (2005) Identification and characterization of a new conserved motif within the presequence of proteins targeted into complex diatom plastids. Plant J 41:175–183

Kitade Y, Asamizu E, Fukuda S, Nakajima M, Ootsuka S, Endo H, Tabata S, Saga N (2008) Identification of genes preferentially expressed during asexual sporulation in Porphyra yezoensis gametophytes (bangiales, rhodophyta). J Phycol 44:113–123

Kitade Y, Iitsuka O, Fukuda S, Saga N (2004) *Porphyra yezoensis* as a model plant for genome sciences. Jpn J Phycol 52:129–131

Klarzynski O, Plesse B, Joubert J-M, Yvin J-C, Kopp M, Kloareg B, Fritig B (2000) Linear b-1,3 glucans are elicitors of defense responses in tobacco. Plant Physiol 124:1027–1037

Kroth PG, Chiovitti A, Gruber A, Martin-Jezequel V, Mock T, Parker MS, Stanley MS, Kaplan A, Caron L, Weber T, Maheswari U, Armbrust EV, Bowler C (2008) A model for carbohydrate metabolism in the diatom *Phaeodactylum tricornutum* deduced from comparative whole genome analysis. PLoS ONE 3:e1426

La Claire JW 2nd. (2006) Analysis of expressed sequence tags from the harmful alga, *Prymnesium parvum* (Prymnesiophyceae, Haptophyta). Mar Biotechnol 8:534–546

LaJeunesse TC, Lambert G, Andersen RA, Coffroth MA, Galbraith DW (2005) *Symbiodinium* (Pyrrhophyta) genome sizes (DNA content) are smallest among dinoflagellates. J Phycol 41:880–886

Lang M, Apt KE, Kroth PG (1998) Protein transport into "complex" diatom plastids utilizes two different targeting signals. J Biol Chem 273:30973–30978

Le Gall Y, Brown S, Marie D, Mejjad M, Kloareg B (1993) Quantification of nuclear DNA and G-C content in marine macroalgae by flow cytometry of isolated nuclei. Protoplasma 173:123–132

Lee H, Lee HK, An G, Lee YK (2007) Analysis of expressed sequence tags from the red alga *Griffithsia okiensis*. J Microbiol 45:541–546

Lin CM, Larsen J, Yarish C, Chen T (2001) A novel gene transfer in *Porphyra*. J Phycol 37:31

Lobanov AV, Fomenko DE, Zhang Y, Sengupta A, Hatfield DL, Gladyshev VN (2007) Evolutionary dynamics of eukaryotic selenoproteomes: large selenoproteomes may associate with aquatic life and small with terrestrial life. Genome Biol 8(9):R198

Lovejoy C, Massana R, Pedrós-Alió C (2006) Diversity and distribution of marine microbial eukaryotes in the Arctic Ocean and adjacent seas. Appl Environ Microbiol 72:3085–3095

Manley SL (2002) Phytogenesis of halomethanes: A product of selection or a metabolic accident?. Biogeochem 60:163–180

Massana R, Balague V, Guillou L, Pedros-Alio C (2004) Picoeukaryotic diversity in an oligotrophic coastal site studied by molecular and culturing approaches. FEMS Microbiol Ecol 50:231–243

Massana R, Guillou L, Díez B, Pedrós-Alió C (2002) Unveiling the organisms behind novel eukaryotic ribosomal DNA sequences from the ocean. Appl Environ Microbiol 68:4554–4558

Matsuzaki M, Misumi O, Shin-I T, Maruyama S, Takahara M, Miyagishima SY, Mori T, Nishida K, Yagisawa F, Nishida K, Yoshida Y, Nishimura Y, Nakao S, Kobayashi T, Momoyama Y, Higashiyama T, Minoda A, Sano M, Nomoto H, Oishi K, Hayashi H, Ohta F, Nishizaka S, Haga S, Miura S, Morishita T, Kabeya Y, Terasawa K, Suzuki Y, Ishii Y, Asakawa S, Takano H, Ohta N, Kuroiwa H, Tanaka K, Shimizu N, Sugano S, Sato N, Nozaki H, Ogasawara N, Kohara Y, Kuroiwa T (2004) Genome sequence of the ultrasmall unicellular red alga *Cyanidioschyzon merolae* 10D. Nature 428:653–657

McGinn PJ, Morel FM (2008) Expression and inhibition of the carboxylating and decarboxylating enzymes in the photosynthetic C4 pathway of marine diatoms. Plant Physiol 146:300–309

McHugh DJ (2003) A guide to the seaweed industry. FAO Fisheries Technical Paper No. 441. FAO, Rome, 105 pp.

Medlin LK, Metfies K, Mehl H, Wiltshire K, Valentin K (2006) Picoeukaryotic plankton diversity at the Helgoland time series site as assessed by three molecular methods. Microb Ecol 52:53–71

Merchant SS, Prochnik SE, Vallon O, Harris EH, Karpowicz SJ, Witman GB, Terry A, Salamov A, Fritz-Laylin LK, Maréchal-Drouard L, Marshall WF, Qu LH, Nelson DR, Sanderfoot AA, Spalding MH, Kapitonov VV, Ren Q, Ferris P, Lindquist E, Shapiro H, Lucas SM, Grimwood J, Schmutz J, Cardol P, Cerutti H, Chanfreau G, Chen CL, Cognat V, Croft MT, Dent R, Dutcher S, Fernández E, Fukuzawa H, González-Ballester D, González-Halphen D, Hallmann A, Hanikenne M, Hippler M, Inwood W, Jabbari K, Kalanon M, Kuras R, Lefebvre PA, Lemaire SD, Lobanov AV, Lohr M, Manuell A, Meier I, Mets L, Mittag M, Mittelmeier T, Moroney JV, Moseley J, Napoli C, Nedelcu AM, Niyogi K, Novoselov SV, Paulsen IT, Pazour G, Purton S, Ral JP, Riaño-Pachón DM, Riekhof W, Rymarquis L, Schroda M, Stern D, Umen J, Willows R, Wilson N, Zimmer SL, Allmer J, Balk J, Bisova K, Chen CJ, Elias M, Gendler K, Hauser C, Lamb MR, Ledford H, Long JC, Minagawa J, Page MD, Pan J, Pootakham W, Roje S, Rose A, Stahlberg E, Terauchi AM, Yang P, Ball S, Bowler C, Dieckmann CL, Gladyshev VN, Green P, Jorgensen R, Mayfield S, Mueller-Roeber B, Rajamani S, Sayre RT, Brokstein P, Dubchak I, Goodstein D, Hornick L, Huang YW, Jhaveri J, Luo Y, Martínez D, Ngau WC, Otillar B, Poliakov A, Porter A, Szajkowski L, Werner G, Zhou K, Grigoriev IV, Rokhsar DS, Grossman AR (2007) The Chlamydomonas genome reveals the evolution of key animal and plant functions. Science 318(5848):245–250

Metfies K, Medlin LK (2008) Feasibility of transferring fluorescent in situ hybridization probes to an 18S rRNA gene phylochip and mapping of signal intensities. Appl Environ Microbiol 74:2814–2821

Mitman GG, van der Meer JP (1994) Meiosis, blade development, and sex determination in *Porphyra purpurea* (Rhodophyta). J Phycol 30:147–159

Mock T, Samanta MP, Iverson V, Berthiaume C, Robison M, Holtermann K, Durkin C, Bondurant SS, Richmond K, Rodesch M, Kallas T, Huttlin EL, Cerrina F, Sussman MR, Armbrust EV (2008) Whole-genome expression profiling of the marine diatom Thalassiosira pseudonana identifies genes involved in silicon bioprocesses. Proc Natl Acad Sci U S A 105(5):1579–1584

Mock T, Thomas DN (2005) Recent advances in sea-ice microbiology. Environ Microbiol 7: 605–619

Moreira D, López-García P (2002) The molecular ecology of microbial eukaryotes unveils a hidden world. Trends Microbiol 10:31–38

Morse D, Salois P, Markovic P, Hastings WJ (1995) A Nuclear-Encoded Form II RuBisCO in Dinoflagellates. Science 268:1622–1624

Moulager M, Monnier A, Jesson B, Bouvet R, Mosser J, Schwartz C, Garnier L, Corellou F, Bouget FY (2007) Light-dependent regulation of cell division in *Ostreococcus*: evidence for a major transcriptional input. Plant Physiol 144:1360–1369

Muhlin JF, Engel CR, Stessel R, Weatherbee RA, Brawley SH (2008) The influence of coastal topography, circulation patterns, and rafting in structuring populations of an intertidal alga. Mol Ecol 17:1198–1210

Müller DG (1967) Generationswechsel, Kernphasenwechsel und Sexualität der Braunalge *Ectocarpus siliculosus* im Kulturversuch. Planta 141:39–54

Nagasaki K (2008) Dinoflagellates, diatoms, and their viruses. J Microbiol 46:235–243

Nassoury N, Cappadocia M, Morse D (2003) Plastid ultrastructure defines the protein import pathway in dinoflagellates. J Cell Sci 116:2867–2874

Nikaido I, Asamizu E, Nakajima M, Nakamura Y, Saga N, Tabata S (2000) Generation of 10,154 expressed sequence tags from a leafy gametophyte of a marine red alga, *Porphyra yezoensis*. DNA Res 7:223–227

Nosenko T, Bhattacharya D (2007) Horizontal gene transfer in chromalveolates. BMC Evol Biol 7:173

Not F, Valentin K, Romari K, Lovejoy C, Massana R, Töbe K, Vaulot D, Medlin LK (2007) Picobiliphytes: a marine picoplanktonic algal group with unknown affinities to other eukaryotes. Science 315:253–255

Nowack EC, Melkonian M, Glöckner G (2008) Chromatophore genome sequence of *Paulinella* sheds light on acquisition of photosynthesis by eukaryotes. Curr Biol 18:410–418

Ohme M, Miura A (1988) Tetrad analysis in conchospore germlings of *Porphyra yezoensis* (Rhodophyta, Bangiales). Plant Sci 57:135–140

Palenik B, Grimwood J, Aerts A, Rouzé P, Salamov A, Putnam N, Dupont C, Jorgensen R, Derelle E, Rombauts S, Zhou K, Otillar R, Merchant SS, Podell S, Gaasterland T, Napoli C, Gendler K, Manuell A, Tai V, Vallon O, Piganeau G, Jancek S, Heijde M, Jabbari K, Bowler C, Lohr M, Robbens S, Werner G, Dubchak I, Pazour GJ, Ren Q, Paulsen I, Delwiche C, Schmutz J, Rokhsar D, Van de Peer Y, Moreau H, Grigoriev IV (2007) The tiny eukaryote *Ostreococcus* provides genomic insights into the paradox of plankton speciation. Proc Natl Acad Sci USA 104:7705–7710

Parfrey LW, Barbero E, Lasser E, Dunthorn M, Bhattacharya D, Patterson DJ, Katz LA (2006) Evaluating support for the current classification of eukaryotic diversity. PLoS Genet 2:e220

Park E-J, Fukuda S, Endo H, Kitade Y,Saga N (2007) Genetic polymorphism within *Porphyra yezoe*nsis (Bangiales, Rhodophyta) and related species from Japan and Korea detected by cleaved amplified polymorphic sequence analysis. Eur J Phycol 42:29–40

Patron NJ, Waller RF, Archibald JM, Keeling PJ (2005) Complex protein targeting to dinoflagellate plastids. J Mol Biol 348(4):1015–1024

Peters AF, Marie D, Scornet D, Kloareg B, Cock JM (2004) Proposal of *Ectocarpus siliculosus* as a model organism for brown algal genetics and genomics. J Phycol 40:1079–1088

Peters AF, Scornet D, Ratin M, Charrier B, Monnier A, Merrien Y, Corre E, Coelho SM, Cock JM (2008) Life-cycle-generation-specific developmental processes are modified in the immediate upright mutant of the brown alga *Ectocarpus siliculosus*. Development 135: 1503–1512

Piganeau G, Desdevises Y, Derelle E, Moreau H (2008) Picoeukaryotic sequences in the Sargasso Sea metagenome. Genome Biol 9:R5

Piganeau G, Moreau H (2007) Screening the Sargasso Sea metagenome for data to investigate genome evolution in *Ostreococcus* (Prasinophyceae, Chlorophyta). Gene 406:184–190

Reddy CRK, Iima M, Fujita Y (1992) Induction of fastgrowing and morphologically different strains through intergeneric protoplast fusions of *Ulva* and *Enteromorpha* (Ulvales, Chlorophyta). J Appl Phycol 4:57–65

Reyes-Prieto A, Moustafa A, Bhattacharya D (2008) Multiple genes of apparent algal origin suggest ciliates may once have been photosynthetic. Curr Biol 18:956–962

Reyes-Prieto A, Weber AP, Bhattacharya D (2007) The origin and establishment of the plastid in algae and plants. Annu Rev Genet 41:147–168

Rodriguez F, Derelle E, Guillou L, Le Gall F, Vaulot D, Moreau H (2005) Ecotype diversity in the marine picoeukaryote *Ostreococcus* (Chlorophyta, Prasinophyceae). Environ Microbiol 7: 853–859

Roeder V, Collen J, Rousvoal S, Corre E, Leblanc C, Boyen C (2005) Identification of stress gene transcripts in *Laminaria digitata* (Phaeophyceae) protoplast cultures by expressed sequence tag analysis. J Phycol 41:1227–1235

Romari K, Vaulot D (2004) Composition and temporal variability of picoeukaryote communities at a coastal site of the English Channel from 18S rDNA sequences. Limnol Oceanogr 49:784–798

Rusch DB, Halpern AL, Sutton G, Heidelberg KB, Williamson S, Yooseph S, Wu D, Eisen JA, Hoffman JM, Remington K, Beeson K, Tran B, Smith H, Baden-Tillson H, Stewart C, Thorpe J, Freeman J, Andrews-Pfannkoch C, Venter JE, Li K, Kravitz S, Heidelberg JF, Utterback T, Rogers YH, Falcón LI, Souza V, Bonilla-Rosso G, Eguiarte LE, Karl DM, Sathyendranath S, Platt T, Bermingham E, Gallardo V, Tamayo-Castillo G, Ferrari MR, Strausberg RL, Nealson K, Friedman R, Frazier M, Venter JC (2007) The Sorcerer II global ocean sampling expedition: northwest Atlantic through eastern tropical Pacific. PLoS Biol 5:398–430

Serrao EA, Pearson G, Kautsky L, Brawley SH (1996) Successful external fertilization in turbulent environments. Proc Natl Acad Sci U S A 93:5286–5290

Shimizu Y, Kitade Y, Saga N (2004) A nonradioactive whole-mount in situ hybridization protocol for *Porphyra* (Rhodophyta) gametophytic germlings. J Appl Phycol 16:329–333

Siaut M, Heijde M, Mangogna M, Montsant A, Coesel S, Allen A, Manfredonia A, Falciatore A, Bowler C (2007) Molecular toolbox for studying diatom biology in *Phaeodactylum tricornutum*. Gene 406:23–35

Siegel BZ, Siegel SM (1973) The chemical composition of algal cell walls. CRC Crit Rev Microbiol 3:1–26

Slapeta J, López-García P, Moreira D (2006) Global dispersal and ancient cryptic species in the smallest marine eukaryotes. Mol Biol Evol 23:23–29

Spector DL (1984) Dinoflagellate nuclei. In: Spector DL (ed) Dinoflagellates, Academic Press Inc, New York, pp 107–147

Suttle CA (2005) Viruses in the sea. Nature 437:356–361

Tanikawa N, Akimoto H, Ogoh K, Chun W, Ohmiya Y (2004) Expressed sequence tag analysis of the dinoflagellate *Lingulodinium polyedrum* during dark phase. Photochem Photobiol 80:31–35

Tyler BM, Tripathy S, Zhang X, Dehal P, Jiang RH, Aerts A, Arredondo FD, Baxter L, Bensasson D, Beynon JL, Chapman J, Damasceno CM, Dorrance AE, Dou D, Dickerman AW, Dubchak IL, Garbelotto M, Gijzen M, Gordon SG, Govers F, Grunwald NJ, Huang W, Ivors KL, Jones RW, Kamoun S, Krampis K, Lamour KH, Lee MK, McDonald WH, Medina M, Meijer HJ, Nordberg EK, Maclean DJ, Ospina-Giraldo MD, Morris PF, Phuntumart V, Putnam NH, Rash S, Rose JK, Sakihama Y, Salamov AA, Savidor A, Scheuring CF, Smith BM, Sobral BW, Terry A, Torto-Alalibo TA, Win J, Xu Z, Zhang H, Grigoriev IV, Rokhsar DS, Boore JL (2006) *Phytophthora* genome sequences uncover evolutionary origins and mechanisms of pathogenesis. Science 313:1261–1266

Vaulot D, Eikrem W, Viprey M, Moreau H (2008) The diversity of small eukaryotic phytoplankton ($\leq$3 mum) in marine ecosystems. FEMS Microbiol Rev 32:795–820

Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA, Wu D, Paulsen I, Nelson KE, Nelson W, Fouts DE, Levy S, Knap AH, Lomas MW, Nealson K, White O, Peterson J, Hoffman J, Parsons R, Baden-Tillson H, Pfannkoch C, Rogers YH, Smith HO (2004) Environmental genome shotgun sequencing of the Sargasso Sea. Science 304:66–74

Waaland JR, Dickson LG, Watson BA (1990) Protoplast isolation and regeneration in the marine alga Porpyra nereocystis. Plantation 181:522–528

Waaland JR, Stiller JW, Cheney DP (2004) Macroalgal candidates for genomics. J Phycol 40:26–33

Wilson WH, Schroeder DC, Allen MJ, Holden MTG, Parkhill J, Barrell BG, Churcher C, Hamlin N, Mungall K, Norbertczak H, Quail MA, Price C, Rabbinowitsch E, Walker D, Craigon M, Roy D, Ghazal P (2005) Complete Genome Sequence and Lytic Phase Transcription Profile of a Coccolithovirus. Science 309:1090–1092

Worden AZ, Lee J-H, Mock T, Rouzé P, Simmons MP, Aerts AL, Allen AE, Cuvelier ML, Derelle E, Everett MV, Foulon E, Grimwood J, Gundlach H, Henrissat B, Napoli C, McDonald SM, Parker MS, Rombauts S, Salamov A, Von Dassow P, Badger JH, Coutinho PM, Demir E, Dubchak I, Gentemann C, Eikrem W, Gready JE, John U, Lanier W, Lindquist EA, Lucas S, Mayer KFX, Moreau H, Not F, Otillar R, Panaud O, Pangilinan J, Paulsen I, Piegu B, Poliakov A, Robbens S, Schmutz J, Toulza E, Wyss T, Zelensky A, Zhou K, Armbrust EV, Bhattacharya D, Goodenough EW, Van de Peer Y, Grigoriev IV (2009) Green Evolution and Dynamic Adaptations Revealed by Genomes of the Marine Picoeukaryotes. Science 324:268–272

Yan X, Fujita Y, Aruga Y (2000) Induction and characterization of pigmentation mutants in *Porphyra yezoensis* (Bangiales, Rhodophyta). J Appl Phycol 12:69–81

Yoon HS, Grant J, Tekle YI, Wu M, Chaon BC, Cole JC, Logsdon JM Jr, Patterson DJ, Bhattacharya D, Katz LA (2008) Broadly sampled multigene trees of eukaryotes. BMC Evol Biol 8:14

Yoon HS, Hackett JD, Bhattacharya D (2006) A genomic and phylogenetic perspective on endosymbiosis and algal origin. J Appl Phycol 18:475–481

Yoon HS, Hackett JD, Ciniglia C, Pinto D, Bhattacharya D (2004) A molecular timeline for the origin of photosynthetic eukaryotes. Mol Biol Evol 21:809–818

Zaslavskaia LA, Lippmeier JC, Kroth PG, Grossman AR, Apt KE (2000) Transformation of the diatom *Phaeodactylum tricornutum* (Bacillariophyceae) with a variety of selectable marker and reporter genes. J Phycol 36:379–386

Zaslavskaia LA, Lippmeier JC, Shih C, Ehrhardt D, Grossman AR, Apt KE (2001) Trophic conversion of an obligate photoautotrophic organism through metabolic engineering. Science 292:2073–2075

Zhang Z, Green BR, Cavalier-Smith T (1999) Single gene circles in dinoflagellate chloroplast genomes. Nature 400:155–159

Chapter 7
Genomic Approaches in Aquaculture and Fisheries

M. Leonor Cancela, Luca Bargelloni, Pierre Boudry, Viviane Boulo, Jorge Dias, Arnaud Huvet, Vincent Laizé, Sylvie Lapègue, Ricardo Leite, Sara Mira, Einar E. Nielsen, Josep V. Planas, Nerea Roher, Elena Sarropoulou, and Filip A.M. Volckaert

Abstract Despite the enormous input into the worldwide development of fish and shellfish farming in the recent decades, in part as an attempt to minimize the impact of fishing on already overexploited natural populations, the application of genomics to aquaculture and fisheries remains poorly developed. Improving state-of-the-art genomics research in various aquaculture systems, as well as its industrial applications, remains one of the major challenges in this area and should be the focus of well developed strategies to be implemented in the next generation of projects. This chapter will first provide an overview of the genomic tools and resources available, then discuss the application of genomic approaches to the improvement of fish and shellfish farming (e.g. breeding, reproduction, growth, nutrition and product quality), including the evaluation of stock diversity and the use of selection procedures. The chapter will also discuss the use of genomic approaches to study and monitor natural fish and shellfish populations and to understand interactions within their ecosystems.

7.1 Introduction

Although initial efforts in genomics were more directed to humans and mammals as biomedical model organisms, this has extended, among others, to a wide range of fish species, in particular teleosts, which are known to occupy multiple, diversified environments within aquatic ecosystems. Fish have also gone through diverse evolutionary changes, making them valuable experimental model organisms not only because of their economical relevance but also because of comparative genomics, providing new perspectives towards our understanding of molecular and organismal

M.L. Cancela (✉)
Laboratory of Molecular Biology of Marine Organisms, Centre of Marine Sciences (CCMAR), University of Algarve, P-8005-139 Faro, Portugal; Faculty of Marine and Environmental Sciences, University of Algarve, P-8005-139 Faro, Portugal
e-mail: lcancela@ualg.pt

J.M. Cock et al. (eds.), *Introduction to Marine Genomics*,
Advances in Marine Genomics 1, DOI 10.1007/978-90-481-8639-6_7,

evolution. Furthermore, since present day species have evolved following teleost-specific genome duplications, researchers now study how thousands of duplicate sets of genes have evolved, in particular (i) the acquisition of new sequence motifs leading to multidomain proteins, (ii) the diversification and multiplication of complex RNA processing mechanisms leading to a plethora of protein isoforms and variants, (iii) the acquisition of new regulatory elements, leading to a diversification of regulatory mechanisms and spatial-temporal changes in gene expression, or (iv) the acquired functionality of non-coding sequences.

Recently, several fish species, including zebrafish (*Danio rerio*), medaka (*Oryzias latipes*), Japanese pufferfish (*Takifugu rubripes*), spotted green pufferfish (*Tetraodon nigroviridis*), Atlantic salmon (*Salmo salar*), three-spined stickleback (*Gasterosteus aculeatus*), and to a lesser extent some shellfish (oyster and mussel) have had their genomes published (only partially for nuclear genomes of shellfish), and several more are in progress. In addition, growing sets of large EST (Expressed Sequence Tag) collections are being produced for many and diverse fish species (e.g. those recently produced for gilthead seabream (*Sparus aurata*), and European sea bass (*Dicentrarchus labrax*), and shellfish (Pacific oyster *Crassostrea gigas*, blue mussel *Mytilus galloprovincialis*) by the Marine Genomics Europe Network of Excellence). With these draft genomes and their complementary ESTs, it has been possible to begin aligning and comparing these sequenced genomes and deduce their gene structure and corresponding protein sequences. However, once specific sequences are identified and the genome annotated, one wonders whether the function of a given gene/protein/DNA motif is as predicted, or why specific sets of genes/regulatory elements have been targets of positive selection. In contrast to fish/shellfish genome sequencing and comparative analysis, which have taken a leap forward recently, functional genomic tools have lagged behind and therefore fish/shellfish-derived in vitro and in vivo tools for functional analysis remain an important aspect requiring further efforts towards a continued development.

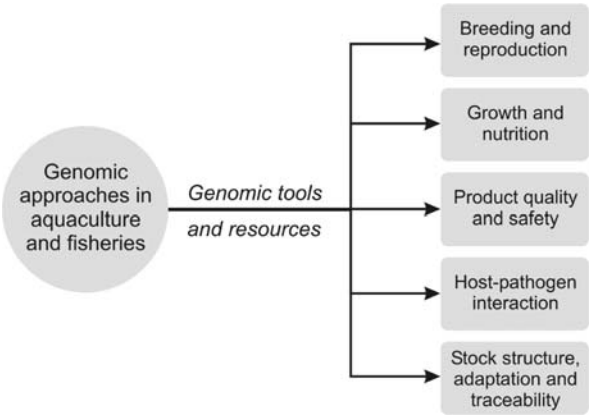

Fig. 7.1 Overview of genomic approaches in aquaculture and fisheries (Figure credit: V. Laizé)

Despite the enormous, worldwide development of aquaculture in the last decade, the application of genomics to aquaculture remains poorly exploited. Improving state-of-the-art genomics research in various aquaculture systems, and its industrial applications remain major challenges and must be the focus of the next generation of projects. This chapter will discuss the application of genomic approaches to the improvement of fish and shellfish aquaculture, including the evaluation of stock diversity and the use of selection procedures. The chapter will also discuss the use of genomic methods to study and monitor natural fish and shellfish populations and to understand interactions within their ecosystems (Fig. 7.1).

7.2 Genomic Tools and Resources

In this section, available (or soon-to-be available) tools for structural and functional genomics in marine fish and shellfish will be described. As the reader will note, most of these tools have been developed for farmed or fished species. One obvious reason for this limitation is the fact that the development of genomic tools requires a relevant economic investment, which has been possible so far only for commercially important species and model species. However, the introduction of new technologies, e.g. ultra high-throughput DNA sequencing (Margulies et al. 2005), has dramatically lowered the costs of EST production, generation of BAC clones, and whole-genome sequencing, therefore more and more marine fish and shellfish species are expected to enter the genomic arena.

7.2.1 Genetic Linkage Maps

Historically, the first portrait of an animal genome was obtained in the form of a genetic linkage map. Any polymorphic locus, i.e. a locus where at least two distinguishable alleles are observed, can be positioned on a genetic map (Fig. 7.2). In the last 30 years, different types of markers have been successively used: allozymes, minisatellites, RAPDs (Random Amplification of Polymorphic DNAs), AFLP (Amplified Fragment Length Polymorphisms), SSRs (Simple Sequence Repeats) and SNPs (Single-Nucleotide Polymorphisms) (Schlötterer 2004; see Glossary and Chapter 3 for detailed explanations about these markers). Allozyme markers have been widely used as they are highly polymorphic in most species, especially bivalve shellfish (Solé-Cava and Thorpe 1991). They have permitted a large number of population genetic studies and are still used, notably in combination with other types of markers (e.g. Nikula et al. 2008). Heterozygote deficiencies and relationships between their heterozygosity and fitness-related traits have been frequently observed in wild or farmed populations using these markers (Raymond et al. 1997, Bierne et al. 2000). Consequently, the non-neutrality of allozyme markers has been strongly debated in bivalves (e.g. McDonald et al. 1996).

To date, most markers are based on DNA technologies. At present, the most popular types are the microsatellites, also known as SSR loci. SSR loci consist of

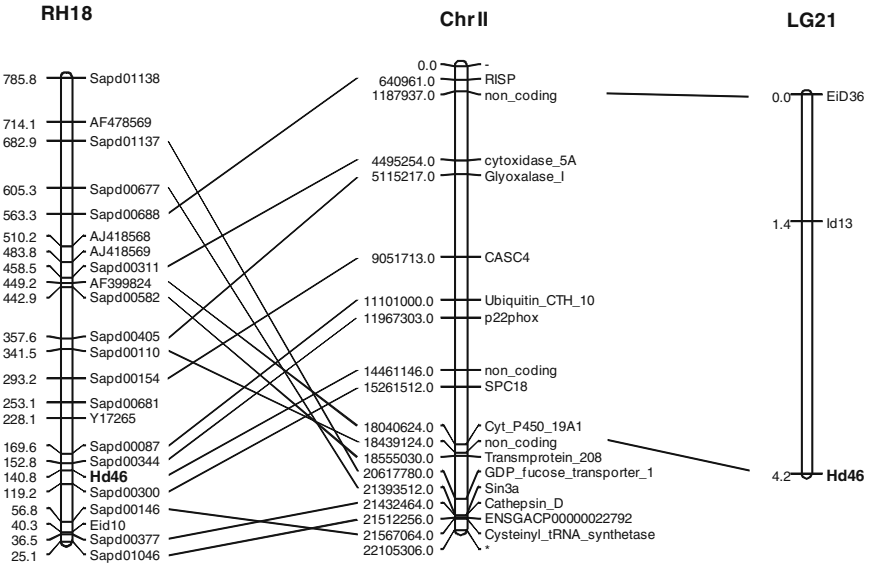

Fig. 7.2 Comparative mapping. RH18 (radiation hybrid group 18) from seabream radiation hybrid map (Sarropoulou et al. 2007), LG21 (linkage group 21) from seabream genetic linkage map (L. Bargelloni, unpublished data) and Chr II (chromosome II) from stickleback genome (www.ensembl.org). Marker names on RH map correspond to unique transcripts from SAPD database (identified with prefix Sapd), publicly available genes or SSR markers (identified with GenBank accession numbers) and unpublished SSR markers (Eid10 and Hd46). Any PCR-amplifiable sequence can be mapped onto the RH map (see text), allowing for a higher marker density. On the other hand, only polymorphic SSR markers (Eid36, Id13 and Hd46) can be located on the linkage map. In *bold*, the single SSR locus present in both maps. Correspondences between individual seabream markers and putative homologues of seabream markers in stickleback genome are shown. Several changes/errors in gene order between the two species are evident. (Figure credit: L. Bargelloni)

a variable number of short repeats (2–6 nucleotides). SSRs are usually uniformly distributed in genomes, at a relatively high frequency (1 every 1.5–6 kb) (Zane et al. 2002, Chistiakov et al. 2006). The repetitive nature of the SSR core sequence causes the frequent gain or loss of one or more repeat unit, creating alleles of different length (Ellegren 2004). However, imperfect microsatellites are often observed, notably in shellfish where very high (>50) number of alleles per locus are often reported (e.g. Huvet et al. 2004). For this reason, most SSR loci show high polymorphism and are extremely useful for the construction of genetic linkage maps (Schlötterer 2004). Marine fish and shellfish species for which a genetic linkage map is available were reviewed by Wenne et al. (2007). Most of these maps are based on SSR loci, although in some cases AFLPs are used as well (e.g. sea bass, channel catfish, mussel, flat oyster). As SSR loci were first characterized in non coding regions, efforts have been made to identify and map SSRs with EST sequences (EST-SSR). Recent efforts to generate ESTs in many marine species contribute to

an increasing number of available EST-SSR (e.g. Yu and Li 2008). Another possibility to position genes on a genetic linkage map is to use SNP loci. While SSR loci are well established genetic markers, SNPs represent a recent addition to the genomic toolbox for most animal species. So far, the representation of SNP markers in linkage maps of marine species has been limited to few loci but the development of high throughput SNP genotyping should allow a significant increase in the number of mapped SNPs in the near future. Compared to SSR loci, SNPs are more frequent in the genome, both in coding and non-coding regions, and are bi-allelic, which makes them portable across laboratories, and more amenable to high-throughput analysis and automation. The frequency of SNPs is particularly high in marine bivalves (e.g. one SNP every 60 bp in coding regions and one every 40 bp in non-coding regions; Sauvage et al. 2007), which have high effective populations sizes (i.e. the number of effective breeders at each generation). SNPs are indeed less polymorphic than SSRs, but their frequency in most species and the implementation of high-throughput technologies makes possible the identification of hundreds of thousands SNPs. Likewise, the genotyping of up to 1,000,000 SNP loci is now possible (e.g. Illumina Infinium HD Human 1 M). At a much smaller scale, SNP isolation and genotyping is well advanced, especially in the Atlantic salmon (Hayes et al. 2007), while other marine species are quickly progressing (e.g. Atlantic cod, European sea bass and Pacific oyster).

Whatever the type of marker used for the construction of a linkage map, it is necessary to sample a large number of independent meiotic events in order to measure recombination between loci. Traditionally, this has been achieved through the use of dedicated mapping panels, i.e. experimental populations originating from F1, backcross or F2 crosses between highly divergent strains. As such populations are not yet available for most marine fish and shellfish, mapping panels are often the F1 generation from a cross between two wild animals. Despite these limitations, the high levels of heterozygosity observed in most marine species (having very large effective population sizes) increases the probability of each marker being informative for the construction of the map.

Linkage maps represent a scaffold of markers for genomic studies. These markers can be used in genome-wide scans for linkage with phenotypic characters of economic or scientific interest. Apart from the identification of Quantitative Trait Loci (QTL) in farmed species, one excellent example of the potential of such methodology comes from the identification of markers linked to loci responsible for morphological divergence between sympatric populations of three-spined stickleback. Peichel et al. (2001) used a medium resolution linkage map to analyze the genetic basis of recently evolved changes in skeletal armour and feeding morphologies seen in the benthic and limnetic stickleback species from Priest Lake, British Columbia.

The main limitations of currently available linkage maps are the scarce representation of coding genes among mapped loci and the low level of comparability between linkage maps from different species. Low-to-moderate resolution linkage maps have been developed during the last decade for several fish and shellfish of aquacultural importance (for review, see Wenne et al. 2007). Most of the maps currently available are based on F1 crosses and AFLP markers. Many of them are first

presented as "preliminary" and are complemented in a second step by co-dominant markers (e.g. Pacific abalone: Liu et al. 2006). Microsatellite-based maps (e.g. Pacific oyster: 88 markers mapped; Hubert and Hedegock 2004; Pacific abalone: 167 markers mapped; Sekino and Hara 2007) are more demanding to develop and often still present fewer markers than those based on AFLP markers. SNP-based linkage maps are currently been developed in several species.

Comparative genomics has its foundation in the presence of conserved regions across species, which are putatively homologous and can be used to align and compare genomes. A small proportion of SSR loci are conserved, allowing anchoring across genomes (Stemshorn et al. 2005, Franch et al. 2006), but typically microsatellite markers are species-specific and homologous genomic regions are often not recognizable in other genomes. This is especially the case in shellfish species showing high levels of nucleotide polymorphism leading to poor cross-species amplification (e.g. Hedgecock et al. 2004).

7.2.2 Radiation Hybrid (RH) Maps

RH maps were originally developed precisely to overcome the limitations of genetic linkage maps (Walter et al. 1994). A detailed description of RH map construction is beyond the scope of the present book and the reader should refer to specific publications (e.g. Kwok et al. 1998). Briefly, hybrid cell lines are established, which contain random fragments of the donor genome (the species of interest) mixed with the whole genome of host cells (generally hamster cells). A set of independent cell lines recapitulates the genome of the target organism several-fold. Markers or genes of interest are analyzed by PCR using DNA isolated from each cell line. Statistical tools are applied to determine both the linear order of markers on each chromosome, and the confidence of each placement. Any genomic region that is PCR-amplified (i.e. coding genes, SSR, ESTs) selectively from the donor genome can be located on the RH map (Fig. 7.2). This allows mapping of loci that are not polymorphic. Moreover, protein-coding loci are often well-conserved even between distantly-related species; therefore most markers located on a RH map find their homologue in the genomic map of other species, allowing comparison between genomes.

So far, technical problems (compatibility between host and donor cells, sensitivity of donor cells to irradiation) have limited the development of RH mapping panels in non-mammalian vertebrates, and especially in fish. RH maps are available for only two teleost species, the zebrafish (Geisler et al. 1999, Hukriede et al. 1999) and the gilthead seabream (Senger et al. 2006, Sarropoulou et al. 2007). For a third species, the European sea bass, a RH panel has been produced with an improved methodology that is virtually cell-culture free (F. Galibert, personal communication), and a RH map is being constructed. Such technical improvements are expected to make RH panel development faster and easier for any vertebrate species. At the same time, application of high-throughput methods for RH panel scoring (e.g. McKay et al. 2007, Park et al. 2008) hold the promise to drastically

reduce the most expensive and time-consuming step of RH map construction, i.e. marker genotyping on hybrid cell lines. Rapid and relatively inexpensive RH maps will in turn allow much broader genome comparisons and the transfer of information from complete or high-quality draft genome sequences to less characterized ones (Sarropoulou et al. 2008), speeding up the identification of loci involved in biologically important phenotypes. At the same time, medium-high resolution RH maps coupled with low-coverage (1–2X) whole-genome sequencing will provide essentially the same comparative data with respect to gene order that is derived from high-coverage (greater than 7X) genome sequencing (Hitte et al. 2008). The possibility of developing a RH map for the Pacific oyster is also currently being examined (F. Galibert, personal communication).

7.2.3 BAC-Based Physical Maps

While genetic maps provide an indirect estimate of the distance between two loci, a genome-wide physical map gives an estimate of the true distance, in measurements called base pairs, between items of interest (note that RH maps may be considered either to be a type of physical map or to be similar to genetic maps, with the frequency of association between markers being determined by the irradiation dose rather than by genetic recombination). A physical map usually comprises a set of ordered large-insert clones such as bacterial artificial chromosomes (BACs). Physical maps can be independent of genetic information but are more valuable if linked to genetically mapped markers, and are even more powerful if integrated with genomic sequence data. Current physical maps are based on technologies to detect overlaps among BACs. The most commonly used is BAC fingerprinting, where restriction profiles of individual BACs are used to order clones on the map (Meyers et al. 2004). While BAC libraries have been constructed for several marine fish species, high-resolution BAC-based physical maps are not available yet for marine fish species. A low resolution BAC-based map, linking 84 BAC clones to the existing linkage map (Wang et al. 2007b), has been recently reported (Wang et al. 2008) for the barramundi or Asian sea bass (*Lates calcarifer*). Technical improvements in BAC DNA preparation (Kuhl et al. 2010) coupled with high-throughput DNA sequencing led recently to sequencing of both ends of over 50,000 clones of a *Dicentrarchus labrax* BAC library (Whitaker et al. 2006) and of a *Sparus aurata* BAC library. In this case, BAC-end sequences have been ordered by comparison with a high quality genome sequence draft from the three-spined stickleback (Fig. 7.2). Comparison with a closely-related species genome for ordering clones has been already used to provide a complementary framework to other methods (Gregory et al. 2002) and appears to be a good tool for preliminary assembly of BAC clones (R. Reinhardt, personal communication).

Few BAC libraries are yet available in molluscan shellfish. Cunningham et al. (2006) first reported a BAC library for Pacific and American oysters. More recently, Zhang et al. (2008b) have reported the construction of two BAC libraries for the

Zhikong Scallop (*Chlamys farreri*) and their use to identify genes involved in the innate immune system of molluscs. A 10X genome coverage BAC library of the Pacific oyster has been recently produced and fingerprinted as part of a project coordinated by P. Gaffney (University of Delaware, USA).

7.2.4 High Quality Draft Genome Sequences

To date, the genomes of five teleost species, *D. rerio*, *T. nigroviridis*, *O. latipes*, *T. rubripes*, and *G. aculeatus*, have been sequenced at high coverage. The most recent sequence assemblies for all these genome sequences are available at www.ensembl.org. Two of these species, *T. rubripes* (fugu) and *G. aculeatus* (stickleback), are marine fish, although only the former is fully marine while the latter shows a freshwater and a marine morphotype, returning in some cases to freshwater to breed. Sequence coverage of these two fish genomes is only slightly different, 8.7X for *T. rubripes* and 11X for *G. aculeatus*. Nevertheless, a comparison between the two species presents an interesting example of how information content of genome sequences might vary. The fugu genome is 393 Mb in size and is represented by 7,213 scaffolds, with the largest scaffold being 7 Mb in size. On the other hand, the stickleback genome is larger (approximately 460 Mb), but the assembly shows 21 chromosomes (groups) with an additional 1,822 unplaced supercontigs. It is quite obvious that the stickleback genome offers the means for a better and more meaningful comparison with other fish genomes. In fact, both linkage and RH maps of different fish species have been anchored against the stickleback genome (Franch et al. 2006, Sarropoulou et al. 2007, 2008, Bouza et al. 2007; Fig. 7.2), showing a remarkable level of synteny, with the majority of genomic rearrangements occurring at the intra-chromosomal level. As already mentioned, novel DNA sequencing technologies will likely make it possible to undertake whole-genome sequencing projects in non-model species, however supporting tools such as linkage and physical maps will be still necessary for high quality assembly of these genomes. This approach is currently being developed for the Pacific oyster as a joined effort beween P. R. China (Institute of Oceanology of Chinese Academy of Sciences and Beijing Genomics Institute) and members of the international Oyster Genome Consortium. Following the sequence of the limpet *Lottia gigantea* (see JGI website), the oyster will be among the first members of the Lophotrochozoa to be fully sequenced.

7.2.5 Functional Genomic Tools

In the genomic area, one principle focus is to translate information obtained using large scale sequencing projects into enhanced understanding of genome function related to biological mechanisms and phenotypic variability. From the set of EST collections usually incorporated into a publicly available database where

they are assembled into contigs, subsequent microarray design or assignment of SAGE (Serial Analysis of Gene Expression) or MPSS (Massively Parallel Signature Sequencing) signatures allow transcriptome profiling to be carried out to study marine animal physiology. Functional genomics focuses on dynamic aspects of the genome such as gene transcription, translation, and protein–protein interactions. We will refer here to functional genomics in a restricted way, i.e. transcriptomics. The analysis of a whole-genome transcriptome in a single experiment has become feasible with the introduction of the DNA microarray technology (Schena et al. 1995). Eukaryotic cell transcriptomes proved to be much more complex than thought and thus difficult to represent fully in a single experiment (Birney et al. 2007), but DNA microarrays have indeed given a totally new perspective to gene expression studies. There are several technical options for the construction of gene-expression DNA microarray, e.g. probes might be either spotted cDNAs or oligos, or oligos of various lengths can be synthesized in situ using several different synthesis technologies (Holloway et al. 2002). Until recently the array platforms available for marine fish and shellfish species were spotted cDNA microarrays. The alternative approach is now represented by oligonucleotide probes, designed on all *in silico* available cDNA sequences, which are then synthesized and spotted or directly synthesized on the glass slide.

For two oyster species, *C. virginica* and *C. gigas*, an international group of collaborators has constructed a cDNA microarray (4,460 sequences from *C. virginica* and 2,320 from *C. gigas*) (Jenny et al. 2007). This array is notably used to estimate the response of *C. gigas* families to heat stress challenge (Lang et al. 2008). In Europe, within the framework of Marine Genomics Europe network of excellence and the Aquafirst European project, a larger portion of the Pacific oyster transcriptome was spotted to produce a 10X microarray slide. This slide was used to examine the biological mechanisms involved in the response of lines selected to be resistant or sensitive to summer mortality (Boudry et al. 2008; E Fleury, personal communication). In the blue mussel, a species commonly used as a sentinel to monitor pollution in the marine environment, a first cDNA microarray was designed including 1,714 probes allowing the identification of about 50 signatures of relevant doses of pollutants in mussel tissues (Venier et al. 2006). A low-density oligonucleotide microarray, representing 24 mussel genes selected on the basis of their potential involvement in mechanisms of pollutants and xenobiotic response, was also validated (Dondero et al. 2006). Finally, from microarray data obtained in the intertidal mussel *Mytilus californiacus* across major portions of its biogeographical range, Place et al. (2008) emphasised the usefulness of such transcriptomic tools to marine ecologists for ecological studies including those relevant to the marine estuarine habitat of the bivalves.

The expanding number of EST available in public data bases for many marine species (e.g. *Salmo salar* 433,337, *G. aculeatus* 276,992, *Gadus morhua* 181,734, *D. labrax* 32,755 and *C. gigas* 56,327) have paved the way for a much broader use of oligo-DNA microarrays. Within the European Network of Excellence "Marine Genomics Europe", a pilot study aid at developing such platform for two marine fish (*S. aurata* and *D. labrax*) has been carried out. Using an in

situ synthesis method based on ink-jet technology, two non-overlapping oligos were designed for each unique transcript. For *S. aurata* 19,715 gene transcripts are represented on an array platform that is coupled with a dedicated database (http://enne.cribi.unipd.it:5555/biomart/martview). This array is now fully validated and publicly accessible (Ferraresso et al. 2008). Likewise, 19,048 unique transcripts have been used to develop a *D. labrax* oligo microarray (L. Bargelloni, personal communication). A similar platform albeit with a reduced transcriptome representation is available for two marine flatfish, *Scophthalmus maximus* (P. Martinez, personal communication) and *Solea senegalensis* (Cerda et al. 2008). The drastically reduced cost of large-scale DNA sequencing will greatly increase available ESTs for a large number of marine fish species, making functional genomic tools easily accessible to the majority of research groups. Ultimately, the trend toward cheaper massive sequencing technologies will lead to direct sequencing becoming the method of choice for transcriptome analysis (Sultan et al. 2008).

The very powerful method of MPSS (Brenner et al. 2000) was used to study heterosis (hybrid vigor), a widely observed phenomena in bivalve molluscs. MPSS technology generates short sequence tags from complex RNA samples to analyse gene expression pattern. MPSS offers advantages for organisms with poorly characterized genomes: (1) it requires no previous sequence information and relies on open-based sampling of transcripts, allowing for the identification of novel transcribed sequences, even if a large EST collection is a great advantage to assign MPSS signatures, (2) MPSS is an unbiased, comprehensive and quantitative method as tag counting provides a digital measurement of the abundancies of transcripts, and (3) MPSS is able to detect rare messages (down to about 3 transcripts per million). Up to now, 4.5 M sequence tags have been identified in *C. gigas* genotypes containing 23,274 distinct signatures that allowed the characterisation of some 350 candidate genes for growth heterosis (Hedgecock et al. 2007). SAGE is another method producing signatures to generate transcriptome profiling (Velculescu et al. 1995). This method makes the same assumptions and has the same advantages as MPSS. To date, there are no published articles describing SAGE analysis of marine molluscs but SAGE will be soon applied to the oyster for genome-wide expression profiling of the hemocytes, the immuno competent cells (E. Bachère, personal communication). Finally, a recent interesting comparison of these three methods (Nygaard et al. 2008) has concluded that the oligoarrays can provide quantitative transcript concentrations that are correlated with MPSS and SAGE data but the absolute scale of the measurements differs across the technologies.

7.3 Genomic Approaches in Breeding and Reproduction

The ultimate goal of fish and shellfish selective breeding programmes is to increase the sustainability and profitability of aquaculture, while maintaining the genetic variability in the cultured stocks and limiting its impact on wild population and the environment. Pedigree information, which is required for efficient breeding programmes in order to maximise effective population sizes and to use information

from related individuals to increase the accuracy of predicting breeding values, is often lacking. Molecular markers can be used for this purpose as well as to detect regions of interest in the genomes linked to a trait of interest, i.e. QTL. This section is dedicated to the possibilities and first realizations of incorporating molecular marker information into fish and shellfish breeding programmes by considering their use for (1) genealogical traceability and genetic variability maintenance, and (2) QTL search and marker-assisted selection (MAS).

Practically, two of the main constraints facing effective breeding programmes for fish and shellfish are the fact that (1) in most species early stages are too small to be tagged individually and, at the same time, (2) spatial and technical constraints strongly limit the number of rearing vessels that can be managed at one time. To address this issue, mixtures of equal-aged progenies from different families can be reared together to avoid family-specific environmental effects. Molecular markers can be used subsequently to assign animals to families after the evaluation of their performance, which is of particular interest for individual traits such as growth rate. SSRs are currently the type of marker most commonly used for paternity analysis and traceability in aquaculture species (Herbinger et al. 1995, Fishback et al. 2002, Vandeputte et al. 2004), mainly due to their high levels of variability and power to discriminate at the individual level. Although it depends on the size of the breeding population, typically, 10–20 variable genetic markers are needed to assign >95% of individuals to single pairs of parents (e.g. Vandeputte et al. 2006). However, high frequencies of null alleles are commonly observed in bivalves (Hedgecock et al. 2004), and fish (Castro et al. 2004, 2006) which can lead to difficulties and bias in such analyses. Parentage assignment and relatedness studies have been made using microsatellite markers in several species, on an experimental level involving a limited number of genitors in molluscs (Boudry et al. 2002, Taris et al. 2006, Li and Kijima 2006, McAvoy et al. 2008), seabream (Castro et al. 2007), Atlantic salmon (Norris et al. 2000), Atlantic cod (Herlin et al. 2007, 2008), rainbow trout (Mcdonald et al. 2004), turbot (Borrell et al. 2004) and sole (Porta et al. 2006). However, the suitability of such approaches for mixed-family breeding programs has not been yet demonstrated, despite the large number of markers already available (Li et al. 2003a). AFLPs (Gerber et al. 2000) as well as SNPs have also been considered (Anderson and Garza 2006), with the need for these latter to develop cost-efficient high-throughput genotyping methods (Hayes et al. 2005). As mean SNP density is very high in oysters (Curole and Hedgecock 2005, Sauvage et al. 2007), and likely to be of a similar level in many other marine bivalves, the number of potential SNP markers is extremely high.

Traceability has also become an area of interest for aquaculture species in order to follow individuals back to their origin for estimates of escapees from farms or to identify sources of diseases and/or toxins in market fish (see Section 7.5 of this chapter and Chapter 1 of this book).

Other potential applications for molecular markers are walk-back selection (Li et al. 2003b, Sonesson 2005) and pedigree-assisted selection methods such as animal model-based methods (Lynch and Walch 1998). Bias in estimating genetic parameters is expected to arise due to the low level of self-fertilisation that can

occur in some shellfish species. This is particularly true for scallop (Martinez and di Giovanni 2007) but also for the European flat oyster (Lallias et al. 2008). Therefore, molecular markers can be used to discriminate between selfed and outcrossed families or individuals within a family.

The choices made at the founding of a breeding programme are critical for its long term success. This is especially true for the choice of the founder individuals. Hence lack of adequate base populations is the main reason for the lack of selection response observed in some fish (Gjedrem 2000) and shellfish (Naciri-Graven et al. 2000). Increasing the effective population size in a breeding programme will decrease random genetic drift and increase likelihood of showing response to selection. Molecular markers can be used to infer relatedness between individuals available as candidate bloodstock to generate the first generation of offspring or at any level of reproduction in a breeding programme (Hayes et al. 2006), and thereby avoid mating among close relatives and prevent negative consequences of high levels of inbreeding and genetic load (Gallardo et al. 2004, Camara et al. 2008). Hence, molecular markers are used on a routine basis to monitor genetic diversity of broodstock of oysters (Hedgecock and Davis 2007), seabream (Blanco et al. 2007), trout (Was and Wenne 2002, Gross et al. 2007), salmon (Norris et al. 1999, Koljonen et al. 2002, Rengmark et al. 2006), cod (Pampoulie et al. 2006) and flounder (Liu et al. 2005b). Furthermore, in the European flat oyster (Launey et al. 2001) and Japanese flounder (Sekino et al. 2002), microsatellites were used to demonstrate a loss of genetic variability in mass selected populations.

Currently, MAS does not yet play a major role in genetic improvement programmes in any of the agricultural sectors, and this is particularly so in aquaculture. Traditionally, fish and shellfish selective breeding programmes have targeted traits that can be easily individually recorded and improved using mass selection (body weight, growth, etc.). However traits that are difficult, expensive or time consuming to score or that express late in the life of the organism (disease resistance, carcass quality, feed efficiency, sexual maturation, etc.) may be considered as good candidates to perform MAS in the context of the recent and important development of molecular markers and high through-put genotyping techniques. Furthermore, compared with genetic modification, the use of MAS is more relaxed, at the level of research and development, as well as field testing, commercial exchanges, or public acceptance for which the technology is not an issue.

As a prerequisite for MAS, there must be a known association between genetic markers and genes affecting the phenotype (trait) of interest. Searching for those associations corresponds to searching for QTL. To date only two published papers report the identification of QTLs in shellfish species compared with more than 20 in fish species. For example, in fish, QTLs have been detected affecting cold tolerance and body weight in several species of *Oreochromis* (Cnaani et al. 2003, Moen et al. 2004), body weight in salmonids (Reid et al. 2005 and references therein), disease resistance (Ozaki et al. 2001, Moen et al. 2004, Rodriguez et al. 2004, Khoo et al. 2004, Cnaani et al. 2004), or upper thermal resistance in salmonids (Somorjai et al. 2003 and references therein). In shellfish, Yu and Guo (2006) identified QTLs for resistance to *Perkinsus marinus* in the American oyster *C. virginica* and Liu

et al. (2007) described QTLs for shell, muscle, gonad, digestive gland and gill weight in Pacific abalone *Haliotis discus hannai*. In addition, positive results have recently been obtained in European flat oyster *Ostrea edulis* for *Bonamia ostreaea* resistance (Lallias et al. 2008), and in the Pacific oyster *C. gigas* for summer mortality resistance (C. Sauvage, personal communication) and heterosis for growth (D. Hedgecock, personal communication).

The candidate gene approach is another alternative, which consists in looking for variation at genes with a known role in the physiology underlying complex traits to explain phenotypic variability for those traits. Hence, direct association between gene polymorphism and phenotypic variation in growth has recently been reported in oysters (Prudence et al. 2006). Physiological causes of such differences have also been investigated in relationship with feeding-related traits and specific amylase activity revealed an association between specific alleles of amylase genes (Huvet et al. 2008). Similarly, relationship of both glutamine synthetase (amino acid metabolism) and delta-9 desaturase (lipid metabolism) genes with resistance to summer mortality was reported by David et al. (2007). In the Arctic charr, candidate genes have been studied for growth related traits using ten conserved gene sequences known to be related to the growth hormone axis (Tao and Boulding 2003) and one SNP was found to be associated with growth rate. In the Atlantic salmon, specific alleles or heterozygotes at genes of the major histocompatibility complex (MHC) were associated with resistance and susceptibility to the infectious haematopoietic necrosis virus (Langefors et al. 2001, Lohm et al. 2002, Arkush et al. 2002, Grimholt et al. 2003, Bernatchez and Landry 2003).

In addition to these a priori approaches, differential gene expression studies can provide new candidate genes. Differential gene expression between oysters selected to be resistant or sensitive to summer mortality (Huvet et al. 2004), or exposure to pollutants (Boutet et al. 2004, Tanguy et al. 2005), have led to the identification of large numbers of candidate ESTs. Until now, Suppression Subtractive Hybridization (SSH) has been the method the most frequently used to identify genes differentially expressed between contrasting individuals. However, novel high-throughput transcriptome analysis methods such as microarrays (Jenny et al. 2007), MPSS (Hedgecock et al. 2007), or SAGE are likely to increasingly contribute to the identification of genes of interest in the near future. In salmonids, a microarray has been used to study gene expression in fish exposed or not to *Piscirickettsia salmonis* (Rise et al. 2004). Finally, through the co-localization of QTLs and candidate genes, there should be a mutual reinforcement of both approaches that would benefit the ultimate progression of improvement programs based on such knowledge.

Although molecular markers can already be used in fish and shellfish breeding programmes to trace individuals for easier rearing, escapes estimation, or optimizing population size in bloodstocks, on a case by case study basis, QTL mapping and MAS are not as well advanced in aquaculture species as in farmed terrestrial plants and animals. However, the merging efforts between genetics and genomics are expected to allow the detection of variation affecting complex traits in fish and shellfish species and their use for increasing the usefulness of MAS schemes.

7.4 Genomic Approaches in Growth and Nutrition

7.4.1 Introduction

One of the main objectives of fish aquaculture is to produce fish with an optimal growth rate. In the wild, the overall fitness of fish populations, and particularly their reproductive fitness, depends on the ability of fish to achieve a certain growth rate. From a physiological point of view, fish growth is a complex process that depends on other mutually interdependent physiological processes, such as development, nutrition and metabolism. Fish growth is commonly viewed as an increase in fish length and, in particular, of its muscle mass in parallel to the bone structure and organs. Skeletal muscle is the most important tissue for growth since it may represent more than 50% of the body mass of the fish and is the tissue destined for human consumption. Most fish species have the ability to grow continuously throughout their lifetime (Mommsen 2001). Growth is dependent on the production and development of muscle fibers and on their metabolic capacity, which in turn depends on food intake as well as on the supply, transport and utilization of nutrients. Therefore, tissues directly involved in nutrition, such as the intestine, liver and adipose tissue, play an essential role in skeletal muscle growth. The liver, due to its capacity to store carbohydrates as glycogen and to use them, to produce glucose by gluconeogenesis and to synthesize and store lipids, is possibly the most important organ contributing to skeletal muscle metabolism. Since growth is a physiological process with multiple tissue contributions, a global and multidisciplinary approach is required in order to have a complete view of all the factors contributing to muscle growth. At the present time, this integrative physiological approach (i.e. Systems Biology) can be used to integrate all the information obtained from the application of transcriptomic, proteomic and metabolomic tools and reconstruct the pathways and functional networks that govern the process of muscle growth.

7.4.2 Transcriptomic Changes in Skeletal Muscle Related to Muscle Growth

Over the last few years, a growing tendency to use high-throughput technologies to study muscle growth in fish has been observed. There are already a few transcriptomic studies related to muscle growth and most of them are restricted to salmonid species. The first approach to study the transcriptome of the growing fish muscle consisted in the analysis of gene expression in transgenic salmon for growth hormone (GH). Overexpression of GH in white skeletal muscle caused muscle hyperplasia accompanied by increased expression of genes involved in transcription, muscle fiber formation and muscle structure, as assessed by substractive hybridization (Hill et al. 2000). A more recent study revealed that transgenic salmon for GH have a higher number and higher proliferation rates of muscle stem cells that can be directly stimulated by GH in vitro; both processes are linked to changes

in the mRNA expression of different myogenic factors (Levesque et al. 2008). However, until very recently, global transcriptional changes in muscle of GH transgenic salmon had not been evaluated using high-throughput approaches. A recent study has reported on the effects of exogenous bovine GH administration on the rainbow trout muscle transcriptome using the GRASP 16 K cDNA microarray (Gahr et al. 2008). In this study, GH was injected in fish from two different rainbow trout families, selected for low- and high- growth rates, and different sets of genes up- and down-regulated were identified in white muscle mediating cellular processes such as cell cycle, immune response, metabolism or protein degradation. Also recently, the muscle transcriptome was investigated in another growth model (fasting and refeeding) using a custom cDNA microarray generated in the INRA-GADIE Resource Center containing 9,023 gene sequences (Rescan et al. 2007). In the early phases of muscle growth recovery, an induction of genes involved in RNA processing, translation, cell proliferation followed by a later phase involved in Golgi and endoplasmic reticulum dynamics and muscle remodeling was observed.

To date, no studies on gene expression related to muscle growth using microarrays have been performed in model organisms like zebrafish and fugu, with fully sequenced genomes. In fugu, subtractive hybridization and quantitative PCR were used to identify differentially expressed transcripts in skeletal muscle from young versus adult individuals (Fernandes et al. 2005). The adult skeletal muscle stops muscle fiber formation, inhibiting the recruitment of muscle fibers, whereas the young skeletal muscle has an active fiber formation. Fernandes et al. identified four novel genes that were expressed at higher levels in fast muscle of adult fugu (Fernandes et al. 2005). These novel genes probably act as growth inhibitors but they do not have significant homology with any known gene. This study provided the initial tools to develop high-throughput strategies to study growth in fugu since they have generated different ESTs libraries that could be used to create cDNA or oligonucleotide microarrays. It is surprising the absence of microarray studies in the growing muscle of zebrafish (hypertrophic growth), since there are oligonucleotides microarrays commercially available (i.e. Agilent) and its genome is fully sequenced and partially annotated.

7.4.3 Transcriptomic Changes in Skeletal Muscle Related to External Factors

External factors such as temperature or infectious pathogens may have an effect on different physiological processes that directly could affect muscle growth. Using high-throughput technologies, the skeletal muscle transcriptome has been studied in response to changes in ambient temperature (Cossins et al. 2006, Gracey et al. 2004, Malek et al. 2004), vitellogenesis-induced atrophy of skeletal muscle (Salem et al. 2006) and vaccination (Purcell et al. 2006). The effects of temperature changes on the skeletal muscle transcriptome were initially studied in zebrafish (Malek et al. 2004). A temperature shift of 10°C provoked an increase in the expression

of genes involved in mitochondrial metabolism, oxidative stress, as well as heat shock proteins 70 and 90. However, these changes in gene expression have not been related with changes in muscular growth by hypertrophy. A very interesting study on the effects of temperature acclimatation was performed in carp (*Cyprinus carpio*) using a custom cDNA array containing 13,349 sequences (Gracey et al. 2004). The authors studied 7 tissues and reported that a decrease in temperature elicits a common basic response in all tissues related to metabolism and an additional, more specific, response related to the function of each tissue. In particular, skeletal muscle undergoes a down-regulation of genes involved in remodeling of the contractile system and an up-regulation of genes involved in protein degradation. Reviews on this topic have been written by Gracey and co-workers (Cossins et al. 2006, Gracey 2007), but the accepted hypothesis about the transcriptional effects of temperature acclimatation is that the muscle suffers a reduction in activity coupled with protein degradation that leads to atrophy. Using a physiological model of muscle atrophy in rainbow trout and assessing changes in gene expression with the GRASP 16 K microarray platform, Salem et al. have identified differentially expressed genes, representing 1% of the total of the genes present in the microarray (Salem et al. 2006). During vitellogenesis, trout skeletal muscle suffers a marked atrophy that is reflected by transcriptional changes that affect different cellular processes, protein degradation being one of the most relevant. This model could be useful for the identification of genes or gene patterns important for evaluating the function of skeletal muscle.

The skeletal muscle is the preferred tissue for DNA vaccine administration due to its easy access and to its high production of expressed antigens. Concerning growth, expression of elevated amounts of antigens in muscle could cause an activation of the immune system that could lead to growth defects. Purcell et al. have evaluated changes in gene expression after DNA vaccination against infectious hematopoietic necrosis virus (IHNV) using the GRASP 16 K microarray (Purcell et al. 2006). Vaccination caused a strong immune response as evidenced by the increase in the expression of genes involved in antigen presentation and viral response in skeletal muscle. In addition, an increase in the expression of marker genes for leukocytes was observed sugesting an infiltration of leukocytes in the muscle tissue. Therefore, microarray analysis of vaccinated skeletal muscle indicates that the activation of a transcriptional response in this tissue could alter its metabolic and/or functional status.

7.4.4 Genomic Approaches to the Study of Hepatic Function

In teleost fish, the liver plays a key role in nutrition and metabolism as an important site for the synthesis of energy reserves in the form of lipids, carbohydrates and proteins. In addition, the liver is essential for the response to fasting by mobilizing energy reserves into the bloodstream and providing nutrients to the tissues and, therefore, is important for the maintenance of homeostasis. In relation to its key role in homeostasis, the liver is one of the main targets for metabolic hormones

such as insulin and glucagon. The liver also plays an important role during growth due to its ability to respond to GH by synthesizing and secreting insulin-like growth factor I (IGF-I), an important mediator of GH growth-promoting effects. Due to the central physiological role of the liver and to the complexity of its function and regulation, genomic approaches are very well suited to assist us in understanding its function and its response to physiological or environmental alterations. To date, several studies have described the response of the hepatic transcriptome and proteome in relation to growth and nutrition in teleost fish.

7.4.4.1 Transcriptional Changes in the Liver in Relation to Growth and Nutrition

Recent studies have addressed the response of the hepatic transcriptome to the growth-promoting actions of GH in salmonid fish using the GRASP 16 K microarray platform. In GH transgenic coho salmon, one of the major transcriptional changes observed was an increase in the expression of genes involved in mitochondrial activity (Rise et al. 2006). It was postulated that this could be due to the higher energy demand of GH transgenic fish with a higher metabolic rate related to their higher growth rate. In addition, hemoglobin genes were also induced in the liver of GH transgenic fish, which was attributed to the possible need to synthesize hemoglobin in relation to the higher metabolic rate of these animals. Surprisingly, this microarray analysis detected only a small set of genes coding for hepatic enzymes, most notably fatty acid desaturase and prostaglandin D synthase. Similarly, a more recent study evaluating the transcriptional response of the rainbow trout liver to a short-term (3 days) treatment with GH and using the same GRASP microarray also reported changes in a small set of genes, very few of those corresponding to hepatic enzymes (Gahr et al. 2008). In this study, the two major functional categories that responded to administration of GH were genes involved in metabolism and immune response. In view of the scarce information on the liver transcriptome in response to GH, it is possible that an important part of the regulation is at the level of translation or post-translation modifications. Therefore, it will be necessary to study the physiological response of the liver also by a proteomic approximation.

From a nutritional point of view, changes in the hepatic transcriptome in response to the adaptation to diets with vegetable oil have been examined. In particular, the hepatic response of Atlantic salmon to diets containing 75% vegetable oil has been studied using a microarray platform constructed with Atlantic salmon ESTs (Jordal et al. 2005). The results showed clear evidence of changes in the expression of genes involved in hepatic liver metabolism. On one hand, an increase in the expression of fatty acid desaturases $\Delta 5$ and $\Delta 9$ was observed. On the other hand, the expression of long-chain acetyl CoA synthase decreased in the liver of fish adapted to the vegetable oil rich diet. Furthermore, several mitochondrial genes and proteins from the external mitochondrial membrane were down-regulated in response to the vegetable oil rich diet. These observations suggested that these fish, as a result of their nutritional challenge, decreased their capacity for β-oxidation in the liver. More recently,

fish oil replacement (100%) with vegetable oil for 62 weeks from first feeding in rainbow trout has also been shown to decrease the expression of two genes involved in lipid metabolism: fatty acid synthase and long-chain fatty acid elongase (Panserat et al. 2008b). In addition to changes in genes involved in lipogenesis, complete fish oil replacement caused changes in the hepatic expression of genes involved in steroid synthesis, xenobiotic detoxification, protein catabolism and transcriptional regulation, interestingly in the absence of changes in body weight, feed efficiency and feed intake (Panserat et al. 2008a,b). In contrast, replacement of animal proteins (fish meal) with plant proteins in the diet caused a reduction of growth rates and feed efficiency and increased feed intake in rainbow trout, which was accompanied by altered expression of genes involved in protein and amino acid metabolism (Panserat et al. 2008b). Not surprisingly, similar changes were observed in the liver of rainbow trout after a 3-week fasting period: decreased expression of genes involved in protein biosynthesis and increase in the machinery needed for protein degradation (Salem et al. 2007).

7.4.4.2 Changes in the Liver Proteome in Relation to Nutrition and Growth

In view of the limited transcriptional response of the teleost liver to growth and nutritional stimuli, studies on the hepatic proteome are needed to detect post-transcriptional changes. To date, studies have evaluated the response of the hepatic proteome of rainbow trout to fasting and to diets in which fish protein is replaced by protein of plant origin. In response to fasting, 24 differentially expressed proteins were detected in the rainbow trout liver. Among the proteins with increased abundance after fasting were enolase and cytochrome C oxidase, on one hand, and cathepsin D, on the other hand, most likely related to the higher energy requirements and protein degradation, respectively, that takes place during fasting (Martin et al. 2001). Using the same proteomic approach, effects of dietary substitution of fish meal by vegetable protein sources were tested on growth performance and liver protein content in trout (Martin et al. 2003, Vilhelmsson et al. 2004). In one study, adaptation of trout to a diet with a partial (30%) substitution of fish meal with soybean meal for 12 weeks resulted in unaltered growth rates but increased protein catabolism, higher protein turnover and higher protein catabolism, accompanied by changes in the abundance of 33 protein spots (Martin et al. 2003). Although not all spots were identified, this study reported changes in the abundance of structural proteins (e.g. keratin and tubulin), lipid binding proteins (e.g. apolipoprotein A) and, most notably, heat shock proteins, which are indicative of a stress response probably induced by the presence of anti-nutritional factors in soy (Martin et al. 2003). In a subsequent study, trout fed a diet in which fish meal was substituted with a mixture of vegetable proteins showed a reduction in growth rate, despite unaltered feed intake, and decreased feed efficiency, mostly due to a decrease in protein utilization (Vilhelmsson et al. 2004). These changes in nutritional parameters were accompanied by changes in protein production, in particular, proteins involved in primary energy metabolism (e.g. production of NADPH and ATP), as well as two proteasome subunits, indicative of an increase in protein degradation. Overall, changes

in protein abundance in the livers of fish fed diets containing protein from vegetable sources suggest that these fish have higher energy demands than fish fed fish meal-based diets.

7.4.5 Conclusions and Future Directions

Transcriptomic and proteomic approaches are now being implemented in growth and nutritional studies in fish. Although only now beginning, application of these technologies to aquaculture and fisheries in the very near future will reside (1) in the identification of genes for marker-assisted selection programs designed to select fish with desired growth rates and nutritional characteristics, (2) in diagnostic of the growth and nutritional status of fish and (3) in assisting on the formulation of new and less-environmentally impacting fish diets. Despite the overall positive trend, published microarray analyses have to date yielded data in the form of, or focused only on, a small number of genes. This is due, in part, to the particular characteristics of available microarray platforms, but also to limitations regarding the supporting bioinformatic analysis. Therefore, most studies have not performed gene expression analyses beyond the single gene level. In order to fully exploit the power of the transcriptomic approach, future genomic studies in aquaculture and fisheries should proceed with an in-depth statistical analysis of functional categories and, most importantly, with a meta-analysis of a large number of microarray experiments on a variety of different conditions or treatments. It is with this type of meta-analysis that true molecular signatures of a given process can be identified and used successfully to describe the physiological status of fish. Finally, strong emphasis should also be placed on extending transcriptomic, as well as proteomic, studies to non-salmonid species.

7.5 Genomic Approaches in Product Quality and Safety

Over recent decades, the strong trend towards healthier eating habits has resulted in a promotion of the consumption of seafood products, since fish and shellfish are identified as an excellent source of important nutrients, including proteins (high biological value, easy digestible), vitamins (A, D and B12), trace minerals (selenium, iodine) and fatty acids (long chain n-3 polyunsaturated fatty acids (n-3 PUFA or omega-3 fatty acids) in which eicosapentaenoic acid, EPA (C20:5 n-3) and docosahexaenoic acid, DHA (C22:6 n-3) predominate). All these characteristics, but mainly a higher intake of n-3 PUFA, have been associated with an essential role in protection against a number of diseases such as cardio-vascular risk, diabetes type 2 or neurodegenerative disorders (Bourre 2005, Calder 2008).

Fish and shellfish demand will continue to grow, in part due to population growth but also because demand is stimulated by economic growth, particularly in societies where consumption per person is low to moderate. In 2020, per capita

seafood consumption is projected to grow throughout the developing world, while developed-country consumption remains virtually unchanged (Delgado et al. 2003). Although provision of seafood from capture fish is declining and partly not sustainable, seafood from aquaculture will potentially overcome this supply issue. It can deliver a product of defined quality, composition and safety to the market in all seasons of the year enabling a greater penetration of fish products in the diet of consumers.

7.5.1 Seafood Quality Has a Multifactorial Background

Quality as a concept must now be considered, as a convergence between consumers' wishes and needs and the intrinsic and extrinsic quality attributes of food products. The increasing number of quality attributes which must be considered, increasing globalization and the heterogeneity in consumption habits between countries are making this convergence progressively more difficult. Fish quality is a broad and complex concept embracing many components, which have differing relative importance for producers, processors, retailers, distributors, caterers, consumers, regulatory authorities and legislators. Recent studies regarding consumers concerns about seafood clearly point out that strongest interest was displayed for aspects related to safety guarantee, quality marks and health benefits (Pieniak et al. 2006, Werbeke et al. 2007). Flesh quality in fish is species-dependent, results from a complex set of intrinsic traits such as muscle chemical composition (fat content, fatty acid profile, glycogen stores, oxidative stability, color) and muscle cellularity (Johnston 1999) and is strongly influenced by a variety of extrinsic factors such as feeding, pre- and post-slaughter handling, processing and storage procedures. The trait of flesh quality has proven difficult to improve by traditional selection because its hereditability is low and the measure for the quality trait is difficult, expensive and in most cases only possible after slaughter. Furthermore, it is now clear that flesh quality has a multifactorial background and is controlled by an unknown number of QTL.

In this section we will try to summarize the information currently available regarding the application of genomic and proteomic approaches to the improvement of flesh quality in aquatic species.

7.5.2 Fish Quality Traits Assessed by Genomic and Proteomic Methods

Selection research in fish has often focused on basic physical traits with high heritability, such as growth rate in salmonids, and this has already made a marked impact on the aquacultured phenotype. Other traits, especially those linked to quality, have received less attention in selection programs. Quality traits are often highly complex, being the end results of a number of physiological processes, each of

which may be regulated by various hormones. A better understanding of the biological basis for quality traits is therefore of great importance in order to successfully produce tailor-made seafood in a predictable way.

7.5.2.1 Colour

Muscle colouration is a complex trait and cannot be deduced from external morphology, which has prevented selection using traditional quantitative genetic methods. This problem has been partially solved in some breeding programs by using scores from full sibs for breeding value predictions and selection (Gjedrem 2000). Flesh colour has medium to low heritability in the main fish species. Selective breeding for this trait is difficult, since phenotypic evaluation requires individuals to be sacrificed for scoring. A single RAPD polymorphism segregating for flesh colour in Coho salmon was used to derive a molecular single locus SCAR marker (Oki206, GenBank accession AY661427) associated with muscle colour traits and suggested as potentially useful in marker-assisted selection (Araneda et al. 2005, Lam et al. 2007). Genetic correlations between muscle carotenoid content and perceived and colorimetric traits were also lower, suggesting that the retention of pigment is not under the control of the same genes as perceived colour. It may be that fillet colour is influenced by factors other than pigment retained in the flesh.

A better understanding of the interaction between carotenoids (e.g. astaxanthin) and salmon muscle proteins is important to achieve better retention of carotenoid in salmon flesh. Knowledge regarding the mechanisms associated to astaxanthin binding to fish muscle is extremely scarce (Matthews et al. 2006). Proteome analysis of membrane bound astaxanthin transport proteins in adult Atlantic salmon revealed some up-regulated proteins in response to dietary astaxanthin (Saha 2005). Further investigations are necessary to elucidate the transport mechanism of astaxanthin from blood plasma to the muscle proteins.

7.5.2.2 Texture (as Muscle Cellularity)

Texture is one of the major criteria of flesh quality. It is a sensory characteristic for the consumer and an important attribute for the mechanical processing of fillets. Because soft texture is frequently reported, the industry is requesting methods able to measure fish texture, and is also seeking answers to what causes fillet softness. Textural properties depend both on chemical composition and structural properties, in particular of myofibrillar and connective tissue proteins. The muscle cellularity (fiber number and distribution) has been found to affect the texture in fish (Johnston 1999, Kiessling et al. 2006). Intra-species comparison has shown that muscle fibres measured as average fibre cross-section area, increases with decreasing sensory firmness in cooked fish (Hurling et al. 1996). Also in fresh and smoked Atlantic salmon and in fresh brown trout, studies have shown a weak decrease in flesh firmness as the size of the fibres increases (or the fibre density decreases) (Johnston et al. 2004, Bugeon et al. 2003).

The muscle fibre number in mammals and avian species is determined by prenatal events (maternal nutrition, bioactive agents, abiotic factors, breed and genotype), but in fish it is also affected by factors post hatching (Johnston 2006). Although the cell biology of myogenesis in teleosts is distinct from that described in mammals, the genes involved in growth regulation are apparently highly conserved (Watabe 2001). The completion of the genome sequences of the Japanese pufferfish (*Takifugu rubripes*) and the zebrafish (*Danio rerio*) provide new opportunities for understanding the molecular regulation of post-embryonic muscle growth. The zebrafish periostin gene was recently found to be important for the adhesion of muscle fiber bundles to the myoseptum and for the differentiation of muscle fibers (Kudo et al. 2004). In rodents, periostin (also named osteoblast-specific factor 2 or Osf2) was found to be strongly upregulated during the muscle regeneration process (Goetsch et al. 2003), and has been considered a candidate gene for enhanced muscle growth in pigs (Bílek et al. 2008). Its potential as a flesh quality trait still remains to be assessed in fish.

Such studies have considerable economic importance in aquaculture since the plasticity of muscle growth under different production conditions is a major factor in determining quality, in particular the texture and processing characteristics of the flesh (Johnston 1999). Furthermore, since the vast majority of teleosts are ectothermic with external fertilisation, environmental factors have far more impact on muscle growth than in mammals (Johnston 2006).

7.5.2.3 Texture (as Affected by Postmortem Degradation)

The texture of fish fillets undergoes rapid changes in the post-mortem stage. A vast body of literature shows that texture, measured as shear force, varies also with farming practices including harvesting stress, slaughter method, dietary factors, storage time and storage temperature (Sigholt et al. 1997, Johnston et al. 2002, Bencze-Røra et al. 2003, Bugeon et al. 2003, Espe et al. 2004). Despite a large bibliography on degradation patterns during post-mortem storage of fish, many uncertainties remain regarding how degradation of specific proteins relates to flesh firmness.

A detailed characterisation of post-mortem changes in fish muscle would benefit from experimental approaches and technologies aimed at parallel analysis of numerous genes and proteins simultaneously. Genomic and proteomic technologies offer a comprehensive approach to study biochemical systems by expanding the investigation from single to multiple genes/proteins simultaneously. Recently, a subtracted cDNA library was used to identify specific genes whose expression is increased in post-mortem muscle of rainbow trout (*Oncorhynchus mykiss*) during on-ice storage (Saito et al. 2006). Of the 200 cDNAs analyzed, 82 had significant homologies to other previously identified fish genes such as troponin I and glyceraldehyde-3-phosphate dehydrogenase (GAPDH). Comparison of gene expression profiles by dot blot hybridization confirmed an increase of mRNA in muscle after 3 h of on-ice storage compared to that at 0 h after death. Real-time reverse transcriptase-PCR analysis indicated that the cells of muscle tissues are able to synthesize troponin I and GAPDH mRNAs for at least 24 h, and maybe up to 48 h, in fish kept on ice.

These results suggest that the transcription profile at cellular level can be a sensitive indicator of freshness during on-ice storage.

Other studies involving proteome analysis have assessed post-mortem changes in seafood (Kjærsgård and Jessen 2003, Martinez and Friis 2004, Morzel et al. 2000, Verrez-Bagnis et al. 2001, Schiavone et al. 2008), demonstrating the complexity of proteolysis in seafood during storage and processing. To assess post mortem muscle integrity, the effect of two different pre-slaughter procedures (limited or 15 min intense muscular activity) on muscle trout proteins was investigated through a proteome analysis (Morzel et al. 2006). Persistent under-representation of desmin, a key cytoskeletal protein, in fish submitted to intense muscular activity suggests that such a pre-slaughter treatment can affect the composition of post-mortem muscle integrity.

Frozen storage is an excellent option to increase the shelf life of fish and meat products but prolonged storage, high frozen storage temperatures, or fluctuating temperatures have a negative effect on product quality such as development of rancid taste and odor in fatty fish (Min and Ahn 2005). Frozen storage also leads to a decrease in protein solubility (Saeed and Howell 2002) and changes in texture quality, such as increased toughness and loss of juiciness (Mackie 1993). Recent studies in rainbow trout and cod identified proteome changes associated to degradation and oxidation processes during frozen storage (Kjærsgård et al. 2006a, b) concluding that both structural (MHC, actin and tropomyosin) and cytoplasmic (creatine kinase and enolase) proteins were being heavily oxidized. On the contrary, nucleoside diphosphate kinase (NDPK) appeared to be only weakly carbonylated, but its solubility was dependent on the frozen storage temperature, opening the possibility for its usage as a biomarker of frozen storage conditions.

The increasing scientific output over recent years in this area indicates that genomics, proteomics and transcriptomics, with their capacity to monitor multiple biochemical processes simultaneously, are methodologies eminently suitable to find biochemical/metabolic markers useful for predicting/monitoring features such as flesh texture of fish farmed under different environmental conditions and subjected to various ante- and post-mortem practices.

7.5.2.4 Nutritional Quality and Health Value

The nutritional quality and health value of fish is mostly associated with its fat content and fatty acid profile in the muscle. Fish are the only major dietary source for humans of n-3 highly unsaturated fatty acids (HUFA) and with the decline of wild stocks, the proportion of farmed fish is increasing in the human diet. The use of high energy diets in farmed fish (e.g. trout, salmon) can increase energy storage in viscera, liver and to a lesser extent in muscle, with the consequence of excess adiposity, generally not desirable in aquaculture products. In general terms, farmed fish tend to have higher levels of whole body lipids than their wild counterparts. The fatty acid composition of fillets from farmed fish is strongly influenced by the feed lipid composition. The relative n-3 HUFA content of farmed fish tends to be lower than that of wild-caught fish, but the amount provided per portion is likely to be similar,

due to higher fat content. Moreover, as a result of global limits on the supply of fish oil, there is a drive to replace dietary fish oils with plant derived oils, which are rich in C18 PUFA but devoid of the n-3 HUFA abundant in fish oils (Tocher 2003). This issue has raised concerns over reduced levels of n-3 HUFA in farmed fish, which could be considered detrimental or likely to compromise the established nutritional benefits of fish for the consumer.

Experimental evidence suggests that the dependence of marine fish on dietary HUFA is caused by deficiency in the activity of one or more of the key enzymes, D5 and D6 fatty acid desaturases, and fatty acid elongases, required for HUFA biosynthesis (Tocher 2003). Comparison of genes encoding key elements in the fatty acid desaturation and elongation pathways between freshwater and marine species is increasing our knowledge of the molecular genetic basis involved in the modulation of PUFA biosynthesis in fish. This area is currently the object of intense research (Zheng et al. 2005, Tocher et al. 2006, Salem et al. 2007, Izquierdo et al. 2008, Leaver et al. 2008, Panserat et al. 2008a). Although much is known regarding the composition and catabolism of lipids, the molecular components necessary for the biogenesis of lipid droplets have remained obscure. Kadereit et al. (2008) reported the characterization of a conserved gene family important for lipid droplet formation named fat-inducing transcript (FIT1 and FIT2). Through a morpholino antisense approach, they showed that by knocking down FIT2 in zebrafish they induced a blockage of diet-induced accumulation of lipid droplets in intestine and liver, highlighting an important role for FIT2 in lipid droplet formation in vivo. Other genes of interest for regulating the extent and the pattern of fat deposition, such as lipin-1a and leptin, are now being studied in farmed animals, including fish (Hanchuan et al. 2006, He et al. 2008).

7.5.3 Other Emerging Quality Traits

As production methods gain importance to many consumers, issues of ethical production, animal treatment and welfare, and environment-friendly production systems as well as sustainability have more and more influence on seafood product choices (Harlizius et al. 2004). This “ethical quality” concept of fish is mainly associated with aquaculture products, but sustainable exploitation conditions in the case of capture fisheries are also gaining importance in seafood markets. This is largely due to increasing public concern about sustainability issues in the food chain and anticipated regulatory changes, but also because the welfare standards by which a fish is reared then slaughtered may produce an impact upon both production and flesh quality. In general, aquatic animal welfare involves philosophical and ethical interpretations of humane practices (Håstein et al. 2005). However, physical health and biological stress indicators are the most universally accepted measures of welfare. An effective health management program must cover all aspects of aquaculture activity including: real time knowledge of the health status of the fish; identifying and managing risks to fish health; reducing exposure to or the spread of pathogens; and managing the use of drugs and/or chemicals (Hill 2005).

Molecular technologies in general can clearly have a direct positive impact on many of the main elements of fish health management. PCR, real-time PCR and nucleic acid sequence-based amplification (NASBA), by enabling the rapid detection, identification and quantification of extremely low levels of aquatic pathogens, are crucial for a more effective disease management, which in turn leads to a reduction in the use of antibiotics and chemicals in the environment. Microarray technologies offer a new dimension to multiplex screening and understanding of host–pathogens interaction. Recombinant DNA technology permits large-scale, low-cost vaccine production. Moreover, DNA vaccination, proteomics, adjuvant design and oral vaccine delivery are foreseen to foster the development of effective fish vaccines in the future (Adams and Thompson 2006). Several recent reviews have assessed the development and future potential of genomic tools to improve health and stress mediated status in fish (see Dios et al. 2008, Martin et al. 2008, Prunet et al. 2008).

The environmental impact of fish farming is also a major concern and is closely associated with excessive feed wastage and sub-optimal nutrient utilization. The inefficient digestion of phytate phosphorus (main form of phosphorus in plant ingredients) by fish has created environmental concerns associated with phosphorus pollution from aquaculture production facilities. To further complicate this situation, phytate is also known to chelate minerals and proteins, making them nutritionally unavailable to fish (Kaushik 2005). One on going strategy involves supplementation of fish feeds with phytase to degrade phytate into inorganic phosphorus, which can then be directly utilized by fish (Vielma et al. 2000). As a model to examine the feasibility and efficacy of producing fish capable of degrading phytate, Japanese medaka (*Oryzias latipes*) transgenic for an *Aspergillus niger* phytase gene were produced and their ability to utilize phytate phosphorus tested (Hostetler et al. 2005). Cell culture techniques, including transfection, RT-PCR, Northern blot, Western blot, and enzyme activity analysis demonstrated that the protein was expressed and actively secreted, without compromising survival and growth, suggesting that similar transgenic approaches could be used in the future for farmed fish to solve this problem.

7.5.4 Seafood Safety

Safety is among the most important food quality issues. To date, the use of genomics in food safety has concentrated on two main areas, the safety evaluation of food components (Ommen and Groten 2004) and the detection of microorganisms which may cause food spoilage or be hazardous to human health (Abee et al. 2004). Safety evaluation in food is generally focused on both, hazard identification (whether a food item causes an adverse health effect), and hazard characterization (the level of exposure required to elicit an adverse health effect). Gathering appropriate data for hazard analysis can be costly and time consuming, requiring detailed toxicological experimentation in animals. Genomic technologies can offer alternatives to classical

toxicological evaluation. Firstly, their high throughput nature means that it is possible to analyze multiple tissues in a timely and cost-effective manner. In addition, DNA amplification is much less time consuming and can easily provide identification of biological contaminants such as microorganisms or viruses. Secondly, by applying transcriptomics, proteomics and metabolomics the full range of biological responses from gene expression to cellular functions can be studied.

7.5.4.1 Health Hazards in Seafood

The development and application of DNA-based analytical techniques to ensure seafood safety is currently the object of intense research efforts. Major hazards in terms of the safety of seafood products are: toxic compounds (e.g. environmental chemical pollutants, biotoxins), viruses, bacteria and parasites. Nowadays, gene expression measured by DNA microarrays is believed to provide a more comprehensive, sensitive and characteristic insight into toxicity than typical toxicological parameters such as morphological changes, altered reproductive capacity or mortality (Steinberg et al. 2008). In recent years, the generation of EST databases for species important in aquaculture and aquatic toxicology, have resulted in the development of DNA microarray platforms where expression of multiple genes can be assessed simultaneously, following exposure to environmental pollutants and natural environmental chemical stressors. Furthermore, expression profiling is an appealing alternative in ecotoxicological screening, both due to the possibility of monitoring multiple classical toxicological biomarker genes simultaneously and the possible discovery of novel biomarkers (for review see Battershill 2005, Heijne et al. 2005, Brul et al. 2006).

7.5.4.2 Allergenicity in Seafood Products

Seafood allergies are a significant public health concern throughout the world and although certainly overestimated in the public perception, they are most prevalent in coastal populations where the consumption and processing of fish and shellfish are high, such as in Scandinavia and Iberian Peninsula countries. Major seafood allergens identified are the Ca^{2+}-binding proteins, parvalbumin in fish and muscle tropomyosin (Pen a1) and arginine kinase (Pen a2) in shrimp (Lehrer et al. 2003). All these proteins seem fairly resistant to heat, chemical denaturation and proteolysis therefore management of food allergy relies on a strict adherence to an avoidance diet aiming at the total exclusion of the offending food.

Recent research has shown that the engineering of hypoallergenic variants of fish parvalbumin and shrimp tropomyosin by site-directed mutagenesis may represent a candidate approach for specific immunotherapy of seafood allergy (Swoboda et al. 2002, Lehrer et al. 2003, Reese et al. 2005). These studies clearly show that by modifying the amino acid sequence at specific locations, it was possible to reduce significantly (in some cases >99%) the reactivity of mutated epitopes.

7.5.5 Seafood Authentication and Traceability

Traceability is increasingly becoming standard across the agri-food industry, largely driven by recent food crises and the consequent demands for transparency within the food chain. European citizens are entitled by law (CR-EC No. 2065/2001 and 104/2000) to information on the scientific name, method of production (farmed or wild), and the area in which wild fish was caught or farmed fish underwent the final developmental stage. Additional legal requirements for the implementation of traceability systems in the food and feed supply chains in Europe are laid down in the General Food Law, Regulation 178/2002/EC, whose article number 18 referring to traceability has become effective since 1st January 2005. The EU Food Law defines traceability as ''the ability to trace and follow a food, feed, food-producing animal or substance intended to be, or expected to be incorporated into a food or feed, through all stages of production, processing and distribution'' (Martinez et al. 2007).

Traditional analytical techniques which relied on protein analysis have been developed for fish species identification: electrophoretic techniques such as isoelectric focusing or SDS-PAGE; chromatographic techniques and immunological techniques such as immunodiffusion and ELISA. Although most of these methods are of considerable value in certain instances, they are not suitable for routine sample analysis because proteins lose their biological activity after animal death, and their presence and characteristics depend on the cell types. Furthermore, most of them are heat labile. Thus, if the product has been subjected to severe processing conditions (such as sterilization in canning) one has to rely on techniques that target small DNA fragments. This is because proteins and DNA are altered so much that they do not render recognizable patterns. For an updated and extensive overview on the PCR-based methods and the application of proteome analyis to fish and fishery products authentication please refer to Martinez and Friis (2004), Martinez et al. (2005) and Gil (2007). An overview of barcoding techniques used in fisheries is also presented in Chapter 1.

Recently, it became clear that to ensure a validation of the authentication methodologies and specifically when using DNA based techniques, quantifiable reference materials in the form of plasmids needed to be developed. In the framework of the EU SEAFOODPlus project, a pool of plasmidic standards has been prepared as reference materials for usage in DNA techniques for fish authentication. An Interlaboratory Ring Test, involving 12 research centres and institutions having genetic fish identification services and offering authentication analysis to fish industry, has been used to validate such standards. This development has been submitted for a patent application. At the same time, a dynamic DNA database including more than 700 DNA sequences from 53 commercial fish species has been made accessible free by internet (http://www.azti.es/DNA_database). Several commercial kits for fish species identification already exist in the market and show potential for field screening purposes in inspection programs (Gil 2007).

Although still in its infancy, a growing global gene expression profiling at the mRNA or protein level is undoubtedly providing us with a better understanding

of gene regulation that underlies certain biological functions associated with the delivery of seafood, safe, with consistent quality and produced according to sound sustainability standards.

7.6 Genomic Approaches in Host–Pathogen Interaction

7.6.1 Host–Parasite Interactions in Fish

In intensively cultured fish species like the European sea bass, the gilthead seabream, the Atlantic salmon or the Senegal sole, infections are more likely to occur and may cause significant economic losses, either directly through pathogen-induced mortality, or indirectly through the impossibility of selling infected animals. Most teleost pathogens, classified in viral, bacterial and parasitical, are species-specific (Table 7.1).

Table 7.1 Most common infection agents encountered in farmed finfish

Infection	Type	Host	Publication
Vibrio anguillarum	Bacterial	Marine fish, salmonids and anadromous species	Samuelsen et al. (2006) Lopez-Castejon et al. (2007) Sepulcre et al. (2007) Boesen et al. (1999)
Aeromonas salmonicida	Bacterial	Salmonids in fresh and seawater	Emmerich and Weibel (1894)
Renibacterium salmoninarum	Bacterial	Salmonids in fresh and seawater	Fryer and Sanders (1981)
Yersinia ruckeri	Bacterial	Salmonids in freshwater	Rucker (1966)
Edwardsiella ictaluri	Bacterial	Catfish, particulary channel catfish *Ictalurus punctatus*	Hawke et al. (1981)
Photobacterium damselae piscicida	Bacterial	Marine fish, e.g. yellowtail *Seriola quinqueradiata*	Toranzo et al. (1991)
Nodavirus (small viruses with a simple architecture)	Virus	Marine fish (also detected in some freshwater fish e.g. *Acipenser sp.* and *Poecilia reticulata*)	Frerichs et al. (1996) Hegde et al. (2003) Athanassopoulou et al. (2004)
Monogenean (flatworms) e.g. *Diplectanum aequans*	Parasite	All fish species	Buchmann and Lindenstrom (2002) and Faliex et al. (2008)

7.6.2 Transcriptomic Characterization of Host Immune Response

7.6.2.1 EST Analysis to Identify Genes Involved in Host Immune Response

EST analysis is one of the most rapid methods for gene discovery and identification, providing useful data for a first insight into gene expression profiles, alternative splicing or differential polyadenylation analysis, as well as for identification of Type I markers. Furthermore, ESTs are the basis for comparative mapping approaches and development of microarrays. Most published data (e.g. Gong et al. 1994, Douglas et al. 1999, Karsi et al. 2002, Sarropoulou et al. 2005a, Bai et al. 2007, Li et al. 2007; Marine Genomics Europe 2004–2008) resulted from EST projects using non-challenged tissues in order to obtain a baseline unigene catalogue. However, to identify transcripts of genes expressed during immune response, ESTs have to be retrieved from cDNA libraries of infected tissues. Studies aiming to identify and isolate transcripts involved in fish immune response have recently been performed for common carp (Kono et al. 2003, Sakai et al. 2005) and European sea bass (Sarropoulou et al. 2009). In the latter, levels of gene expression were analysed using ESTs retrieved from six cDNA libraries derived from tissues infected with *V. anguillarum* (liver, spleen, head kidney, peritoneal exudate, gill and intestine) and from four cDNA libraries derived from tissues infected with nodavirus (head kidney, spleen, brain and liver). Genes differentially expressed upon infection were isolated and further analysed by real-time PCR, thus confirming some of them to be biomarkers for bacterial and viral infections in fish. Altogether, the increasing availability of sequence information from various teleost species, as well as the corresponding functional expression data, are of importance to better understand immunity-related mechanisms of teleost and their evolution.

7.6.2.2 Microarray Analysis to Identify Genes Involved in Host Immune Response

DNA microarrays, which allow high-throughput expression profiling (see Section 7.2.5 Functional genomics tools), have been only used in few studies to identify genes involved in the immune response of aquacultured fish species. One of these studies aimed at profiling gene expression, using a cDNA microarray, during the acute immune response of catfish following infection with Gram-negative bacteria (Peatman et al. 2007). Microarray analysis revealed that expression of the majority of typical acute phase proteins was up-regulated in catfish, together with a set of putative teleost acute phase reactants. A similar study performed in sea bass (E. Sarropoulou, personal communication) showed that expression of several genes involved in iron homeostasis was strongly induced during immune response, suggesting that limitation of free iron may inhibit bacterial growth while avoiding metal-induced cellular damage. In a follow-up study by Peatman et al. (2008) using an oligonucleotide microarray, a detailed analysis of transcripts involved in the MHC I pathway was performed. The authors identified a total of 131 differentially expressed genes, of which 103 were believed to be unique genes. Interestingly,

β-actin, a supposed housekeeping gene and as such frequently used as a reference gene in qPCR assays, was also reported to be amongst the differentially expressed transcripts suggesting that its use under these circumstances may not be adequate. In this study, the authors also reported the stimulation of several signalling pathways upon infection, as well as the existence of significant differences between two closely related species, the blue and the channel catfish. Most of the transcripts differentially expressed between those two species belong to the MHC class I pathway. More recently, microarray profiling was used to characterize the immune response of the yellow tail *Seriola quinquerdiata* upon exposure to immunostimulants such as ConA and lipopolysaccharide (LPS; Darawiroj et al. 2008). Similar studies have been performed on the Atlantic salmon (Ewart et al. 2005, Martin et al. 2006) and the rainbow trout (Tilton et al. 2005, Gerwick et al. 2007). Finally, MacKenzie et al. (2008) have shown in the rainbow trout that it is possible to distinguish the immune response against two different immune agents, i.e. a virus (IHNV) and a bacterial cell wall component, using microarray gene expression profiling.

7.6.2.3 Real-Time PCR to Identify Candidate Markers for Disease Detection

The polymerase chain reaction (PCR), first introduced by Kleppe et al. (1971) and further developed by Mullis and Faloona (1987) allows detection of a given transcript following amplification of a specific part of the genome. The first application using PCR for quantifications was introduced by Higuchi et al. (1993) 10 years later. Today quantitative real-time PCR (qPCR) is commonly used in human diagnostics and in expression studies of various biological systems (Bustin et al. 2005). Accuracy of qPCR depends on factors such as RNA template quality, type of polymerase, primers, reference genes used and data analysis. qPCR has been used in several studies to reveal the role of immune-related genes in the immune response of economically important fish species. For example, qPCR analysis revealed the role of hepcidin, an antimicrobial peptide, in the innate immunity of the gilthead seabream following bacterial infection (Cuesta et al. 2008). Raida and Buchmann (2007) showed by qPCR that expression of both innate and adaptive immune response genes is temperature-dependent in the trout. qPCR has also been used to characterize virulence mechanisms (Sepulcre et al. 2007), prophylaxis and treatment of viral and bacterial-related diseases (Samuelsen et al. 2006).

7.6.3 How Can Genetic Linkage, RH and Physical Maps Contribute to Shedding Light on Fish–Pathogen Interactions?

Besides growth and stress responses, economically important traits for farming of most fish also include immune response and host–pathogen interactions. Regions of the genome corresponding to QTL have been identified and, in some cases, the molecular polymorphism underlying the QTL has been identified in terrestrial

livestock (Stear et al. 2001, Andersson and Georges 2004). However, the lack of appropriate tools for aquaculture species has limited the application of advanced techniques like mapping of QTL and MAS. The few QTL studies performed so far in aquaculture species aimed at improved growth (Moghadam et al. 2007, Wang et al. 2008) and disease resistance (Moen et al. 2007). For QTL detection, molecular tools such as genetic maps, genome maps, and sufficient genomic information (e.g. large collections of ESTs or BAC-end sequences) are required but nowadays, recent developments in bony fishes and aquatic animals are providing the required molecular tools. At present, genome maps of Salmonidae, Cyprinidae, Cichlidae, Moronidae, Paralichthyidae and Ictaluridae are available (Kocher and Kole 2008) as well as molecular markers (Bouza et al. 2007, Sanetra and Meyer 2008), BAC libraries (Matsuda et al. 2001, Whitaker et al. 2006, Wang et al. 2008), expression data (Ewart et al. 2005, Sarropoulou et al. 2005b, Tilton et al. 2005, Gunnarsson et al. 2007, Darawiroj et al. 2008, Darias et al. 2008, Peatman et al. 2008) and mapping panels (Gilbey et al. 2006, Houston et al. 2008). Maps and polymorphic molecular markers are essential molecular tools to better understand host–pathogen interactions and host immune responses at the genome level, and to unravel genes underlying QTL determining host resistance, tolerance and susceptibility to infection. Mapping of candidate genes can significantly narrow down the regions for QTL scanning. Analysis of channel catfish ESTs revealed that approximately 10% of them contained shorts-tandem repeats (Serapion et al. 2004a, b). However none of these STR-containing ESTs represent immune related transcripts or appear to be indirectly in the immune response. Therefore, mapping of immune-related transcripts will enhance the finding of markers linked to QTL related to immune response. Candidate markers can be identified by expression analysis studies, either by microarray technology, qPCR screen or sequencing of cDNA libraries of infected tissues.

7.6.4 Host–Parasite Interactions in Shellfish

7.6.4.1 Improvement of Diagnostic Tools Using Molecular Approaches

Nowadays molecular biology can offer a vast number of tools that allow simple and reliable diagnostics, replacing the more traditional time-consuming techniques requiring greater amounts of sample. PCR has brought to shellfish diagnostic many new advantages and improvements but mostly specificity, swiftness and reliability. PCR has been used to detect *Vibrio* (Hill et al. 1991, Brauns et al. 1991), viruses (Desenclos et al. 1991, Batista et al. 2007), *Listeria* spp. (Jeyasekaran and Karunasagar 1996), *Bonamia ostreae* (Cochennec et al. 2000), *Salmonella* spp. (Dupray et al. 1997), *Perkinsus* spp. (Reece et al. 1997), *Giardia* spp. (Graczyk et al. 1999), *Marteilia* spp. (Le Roux et al. 1999) and *Criptosporidium* spp. (Gomez-Bautista et al. 2000). More complex methods, which use PCR as part of an integrated approach, have been also used to detect shellfish pathogens, e.g. RFLP (Buchrieser et al. 1995, Gomez-Bautista et al. 2000, Hine et al. 2001, Le Chevalier et al. 2003,

Abollo et al. 2006), Enzyme Linked Immuno Sorbent Assay (ELISA; Gonzalez et al. 1999, Schwab et al. 2001, Elandalloussi et al. 2004), or direct improvements of PCR technique itself, like multiplexing (Shangkuan et al. 1995, Brasher et al. 1998, Penna et al. 2001) and real-time PCR (Blackstone et al. 2003, Campbell and Wright 2003, Audemard et al. 2006). By allowing the detection of more than one species/strain per PCR (multiplex PCR), or the precise quantification of an infection (real-time PCR), diagnostic tools based on molecular techniques have become very popular. Recently, a multiplex real-time PCR method has been developed to simultaneously detect and quantify infection by multiple pathogens/strains (Panicker et al. 2004, Nordstrom et al. 2007). The development of high-throughput methods will allow in a near future rapid and simultaneous diagnostics of pathogens in a large number of samples.

7.6.4.2 Molecular Immunity of Bivalves

Bivalves inhabit environments where they must protect themselves from an array of commensal pathogenic and parasitic organisms. Bivalves, like other invertebrates, are deprived of an adaptative immune system (Zinkernagel et al. 1996) and fight pathogen aggression through an innate immune response (Bachère et al. 2004). Internal defence mechanisms can be split into cell-mediated and humoral mechanisms and it has become increasingly apparent that both are interrelated and closely associated with hemocytes, the main immune competent cells (Cheng 1981, Hine 1999). In order to extend our knowledge about the bivalve immune response, genomic approaches have been developed and immune-related genes have been characterized in several species (Gueguen et al. 2003, Tanguy et al. 2008).

Host defence response can be divided into three main steps: (i) pathogen recognition, (ii) activation of signal transduction, which induces the innate response, and (iii) initiation of effector production mechanisms. The recognition phase is based on the ability to discriminate between innate and exogenous agents through receptor systems and molecules which have the ability to recognize invariant Pathogen-Associated Molecular Patterns (PAMPs; Janeway and Medzhitov 2002). PAMPs are exclusive to microbes and are not produced by hosts, emerging as indispensable for microbial fitness (Nürnberger et al. 2004). Nevertheless, activation of defence mechanisms in oyster species could be triggered not only by PAMPs but also by damage-associated molecular patterns (DAMPs), as defined in the Matzinger damage model (Janeway and Medzhitov 2002, Matzinger 2002, Montagnani et al. 2007). Host molecules recognizing PAMPs are called Pattern Recognition Molecules (PRM) or Pattern Recognition Receptors (PRR; Medzhitov and Janeway 2002, Janeway and Medzhitov 2002). In bivalves, some PRRs were already identified, including: peptidoglycan recognition protein (PGRP; Su et al. 2007, Ni et al. 2007, Itoh and Takahashi 2008), several carbohydrate-binding lectins (Kang et al. 2006, Wang et al. 2007a, Yamaura et al. 2008), chitin-binding lectin (Badariotti et al. 2007), LPS-binding proteins (Gonzalez et al. 2005a, 2007, Ni et al. 2007, Bettencourt et al. 2007), receptors like Toll (Tanguy et al. 2004; Qiu et al.

2007a) and C1q-domain-containing proteins (Zhang et al. 2008a). The interactions between PRRs and PAMPs trigger the defence mechanisms.

In the Pacific oyster *Crassostrea gigas*, the characterization of six genes related to the Rel/NF-κB pathway support the concept of a conserved signalling pathway (Gueguen et al. 2003, Escoubas et al. 1999, Montagnani et al. 2004, 2008). In others bivalves, components like the Toll receptor, MyD88 and Rel were also characterized (Tanguy et al. 2004, Qiu et al. 2007a, b, Wu et al. 2007, Bettencourt et al. 2007). The homology between Rel/NF-κB pathways in bivalves and insects suggest a role for this pathway in the regulation of genes involved in innate defence (Lemaitre et al. 1995, Silverman and Maniatis 2001). Similarly, TGF-β or TGF-beta pathway could be involved in the activation of inducible defence systems (Lelong et al. 2007). Finally, the third step of the immune response leads to the expression of effectors characterized by an antimicrobial activity. These molecules are mainly produced by the immune-competent cells, the hemocytes, but also by some epithelial tissues (e.g. gills and mantle), which constitute, in bivalves, the first line of defence. Among these effectors, inhibitors of proteases and antimicrobial peptides have been studied. The inhibitors are known to target microorganism proteases to prevent host infection (Labreuche et al. 2006a,b). In bivalves, several of these molecules have been identified such as α2-macroglobuline (Gueguen et al. 2003, Ma et al. 2005), serine proteases inhibitor (serpin; Gueguen et al. 2003, Tanguy et al. 2004) and metalloprotease inhibitor (TIMP; Montagnani et al. 2001). The first antimicrobial peptides characterized were found in *Mytilus* (Charlet et al. 1996, Hubert et al. 1996) and include four families of peptides: defensin (Hubert et al. 1996, Mitta et al. 1999b), myticin (Mitta et al. 1999a), mytilin (Mitta et al. 2000b) and mytimicin (Mitta et al. 2000a). More recently, antimicrobial peptides were also isolated from other bivalves (Seo et al. 2005, Gueguen et al. 2006, Zhao et al. 2007, Gestal et al. 2007, Bettencourt et al. 2007) supporting the concept that these effectors are present in all phyla of the living kingdom. The amplification of immune effector production in response to infection is likely to be related to transcriptional or post-transcriptional regulation but also to an activation of haematopoiesis, which increases the number of hemocytes (Tirape et al. 2007).

7.6.4.3 Immune Response to *Perkinsus* Infection

Molecular mechanisms involved in bivalve–Perkinsus interaction remain largely unknown. A well-characterised mechanism in bivalves is the one triggered by lectins. Bivalves rely on lectins to recognize infectious agents and trigger resistance mechanisms to prevent invasion. Because this recognition has proven to be highly specific it has been used to monitor *Perkinsus* infection (Kim et al. 2006, 2008). For example, *Perkinsus olseni* (Bulgakov et al. 2004, Kang et al. 2006, Kim et al. 2006, 2008) and *Perkinsus marinus* (Gauthier et al. 2004, Tasumi and Vasta 2007) are not recognized by the same lectins in clam and oyster, respectively. A new galectin was also recently shown to recognize *P. marinus* in particular trophozoites, the virulent stage of the parasite (Tasumi and Vasta 2007).

Extracellular proteases, in particular serine proteases, play a major role in *Perkinsus* pathogenicity and virulence (La Peyre et al. 1996, Faisal et al. 1999, Tall et al. 1999). The presence of protease inhibitors in bivalve plasma have been proposed as a possible mechanism of defense against *Perkinsus* (Xue et al. 2006). Thus, inhibitory activity of oyster plasma was weaker in *C. virginica* than in *C. gigas*, which is also less susceptible to *P. marinus* infection (Faisal et al. 1998). Furthermore, a negative correlation was reported between disease intensity and protease inhibitory activity (Oliver et al. 2000).

Another known defense mechanism of shellfish to *Perkinsus* is the encapsulation of parasite cells (Navas et al. 1992, Montes et al. 1995b, Sagristà et al. 1995; Fig. 7.3). This mechanism underlies the inflammatory response described in clam species such as *R. decussatus* and *R. philippinarum*, where *P. olseni* induces a specific cellular response (Montes et al. 1995a, b) during which granulocytes are recruited and infiltrate into infected tissues. They then synthesize a slightly glycosylated polypeptide released in a polarized manner and organized as a capsule around *Perkinsus* trophozoites. This peptide is not expressed in *Perkinsus*-free clams or upon exposure to other micro-organisms such as bacteria or algae (Montes et al. 1995b). Although the percentage of dead trophozoites caused by this process is quite low (Montes et al. 1996) due to *Perkinsus* cell wall resistance to proteolysis, encapsulation can effectively block trophozoite dissemination (Montes et al. 1995a, Rodriguez and Navas 1995). Nevertheless, this inflammatory reaction can also destroy the blood sinuses causing host death (Montes et al. 1995a).

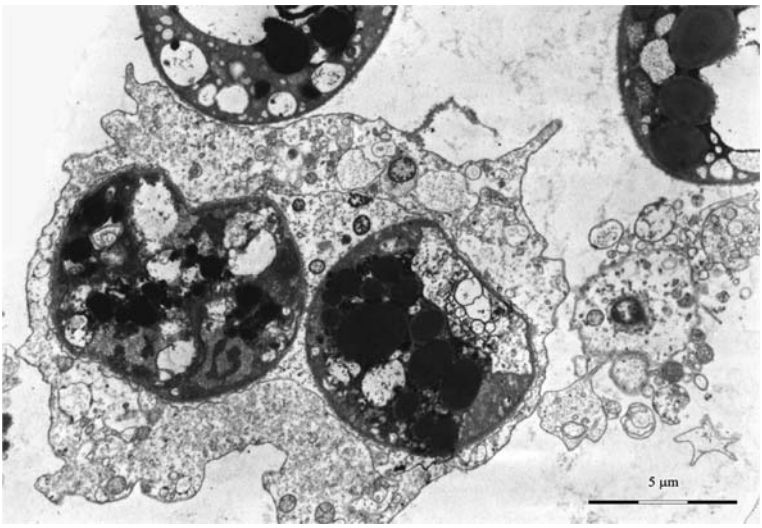

Fig. 7.3 *Perkinsus olseni* trophozoites upon phagocytosis by a clam (*Ruditapes decussatus*) hemocyte. (Figure credit: R. Leite)

7.6.4.4 Immune Response to *Vibrio* Infection

As filter feeders, bivalves are constantly exposed to various pathogenic and/or opportunistic bacteria naturally present in the microflora of coastal environments. Among these bacteria, the Vibrionaceae are the most prevalent in marine environment (Potasman et al. 2002). Reported as commensal bacteria, vibrios are also considered to be opportunistic pathogens associated with mortalities of bivalves, particularly *Crassostrea gigas* (Paillard et al. 2004). Bivalve hemocytes are equipped with both oxidative (Bramble and Anderson 1997, Bachère et al. 2004) and non-oxidative killing systems related to activities of lysosomal enzymes (Hine 1999). Pathogenic *Vibrio* species have the ability to stimulate the oxidative burst in hemocytes of *Crassostrea gigas* (Lambert et al. 2003, Labreuche et al. 2006a,b). Moreover, the bivalve hemocytes are mobile, have a high clumping potential, show chemiotactic activity (Prieur et al. 1990, Canesi et al. 2002) and develop spontaneously cytoplasmic extensions (pseudopods) facilitating adhesion. Hemocytes appear to have a different migration pattern according to bacteria and bivalve species (Howland and Cheng 1982, Kumazawa and Morimoto 1992, Fawcett and Tripp 1994). Authors demonstrated that *Vibrio* induced loss of pseudopods and cell rounding in *Mytilus* (Nottage and Birkbeck 1990, Lane and Birkbeck 1999) and reduced the adhesive capacities of clam hemocytes (Choquet et al. 2003). Similarly, extracellular products of *Vibrio* affected phagocytosis and adhesion of hemocytes in *C. gigas* (Labreuche et al. 2006a). Finally, lysozyme activity appears to be higher in *Vibrio*-challenged clams (Allam et al. 2006). Thus, factors such as bacterial surface ligands and soluble hemolymph components, as well as the ability of bacteria to activate distinct signalling pathways involved in hemocyte response, are reported to be responsible for the persistence of *Vibrio* in marine bivalves (Pruzzo et al. 2005).

Little is known about molecular mechanisms involved in bivalve-*Vibrio* interactions and associated with the immune response. An EST library was recently generated from oyster hemocytes challenged with pathogenic vibrios (Gueguen et al. 2003). Fifty-five sequences out of 1,142 ESTs analysed were classified as immune genes. Among these ESTs, the tissue inhibitor of metalloproteinase (*Cg-Timp*) was found to be highly represented. This gene is expressed specifically in hemocytes (Montagnani et al. 2001) and *Cg-Timp* mRNA accumulation could be induced by secretory/excretory molecules produced by *Vibrio* (Montagnani et al. 2007). In addition, serine protease inhibitor and serine protease were inducible by *Vibrio anguillarum* in scallop (Zhu et al. 2006, 2007). From this library were also identified four cDNAs homologous to molecules of the Rel/NF-κB signal transduction pathway in addition to the two previously characterized genes, *oIKK* and *Cg-Rel* (Escoubas et al. 1999, Montagnani et al. 2004). *Cg-MyD88*, which acts as an important adapter in this pathway, is up regulated after a bacterial challenge (Tirape et al. 2007) whereas *oIKK* and *Cg-Rel* are not affected (Escoubas et al. 1999, Montagnani et al. 2004). Moreover, a Toll receptor gene identified in scallop is up regulated with the treatment of LPS, suggesting that this pathway is involved in immune response to *Vibrio* (Qiu et al. 2007a). LPS binding proteins could play an important role in activation of the immune system. In oyster, a BPI protein (*Cg*-BPI)

was identified (Gonzalez et al. 2007a). The *Cg*-BPI expression is constitutive in the epithelial cells of most tissues and induced in hemocytes after bacterial challenge in adult oyster (Gonzalez et al. 2007) and during development (Tirape et al. 2007). In addition, after infection of oyster with the pathogen *Vibrio aesturianus*, the expression of the antioxidant enzyme *Cg*-SOD, specifically produced in hemocytes (Gonzalez et al. 2005), was strongly decreased in parallel to an increase in Reactive Oxygen Species (ROS) production (Labreuche et al. 2006b). These results are in contrast with other works where bacteria were reported to cause a reduction of ROS production by hemocytes (Lambert et al. 2003). Labreuche et al. (2006b) suggested that these events could lead to an oxidative stress and allow *Vibrio* to escape host cellular responses. In scallop, a member of the catalase family, which includes enzymes involved in eliminating ROS by degradation of hydrogen peroxide, was characterized and found to be up-regulated following challenge with *Vibrio* (Li et al. 2008). Concerning the antimicrobial peptides, modulations of their expression were reported after bacterial challenges (Mitta et al. 1999a, Mitta et al. 2000a,b, Zhao et al. 2007; Gonzalez et al. 2007, Gestal et al. 2007) suggesting a possible action of peptides against *Vibrio*. Thus, the specific role of effectors in the immune response to *Vibrio* remains to be elucidated. It will be very interesting to understand the role of inducible and constitutive genes in the immune system and to determine how bivalves discriminate between pathogen bacteria and various commensal microflora.

7.6.4.5 Status of Transcriptomic Tools

Until recently, most studies aiming at elucidating bivalve-parasite interaction were done using the Pacific oyster. Nowadays, cDNA libraries, EST collections and genomic sequencing of a wide range of bivalve species are becoming common practices in most laboratories and research projects, thus increasing genomic resources worldwide. Major achievements in bivalve genomics resulted from joint research performed within the scope of large consortiums like Marine Genomics Europe (www.marine-genomics-europe.org) or American Marine Genomics consortium (www.marinegenomics.org), whose major purposes were the massive sequencing of EST and BAC libraries in order to identify candidate genes with putative functions in cellular and biochemical processes. Using this approach, hemocyte genes related with immune defense of bacteria-challenged Pacific oysters (Gueguen et al. 2003) and *Perkinsus*-challenged Manila clam (Kang et al. 2006) have been identified. The end of Marine Genomics Europe project in 2008 left a huge legacy in terms of bivalve EST collections, including ESTs from *C. gigas*, *R. decussatus*, *R. philippinarum*, *Mytilus edulis*, and *Bathymodiolus azoricus* (Tanguy et al. 2008). As of August 2008, EST collections from 19 different species have been posted in public databases (www.ncbi.nlm.nih.gov). The most represented species is *C. gigas* with 29,018 ESTs available, followed by *Mytilus californianus* and *C. virginica* with 23,871 and 14,560 ESTs, respectively. Hence, the most represented organisms are oysters, mussels, clams and scallops, all of which are farmed species. In parallel, ESTs from 2 freshwater mussels *Dreissena rostriformis bugensis* and *Dreissena polymorpha*, with no commercial value but becoming increasingly problematic invasive species in many different countries, were collected in order to attempt to

unravel useful information suitable to develop appropriate solutions. EST projects can result not only in gene annotation, but also in the establishment of physical linkage and comparative mapping, analysis of alternative splicing and gene duplication, and microarray development. The development of cDNA microarrays covering genes from *C. gigas* and *C. virginica* exposed to *Perkinsus* and *Vibrio* allowed the characterization of differentially expressed genes in a host-pathogen condition (Jenny et al. 2006). Also within the framework of the MGE network, microarrays containing genes from oysters (*C. gigas*), mussels (*Mytilus* spp.) and clams (*Ruditapes* spp.) are being developed.

In parallel, subtractive libraries provide some highly relevant information about genes differentially expressed during host-parasite interactions. This technique (Diatchenko et al. 1999) offers a powerful and reliable evaluation of both sides in the pathogenic interaction, allowing a meticulous characterization of gene regulation either in the host or in the parasite. With the help of this technique, it has been possible to identify sets of genes differentially expressed in clams exposed to *Vibrio* (Gestal et al. 2007), in oysters exposed to *Perkinsus* (Tanguy et al. 2004) and *Vibrio* (Jenny et al. 2006), and in *Perkinsus* exposed to clam hemocytes (Ascenso et al. 2007). In addition to the techniques described above, which allow the gathering of consistent and abundant information on the characterization of host/parasite transcriptomes, many other reliable tools may be accessed, such as pyrosequencing, SAGE and MPSS, providing strong prospects in a near future.

7.6.4.6 Conclusions

Genomics, in combination with other techniques, can provide powerful approaches and tools towards the understanding of host-parasite interactions and the control of diseases affecting bivalves. In addition to strain selection programs based on host molecular markers, e.g. Pacific oyster selection programs to produce *Haplosporidium nelsoni* resistant strains (Ford and Haskin 1987), there is currently a worldwide effort towards the sequencing of host and pathogen genomes to understand defense and infection mechanisms, respectively. Even though important tools, such as bivalve cell systems, massive high-throughput genomics and microarray technology, are still missing, available data already provide important clues towards the understanding of the immune system and physiological processes related to host defense.

7.7 Genomic Variation, Stock Structure, Adaptation and Traceability in Natural Fish Populations

7.7.1 The Major Issues

The diversity, structure, evolution and functioning of marine populations is determined by the environment in which they live and their individual biological characteristics. The oceanic environment is characterised by its three dimensions, spatial continuity throughout a high-density medium and relatively moderate variation of

the environmental variables. Hence the hydrodynamics of the habitat in which marine organisms live shapes the ecosystem. But the environment in which marine organisms live is also shaped by the ecosystem itself, which is defined by the community of organisms, their composition, interactions and dynamics (Scheffer et al. 2005). Predator-prey and socio-behavioural interactions, host-pathogen dynamics and random forces influence the inhabiting species and populations and hence the full ecosystem make-up. To some extent the influence is exerted through the phenotype of these organisms, where biological traits of importance are their life-history, ecological niche and evolution.

Population genomics is the investigation of the genetic basis of adaptation and speciation at the genome scale (Bonin 2008), or loosely defined it is simply population genetics on a large scale. Its foundation can be traced back to human genetics by separating locus-specific effects from genome-wide effects. Since that time, genomics has been shown to hold much promise for generating knowledge from the poorly and difficult to study species of the open ocean (Derelle et al. 2006, Yooseph et al. 2007).

We will start this section by taking a closer look at some of the key issues on intraspecific evolution of fishes.

1. Since speciation is an ongoing and gradual process, the distinction between species, subspecies and population is not always unambiguous. Hybridisation between taxa is a common feature across the tree of life. This issue is further complicated by the presence of numerous cryptic taxa across the tree of life, including the fishes (Colborn et al. 2001; see Chapter 3 on Barcoding).
2. Fish are free-living organisms and have as a rule free-living larvae (and often free-floating eggs). The small dimension of eggs and larvae makes that they are associated with the plankton. Juveniles and adults in contrast, actively control their movements. However, the behaviour of parents and larvae makes them disperse less than expected, and show some level of retention (Jones et al. 2005). This encourages a pattern of genetic structuring, albeit subtle (see further). Frontal systems, large-scale eddies, depth, coastal topography and bottom type shape these patterns.
3. Many oceanic fish taxa are characterised by huge census numbers (Palumbi 2004). However, recruitment is such that numbers may fluctuate between cohorts (see further). If the breeding cycle is limited to just a few years, large differences in adult numbers may be observed. Variable recruitment results in an effective population size that is several orders of magnitude smaller than the census size (Hauser et al. 2002). It might make these populations less robust to demographic changes than thought previously. “Sweepstake” survival of entire family groups at the early stages of the life history leaves an imprint on the next cohort (Hedgecock 1994). This fits with the classical “Match-Mismatch” hypothesis of Hjort (1914) (reformulated by Cushing 1969), which states that survival of a cohort depends on the overlap between prey (the phyto- and zooplankton) and predator (fish larva), or alternatively the fish larvae as prey for higher level predators.

4. Eggs are produced in large numbers and fertilised externally. The older (and larger) the fish, the more eggs (often of a better quality) are produced (Marteinsdottir and Begg 2002). The most common breeding strategy is mass spawning: some species brood their eggs, such as nest guarding gobies, pouch brooding (e.g. pipefishes) and live-bearers (e.g. *Sebastes* spp.). Despite the huge numbers of progeny and the huge potential for population growth, adult population sizes remain relatively constant. This is largely due to the high selection pressure at the early life stages. Eggs, larvae and young post-larvae are massively preyed upon and even cannibalised. But infections, failed fertilisation, oceanic drift to suboptimal "sink" habitats, maternal effects and genomic incompatibilities also cause extensive mortalities (King 1995). Typically, the contribution of spawners to the next generation is not equally distributed between individuals and genders. Only a few parents contribute significantly to the progeny, while most do not contribute at all.
5. The last point to mention is that given the high level of connectivity, genetic population structure might be real but is also weak, usually an order of magnitude smaller than in continental systems (Ward et al. 1994, Cowen et al. 2006). "Metapopulations" conceptualise the patterns and dynamics of fish communities (Hanski and Gaggiotti 2004). Patterns and dynamics are strongly determined by the landscape in which they spend their early life-history, feed as (sub)adults and mate (Jørgensen et al. 2005). An alternative perspective is the "member-vagrant" concept (Sinclair 1988), where propagules may or may not successfully colonise new habitats.

From the above-mentioned characteristics of marine fish and their habitat the key questions from a genetic perspective can be reformulated as:

- What is the historical and contemporary genetic diversity, its causes and dynamics?
- What are the historical and contemporaneous genetic patterns (neutral evolution), their causes and dynamics?
- What are the historical and contemporary patterns of adaptive evolution, be it either natural or anthropogenic in origin? What are their causes and dynamics?

Some aspects of population genomics have great relevance for the goods and services they deliver to society.

- The tracing of populations is a cornerstone for ecosystem-based fisheries and conservation management.
- The tracing of food stuffs is a commonly applied tool for law enforcement and quality assurance
- The tracing of populations and individuals is of fundamental importance to allied research disciplines (ecology, physiology and behaviour).
- Diversity at the population level has helped to discover and understand novel substances for biotechnological applications.

The field of marine population genetics has received a major impetus since the end of the last millennium. Initially restricted to a small number of viruses, bacteria, model species and man, it is now being extended to an ever growing group of taxa or even communities (metagenomics of microbial and picoplankton communities) (see Chapter 2 on Metagenomics). Fish genomics was initially limited to the study of developmental biology and genetic architecture of model fish such as Japanese pufferfish (*Takifugu rubripes*) and zebrafish (*Danio rerio*). Recently, the number of fishes for which genome wide sequence information is available has risen steadily (Cossins and Crawford 2005, Kocher and Kole 2008). Fish population genomics was applied first in the freshwater environment (e.g. Campbell and Bernatchez 2004) and later on in the oceans (e.g. Larsen et al. 2007). However, relatively few studies deal with marine population genomics, which is in line with the poor overall focus on marine biodiversity (Hendriks and Duarte 2008). The following paragraphs therefore aim at integrating key studies in the field of population genomics to gain an understanding of state-of-the-art and future direction of marine fish population genomics.

7.7.2 State-of the Art in the Population Genomics of Fishes

7.7.2.1 Identifying Population Structure and Dynamics

In order to manage marine fisheries resources sustainably, one of the most pertinent questions to answer is how genetic and genomic variation in exploited fish is distributed in time and space (Carvalho and Hauser 1998). By conducting genetic analyses throughout the targeted species distribution, it is possible to determine whether it consists of one large genetically homogenous unit where individuals are mating randomly, or a smaller or larger number of semi-independent populations or "stocks", with various degrees of reproductive isolation, and their spatial dynamics (e.g. Sinclair 1988, Hanski and Gilpin 1997, Waples and Gaggiotti 2006). At the same time analysis of historical tissue collections available in museums and fisheries institutions all over the world (Nielsen and Hansen 2008) can provide inferences on the temporal dynamics of the identified population structure and the demographic trajectories of individual populations.

Failing to recognize the evolutionary relationships among spawning groups within a species may have dire consequences for fisheries management. It may lead to the over-exploitation of small, isolated, slow growing populations and ultimately to their extirpation (Dulvy et al. 2003) and associated loss of intraspecific biodiversity. Accordingly, management and conservation of fisheries resources has to be population based. Here genetics has traditionally been playing an essential role in delineating population structure to assist fisheries managers.

Population genetic analyses of fish have been conducted for more than five decades. In general the number of populations and levels of genetic differentiation among populations is low for marine compared to freshwater and anadromous fish (Ward et al. 1994). This may be attributed to higher levels of migration and

associated exchange of genetic material (gene flow) among populations. Population genetic theory predicts that the level of genetic divergence among populations is positively correlated with the degree of reproductive isolation. For most species the "genetic distance" (F_{st}[1]) is below 3%. As such, subtle genomic variation is the rule more than the exception (e.g. European eel: Maes et al. 2006; bluefin tuna Rooker et al. 2007; Atlantic herring: Ruzzante et al. 2006), although cases of relatively strong genetic differentiation have been documented (e.g. in Atlantic cod, Nielsen et al. 2003; Atlantic killifish, Duvernell et al. 2008). Accordingly, population genetics of marine fish has constantly been hovering on the brink of detection of a true signal of population differentiation (Waples 1998). The statistical power for inferring population structure in marine fish can be improved by using large samples of individuals and checking for temporal stability. In many cases, however, strong inferences of the mere occurrence of population structure or the relationships among putative populations cannot be made with the limited number of genetic markers commonly employed at present. For example, Koskinen et al. (2004) found that the median number of genetic markers (nuclear microsatellites or SSR) was only six! The theoretical recommended number for unambiguous determination of relationships and distance among populations is at least 30 but could reach several hundreds (see Takezaki and Nei 1996, Pollock et al. 1998). In practice, Koskinen et al. (2004) showed that increasing the number of microsatellite genetic markers from 6 to 17 provided a massive improvement of the power for determining the correct genetic relationships. Obviously, a switch to genomic methods, where hundreds of markers are commonly employed, might revolutionize the field of research. An additional problem relates to determining the evolutionary properties of the genetic markers employed. State of the art genetic markers such as microsatellites are a priori assumed to represent neutral genomic variation, i.e. not subject to direct or hitchhiking selection. Recent cases have shown that microsatellites subject to selection are not uncommon in population genetic studies of fish (Nielsen et al. 2006, Larsson et al. 2007). Accordingly, demographic inferences on migration rates and population sizes based on "neutral" population genetic theory are expected to be biased. In contrast, scanning the genome using multiple genetic markers spread throughout the genome would allow a statistical evaluation of the evolutionary dynamics of individual marker loci (Beaumont and Nichols 1996, Schlötterer 2002) enabling the establishment of a neutral baseline applicable for providing demographic inferences (Luikart et al. 2003). Markers subject to selection on the other hand could be used for inferring local adaptations to specific environmental conditions experienced by local populations (see next paragraph).

Although we are aware of many ongoing studies using large genomic datasets for inferring population structure of marine fish, the number of published studies is relatively scarce. As in many other fields of fisheries science, salmonid research is

[1] F_{st} is the proportion of the total genetic variance which can be ascribed to population differences. F_{st} can attain values between 1, when populations are completely isolated, and 0 when there is no apparent reproductive isolation among samples. Intermediate values represent different levels of migration among populations.

leading the way due to the high value of commercial and recreational fisheries but probably most importantly due to their additional importance for aquaculture breeding and associated development of genomic resources (see Section 7.2 on Genomic tools and resources). For example, Hayes et al. (2007) identified 2,507 SNPs from the alignment of Atlantic salmon ESTs, which can be readily used for population genomic studies.

An example of improving the performance of the number of genomic markers in salmonids can be found in Smith et al. (2007). They compared estimates of population differentiation among 16 collections of Chinook salmon (*Oncorhynchus tshawytscha*) using traditional allozyme loci (22 loci), small tandem repeats/microsatellites (nine loci) and 39 newly developed nuclear SNP loci. The larger number of markers allowed testing for "outlier loci" (Beaumont and Nichols 1996) identifying five SNPs likely to be influenced by selection. Exclusion of these loci, and three other, which had previously been suspected to be under selection, provided a reduction in the estimates of differentiation and an associated increase in the estimated migration among putative populations. Similarly, slight differences in relationships among populations were apparent when comparing results from different marker classes, where SNPs generally, even after exclusion of outliers, displayed higher levels of genetic differentiation. In conclusion, better estimates of population structure could be achieved for this species by increasing the number of markers to, in genomic terms, a relatively modest level.

Also for "classical" marine fish, i.e. widespread species with large population sizes, high fecundity and pelagic eggs and larvae, examples of population genomics has started to emerge. Moen et al. (2008b) identified and characterised 318 SNPs in Atlantic cod from alignment of EST sequences. Overall their results demonstrate and substantiate that Norwegian Coastal Cod and North-East Arctic cod represent two highly isolated populations, which differ not only in neutral genes due to the lack of migration, but have also diverged on a genomic level by adaptation to the differences in local environment that they experience.

As can be seen from the examples above, SNPs are becoming the marker class of choice for population genomic studies of marine fish. Although large numbers of microsatellites are isolated (e.g. Karlsson et al. 2008) the attention of genomic resource development for population genetics is turning more towards SNPs. They possess several attractive features for ecology, evolution and conservation in general (see Morin et al. 2003) but for fisheries management in particular. First of all management of most marine fish species are transnational and management advice is delivered by international bodies such as ICES (International Council for Exploration of the Sea). Accordingly, there is a need for easy calibration of markers among national fisheries labs to provide concerted data sets and help with the validation. Since SNP genotyping is qualitative (i.e. different bases represent different alleles) compared to microsatellite scoring, which is quantitative (i.e. fragment length is translated into alleles with different numbers of base-pairs) calibration is much facilitated. On the down side, SNPs are less variable than microsatellites (fewer alleles, commonly two) requiring 4–12 times more loci than microsatellites to attain the same statistical power (Liu et al. 2005a, Smith and Seeb 2008).

But cases have been reported where eight microsatellites have the same information content as 18 SNPs (Artamonova 2007). This can, however, be compensated for by the high and widespread genomic abundance facilitating huge numbers of markers with good genomic coverage to be developed. Slightly more problematic is the "ascertainment bias" (see Clark et al. 2005 and references therein), which is the selection of loci from an unrepresentative sample of individuals or the use of a method which provides a biased sample of loci. Ascertainment bias may not represent a major obstacle for SNP application in marine fish, as it is generally expected to be particularly serious for highly structured species (see Chapter 3).

The development of new and more numerous genetic markers is not the only way that the genomic revolution can contribute to our understanding of population structure in marine fish. Other methods such as the analysis of population differences in gene expression can provide important pieces to the puzzle (Cossins and Crawford 2005). Microarrays for transcriptional analysis are an example. They have been developed for a number of marine fish, primarily for model or aquaculture species (see Wenne et al. 2007 and references therein) and have also been used to investigate gene expression in natural populations (e.g. killifish, Oleksiak et al. 2002). This approach may provide new insights into the population structure of managed species. Larsen et al. (2007) used a microarray to investigate potential differences in gene expression between individuals from the North Sea and the Baltic Sea, two different physical and biological environments, however with low levels of genetic differentiation. A high number of genes were differentially expressed between fish from the two populations experiencing similar conditions, thus strongly suggesting population subdivision despite low levels of neutral genetic divergence. The results suggest that population based management of marine fish should be enforced even when migration rates appear to be relatively high.

7.7.2.2 Selection and Adaptation in Natural and Exploited Populations

The description of the demography of wild populations has been the primary focus of population and conservation genetics for the last few decades (Beaumont 2005). Although the first genetic studies of marine fish were using markers such as haemoglobin (Sick 1965) shown to be under selection, the quest for demographic information to feed into the current management system of fisheries resources (see section above), has also meant that (presumed) neutral markers are the choice for marine fish species. However, the focus of the whole scientific field is shifting from studying neutral genetic variation alone to include inferences from gene loci known – or suspected – to be under selection. The reasons for this change are many. Generally, it can be said that adaptive variation represents the other side of the evolutionary coin. That is, in order to understand the evolution of a species it is crucial to know both how the genetic diversity is affected by demographic processes (migration and genetic drift), and how traits and the underlying genes are subject to selection, in turn improving the individual′s chances of survival and reproduction. Much has been learned from directly studying trait variation in natural populations in the wild or under controlled conditions. However, many such studies

have been confounded by the difficulty of disentangling environmental from genetic effects (Endler 1986). Accordingly, there has been a great interest in identifying the footprints of selection at the molecular level. Three issues come to mind, namely genomic inferences on (1) the genetic basis and architecture of local adaptations; i.e. the genomic basis of why resident individuals commonly have higher fitness than individuals from other environments (Kawecki and Ebert 2004), (2) demonstrating and understanding the genomic basis of evolution in response to the on-going global change (Reusch and Wood 2007, Gienapp et al. 2008), but also studies of (3) the evolutionary impact of selective harvesting (Coltman 2008, Allendorf et al. 2008). All three issues are highly relevant in relation to marine fish evolution as well as for the management of genetic resources in marine fish. First of all, if marine fish are locally adapted then conservation of the local populations is essential, not only to maintain adaptive genetic diversity within the species, but also to assure the survival of the species in particular geographical areas/habitats and to maximise productivity. Immigrant non-adapted fish would have lower survival and/or growth; reproduction would be strongly impeded. Secondly, global change is expected to alter the distribution and abundance of many marine species and related fisheries (Roessig et al. 2004, Dulvy et al. 2008). Little is known whether the fish populations will be able to adapt genetically over this – in an evolutionary context – rather short period with rapidly changing environmental variables such as temperature (Davis et al. 2005). Finally, fisheries induced evolution has been inferred based on changes in age and size at maturation for fish (Jørgensen et al. 2008), albeit without direct genetic evidence of change at the molecular level. If fisheries are altering important life-history traits, fitness under natural conditions as well as population productivity may be strongly reduced.

A number of studies have demonstrated environmental selection on single protein markers (e.g. Sick 1965, Christiansen and Frydenberg 1974) or DNA markers (e.g. Case et al. 2005, Hemmer-Hansen et al. 2007) suggesting genetic adaptations to the local environment. Although such studies have provided valuable insights into the evolutionary processes in marine fish, they, nevertheless, only infer selection of a particular gene, i.e. not at the genomic level. Instead general information is warranted about the proportion and types of genes subject to selection, and the genetic architecture (number, function and interplay among genes) of phenotypic traits. Therefore, genomic approaches allowing simultaneous assessment of many genes potentially subject to selection would provide a major leap forward in relation to understanding adaptation to natural and anthropogenic drivers of evolution.

Selection in space and time can subsequently be identified by a number of different approaches. The “candidate gene approach” exclusively investigates variation in genes thought to play a major role for particular adaptive traits of interest. Common relatively long genomic sequences are generated from individuals from different populations and subjected to statistical methods which enable the identification of molecular footprints of selection (Guinand et al. 2004). Such a directed approach has many advantages when investigating particular traits with very good “candidates”. A historical example is the characterisation of the natural variation and physiological

impact of the Leucine Aminopeptidase (LAP) gene involved in the osmoregulation of bivalves (e.g. Milkman and Koehn 1977, Koehn and Immermann 1981). However, the number of loci which can be assessed is relatively limited and many candidates may turn out not to be subject to selection and therefore unsuitable. A good example of the inherent problems of using a candidate gene approach can be found in Ryynanen and Primmer (2004), who examined variation in and around the Growth Hormone (*GH*) gene in Atlantic salmon. Despite extensive sequence analysis of one of the most prominent candidates for growth differences in fish (De-Santis and Jerry 2007), they were not able to find any genetic variation in exons, which could be subject to differential selection among populations. Therefore "genome scans" (e.g. Storz 2005) of large numbers of gene associated markers such as SNPs and SSRs has become the preferred genomic tool for identification of adaptive divergence among populations (Beaumont 2005). Loci that display elevated levels of genetic divergence among populations (high F_{st}) are most likely subject to selection. The genes included can be completely random, thus providing evidence of the proportion and types of genes involved in local adaptation. For example, Lemaire et al. (2000) detected in European sea bass caught in inshore and offshore habitats, six loci (all allozymes), among 6 microsatellite and 18 allozymes, with above average F_{st} values. Alternatively, the scan can be "directed" by combining the genome scan with a candidate gene approach, preferentially including genes with known function and suspected to be involved in adaptation. This approach allows the information content of the scan to be maximised and inferences on the genetic architecture of adaptive traits to be made. Mäkinen et al. (2008) conducted a genome scan among seven marine and freshwater populations of the three-spined stickleback (*Gasterosteus aculeatus*). They found strong signatures of directional selection for two of the EST derived microsatellites and the *Eda* associated indels, thus strongly suggesting local adaptation. Likewise, Vasemägi et al. (2005) conducted a genome scan among populations of adult Atlantic salmon from different habitats (i.e. freshwater, brackish and marine). They found that nine EST-associated microsatellites displayed highly divergent patterns of genetic differentiation, and were thus likely to be candidate genes for local adaptation in Atlantic salmon. Still the genome scans presented here are based on a relatively small number of gene markers and could be vastly improved by applying new large-scale sequencing methods such as "sequencing by synthesis" (e.g. 454 sequencing). Naturally, genome scans are not restricted to spatial analysis, but can also be applied on a temporal scale to study potential evolutionary change mediated by global change or selective fisheries. No large-scale temporal genome scans have been published to date. However, Nielsen et al. (2007) investigated potential temporal selection on the *Pan I* (Panthophysin) gene in Atlantic cod. By extracting DNA from up to 69 year old otoliths, they were able to compare levels of genetic differentiation at the *Pan I* locus with a suite of microsatellites and found that the temperature change in the investigated time period and area was too small to have had any profound effect on *Pan I* allele frequencies. With the increasing awareness of global change and the need to understand evolution in order to predict the distribution and abundance of marine fish in a changing world, we expect temporal genome scans to play an increasingly crucial role. We also expect that in the

future genome scans will be imbedded in a framework of landscape or "seascape" genetics (Galindo et al. 2006, Joost et al. 2007), i.e. where the patterns of selection observed are statistically tested for associations with particular environmental variables. Obvious candidate drivers of selection and local adaptation are temperature, salinity and oxygen content. However, a number of other factors such as pollution, infections and predator/prey interactions are also very likely to be driving local adaptation.

The identification of differential selection among populations does not only rely on assessment of variation in the genome, but can also be evaluated through expression phenotypes. Investigating differences among populations in gene expression, by conducting quantitative PCR or microarray studies under controlled conditions, can provide good indications of genetically based gene expression differences among populations. For example, Lucassen et al. (2006) investigated the effect of cold acclimation on RNA expression of mtDNA genes in two populations of Atlantic cod maintained at identical conditions, but experiencing different temperature conditions in their native habitat. They found clear differences among populations, which were ascribed to genetic variation at functional sites between the two populations. Finally, gene expression profiles can also be used to detect evolutionary changes in hatchery-reared fish compared to their wild conspecifics. Roberge et al. (2006) found strong differences in gene expression between wild and farmed salmon using a microarray with 3,557 genes, but patterns of parallel evolution between two farmed populations. Consequently, this highlights the need for avoiding escapes of hatchery-reared fish into the wild, since hybridisation can have large and unexpected effects on gene expression (Roberge et al. 2008).

7.7.2.3 Tracing Natural Populations for Fisheries Enforcement and Traceability

Probably the most serious obstacle for obtaining sustainable fisheries is illegal, unreported and unregulated (IUU) fisheries. Therefore, there is a huge demand for methods which can establish the species and area of origin of fish in all links of the chain, i.e. from catch fisheries or aquaculture to the plate of the consumer. This information can be used for forensic purposes, allowing genetic evidence to be presented in court-cases relating to illegal fishing and or mislabeling of fish products. A number of methods are available which can provide information on catch origin such as lipid composition (Falch et al. 2006), isotope analysis (Guelinckx et al. 2008) and elemental analysis (Campana and Thorrold 2001). As DNA is found in all cells it allows analysis from all types of tissue in fish, even partly degraded samples such as processed food (Rasmussen and Morrissey 2008) and historical samples (Nielsen and Hansen 2008). Accordingly, genetic methods are generally considered the most versatile means for classification of fish to species or local population origin (e.g. Hansen et al. 2001). Species identification is relatively straightforward commonly requiring sequencing of just one gene (see also Section 9.5.5). An example is the *COI* (Cytochrome Oxidase I) gene, which is the chosen marker for the

"Barcoding Of Life" initiative (www.fishbol.org; see Chapter 3). However, inferences on the population of origin of a sample of fish and/or individual fish rarely rely on fixed genetic difference among populations, since they are connected by gene flow. Instead they have to use statistical methods for calculation of the most likely population of origin. Most commonly the genome is used, but also the transcriptome may identify populations (Larsen et al. 2007) or separate wild from culture fish (Roberge et al. 2006). The two most widely used methods for GSI (Genetic Stock Identification) are "Individual Assignment" (IA; see reviews by Hansen et al. 2001 and Manel et al. 2005) and Mixed Stock Analysis (MSA; Pella and Masuda 2001).

The principle of IA is that an individual is assigned to the population, out of a number of potential populations of origin, where its multi-locus genotype has the highest likelihood of occurring. For example, Nielsen et al. (2001) were able to almost unambiguously assign individual cod form the North Sea, Baltic Sea and Barents Sea to their population of origin. This ability has subsequently been used in a court case, where a Danish fisherman was accused of fishing illegally in the North Sea, while he claimed that the fish were legally caught in the Baltic Sea. DNA analysis of five cod from the catch and subsequent assignment tests using 10 microsatellites revealed that all fish were assigned to the North Sea. The fisherman was subsequently convicted (E.E. Nielsen, personal communication). Likewise, Primmer et al. (2000) used assignment tests to exclude an unusually large Atlantic salmon presented at a fishing contest in Lake Saimaa (Finland) as a local fish. Its genotype was 600 times more likely to correspond to populations commonly sold on fish markets in Finland. When confronted with the evidence the angler confessed to the fraud.

In contrast to IA, MSA estimates the most likely proportions of individuals, in a potentially mixed sample, originating from each of a number of sampled base-line populations. Ruzzante et al. (2006) used MSA to estimate the most likely origin of mixed population aggregations of herring in Skagerrak. They found that the proportions of fish from the North Sea, Skagerrak and Kattegat/Western Baltic varied among juveniles and adults and summer/winter, while the results between years were remarkably stable. It is now possible to manage fisheries in the area, to avoid overfishing of vulnerable populations by targeting specific areas and time periods where thriving populations are abundant. MSA has also been employed in forensic cases, in particular for Pacific salmonids. Withler et al. (2004) reported 17 forensic cases of illegal fishing or selling of Sockeye (*Oncorhynchus nerka*) and Chinook salmon from the Fraser River, Canada. Microsatellite analysis and subsequent use of both MSA and IA for fish sampled at the harvester as well as in restaurants revealed fraud most of the time. An additional application of genetic methods for traceability and forensics is for the identification of individuals, parentage and families as used in livestock (Dalvit et al. 2007). Identification of individual fish in wild populations of marine fish is not common, but could be highly relevant for large and/or particularly valuable specimens such as tuna or whales. For example, Baker et al. (2007) assessed unique genotypes of North Pacific minke whales (*Balaenoptera acutorostrata* spp.) sampled at 12 surveys of fish markets in the Republic of South Korea

to estimate the total number of minke whales entering the trade. Using a capture-recapture model they estimated that the most likely number was approximately double of the officially reported number.

These examples are representative of the potential of genetic methods for fisheries enforcement and traceability. However, the genetic analyses are often hampered by the low levels of genetic differentiation commonly found among populations of marine fish. Accordingly, the statistical power for determination of the origin of individual or samples of fish rarely suffices for forensic purposes. Since the power for GSI is ultimately linked to the number of loci applied and the levels of genetic differentiation at marker loci (e.g. Koljonen et al. 2005 and references therein) there are large perspectives for future application of genomic methods.

7.7.2.4 Integrating Evolutionary and Ecological Functional Genomics with the Environment

Divergent selection should act to reduce gene flow and hence contribute to speciation (Coyne and Orr 2004). However, how adaptations identified at the genomic levels relate to the external phenotype in terms of life-history, behavioural and physiological traits, and likewise to the genetic architecture of adaptive trait variation remains largely unknown. Most arguments have been theoretical with surprisingly little empirical knowledge about the genetic basis of adaptation and the role of selection (Orr 2005). Evolutionary and Ecological Functional Genomics (EEFG) explores the evolutionary mechanisms that underlie ecological traits and how these traits affect evolutionary fitness (Feder and Mitchell-Olds 2003). It requires a rather challenging simultaneous use of molecular, cellular, organismal, population and ecological approaches. Initially integrated studies were limited to a few classical model species (e.g. the plant *Arabidopsis thaliana*). The biology of these taxa was sufficiently well known at the various organizational levels to allow an EEFG approach. Understandably, the study of evolution in a natural setting requires more and different types of "model" organisms. Fortunately, the number of taxa with a dense genomic background has increased. Following the ecological models of three-spined stickleback (Kingsley et al. 2004) and mummichog *Fundulus heteroclitus* (Oleksiak et al. 2002), a growing number of fishes relevant to fisheries (and aquaculture) complement the list: the salmonids rainbow trout *Oncorhynchus mykiss* (Rexroad et al. 2005), Atlantic salmon (Moen et al. 2008a), brown trout *Salmo trutta* (Gharbi et al. 2006) and whitefish *Coregonus clupeaformis* (St-Cyr et al. 2008), the perciforms European sea bass (Volckaert et al. 2008), gilthead seabream *Sparus aurata* (Sarropoulou et al. 2005b) and Nile tilapia *Oreochromis niloticus* (Cnaani and Hulata 2008), the gadid cod *Gadus morhua* (Symonds and Bowman 2007), the flatfishes European flounder *Platichthys flesus* (Williams et al. 2003), olive flounder *Paralichthys olivaceus* (Kang et al. 2008) and turbot *Scophthalmus maximus* (Bouza et al. 2007), and the channel catfish *Ictalurus punctatus* (Liu et al. 2008).

Adaptive radiation is most classically studied on external phenotypes and life-history traits, often a single trait at a time. A prime example among fishes where an impressive body of knowledge of fast evolution has accumulated is the three-spined

stickleback (McKinnon and Rundle 2002). Stickleback populations have invaded repeatedly freshwater habitats after the last ice age throughout the northern hemisphere causing rapid genetic divergence. This leads to systematic adaptations of phenotypic traits. For example, the reduction in the pelvic spine (Shapiro et al. 2006) and body armour (Peichel et al. 2001) has been attributed to the *Ptx1* and *Eda* genes respectively. Interestingly, the trait body armour has been lost in freshwater populations many times in parallel due to the sorting of a rare allele of the *Eda* gene in ancestral marine populations (Raeymaekers et al. 2007). The gene seems to give a growth advantage, but it remains unclear whether it affected additional traits or whether linked loci play a role (Barrett et al. 2008). Also mate choice (Boughman 2001), feeding (Schluter 1995) and parasite resistance (Kalbe and Kurtz 2006) have diverged rapidly and in parallel. Parasite resistance in stickleback has been attributed to a trade-off between innate immunity (*MHC* class I and II) and acquired immunity (Wegner et al. 2007). Increasingly, the significance of the gene cluster *MHC* has been acknowledged; *MHC* diversity has been linked to many fitness traits, including mate choice (Milinski et al. 2005).

Further developments of genomics and better insights into the genetic architecture of model organisms have made studies of multi-locus and multi-trait adaptations in non-model organisms feasible. The above mentioned studies by Oleksiak et al. (2002) and Larsen et al. (2007) using natural and managed populations, respectively, represent the first cases in marine fish where transcription profiling was used to discriminate among populations. There is also interest for evolutionary models of parallel adaptive differentiation in physiological and behavioral functions between species and among ecotypes. The dwarf morphotype of the whitefish *Coregonus clupeaformis* and the cisco *C. artedi* locally occupy sympatrically the same limnetic ecological niche in North American freshwater lakes. *C. clupeaformis* has apparent adaptations to the limnetic trophic niche channeled through transcriptional changes in functional genes involved in muscle contraction and energetic metabolism relative to the sympatric normal ecotype (Trudel et al. 2001). As evidenced by a transcriptome analysis of both species using a heterologous microarray, each had the same genes involved in muscle contraction and energy metabolism, but regulation was different (Derome and Bernatchez 2006).

A more detailed analysis of the mechanisms driving adaptive radiation in the dwarf limnetic and normal pelagic ecotype of the whitefish *C. clupeaformis* was involved in the analysis of physiological and behavioural traits. First, a comparative linkage analysis of the dwarf and normal ecotype showed a different genetic architecture between the two types. Rogers et al. (2007) concluded that allopatric divergence during the Pleistocene glaciations caused divergence and that subsequent ecological speciation occurred sympatrically. In a next step a QTL analysis of a F_1 backcross family combined with a genome scan of both natural ecotype pairs showed distinct differences in adaptive divergence and the role of divergent natural selection of traits. Significant QTL were associated with swimming behaviour (which is important for habitat selection and predator avoidance), growth rate, morphology (number of gill rakers and condition) and life history (more specifically the onset of maturity and fecundity). The natural sympatric pairs revealed a

signature of selection at 24 loci, of which eight loci showed a signature of selection in the mapping families (Rogers and Bernatchez 2005, 2007). From complementary transcriptome (expression) profiling (eQTL), six candidate genes related to white muscle modulating swimming activity in the dwarf ecotype of white fish and the sympatric cisco turned out to be upregulated (Derome et al. 2006, see above). Four genes were upregulated in cisco, pointing to its greater physiological potential to exploit the limnetic trophic niche. A follow-up study on the sympatric pairs of dwarf and normal whitefish in two natural lakes through transcriptome analysis of the liver showed that 6.45% of significantly transcribed genes showed regulation either in parallel or in different directions. Dwarf whitefish showed consistently significant overexpression of genes potentially associated with survival through enhanced activity. The normal ecotype showed more overexpressed genes associated with growth. These original results show a first mechanistic genomic basis for major life history trade-offs in both ecotypes. In short, enhanced survival through active swimming increases energetic costs translating into slower growth and reduced fecundity in the dwarf morph (St-Cyr et al. 2008).

7.7.3 A Vision of the Future

Although the concepts of population genetics have been with us for a long time, molecular tools for testing evolutionary hypotheses under challenging oceanic environmental conditions were largely missing. Technological development has progressed to such an extent that the genome, transcriptome and proteome have become accessible for hypothesis testing in species and populations of key ecological relevance under field and laboratory conditions alike. Therefore it is interesting to have a concise look at future developments.

Fundamental knowledge on fish and fisheries genomics is likely to benefit from the advances in genomics. The technical capacity of genomics keeps on progressing towards higher throughput at a lower cost per base pair (Gupta 2008). Hence many hypotheses of population genetics might finally be testable. If genome based monitoring is linked to the latest developments in seabed and water column based monitoring and tracking (Delaney 2007), much is to be expected from real-time genotyping. This will of course vastly increase the amount of data generated. Bioinformatics is foreseen to play an increasingly important role for data management and analysis.

Genomics holds great promise for future applications to identify management units in marine fish. Gene flow is generally high and accordingly, spatial and temporal patterns of genetic differentiation are subtle. The information signal on evolutionary separated units originating from neutral genomic variation is too subtle to understand the full picture of dispersal and connectivity among local populations. The recent focus on adaptive traits holds great promise for understanding the dynamics of population structure in space and time (e.g. Pogson and Fevolden 2003, Larsen et al. 2007). Full genomes of populations adapt to their environment,

but in the first place selection is identified at specific genomic sites, where crucial genes or gene complexes are lodged. It provides great promise for understanding the genomic imprint of man's actions, such as "fishing down the food web" (Pauly et al. 1998), for modifications to the trophic cascade (Scheffer et al. 2005), and for defining biologically meaningful management units and marine protected areas. The incorporation of genomics allows for realistic "genetic monitoring" of populations as an integral part of fisheries management.

The analysis of phenotype in the natural environment was historically limited to the morphotype and some ecophysiological traits. That picture is increasingly complemented now by a well-documented transcriptome and proteome phenotype in a number of species. This will aid to understand the functioning and evolution of fish and their communities. Accordingly, understanding the function and origin of an ecological adaptation has become feasible (Feder and Mitchell-Olds 2003). Unfortunately, fully understanding those patterns remains a tedious process. The bottleneck to fast progress is that information has to be collected from diverse fields of research and put together in a comprehensive model framework. Still, these novel insights will benefit the sustainable resource management of fisheries and the optimisation of aquaculture.

Phenotype is a major focus in aquaculture, where animals ideally feature traits of preference to the producers (e.g. low food conversion, high survival and long shelf life) and consumers (e.g. shape, colour and taste). There are two main strategies to achieve this. First, animals can be selected through the well-established and classical strategies of trait selection (most efficiently through some kind of family selection). Genetic background information on relatedness, origin and diversity can be incorporated to support breeding. Second, a genomics supported breeding strategy of marker-assisted selection may accelerate the domestication process. Here molecular genetic information closely linked to genes of interest serves as selection target. As most fish have a relatively short history of domestication, few traits have been maximised for aquaculture production. Using markers linked to well characterised QTL should speed up the process of domestication. Understanding how genes function and chromosome segments impact traits is high on the breeding agenda, although few cases have reached the demonstration level in fish breeding (A. Sonesson, personal communication).

Globalisation of the food market has raised concerns about food quality. Huge efforts are devoted to monitoring the flow of animal products, including fish and fisheries products, from "field to fork". Customers look for certification of origin and composition; traders and processors prefer guarantees of quality and ownership (see Section 7.5). It is foreseen that routine monitoring and forensic applications in fisheries will vastly expand. It involves a legal and organisational framework, including sampling, genotyping, statistical analysis and interpretation. Genomics will contribute to the increasing use of DNA based methods as it allows for the simultaneous analysis of many samples for a high number of loci, and the identification of selected loci with elevated levels of genetic differentiation. So far no genomic studies of GSI and traceability have been published for marine fish, but genomics will revolutionise the field of fish forensics in the near future.

We may conclude this overview on the population genomics of bony fishes with the observation that the understanding of evolutionary processes in natural and cultured populations is expected to increase in the near future. This will facilitate sustainable management, hopefully assisting to curb global overfishing and assure that the growing world population continues to have fair access to fish protein.

Acknowledgments The inspiration for this chapter originates from the EU-FP6 Network of Excellence *Marine Genomics Europe* (GOCE-CT-2004-505403) and the EU-FP6 STREP projects *Combined genetic and functional genomic approaches for stress and disease resistance marker-assisted selection in fish and shellfish* (Aquafirst; STREP-2004-513692), *Fisheries induced evolution* (FinE; STREP-2007-44276) and EU-FP7 KBBE-2007-1.2-14 project *The structure of fish populations and traceability of fish and fish products* (FishPopTrace; KBBE-2007-1.2-14).

References

Abee T, Van Schaik W, Siezen RJ (2004) Impact of genomics on microbial food safety. Trends Biotechnol 22:653–660

Abollo E, Casas SM, Ceschia G et al (2006) Differential diagnosis of *Perkinsus* species by polymerase chain reaction-restriction fragment length polymorphism assay. Mol Cell Probes 20:323–329

Adams A, Thompson KD (2006) Biotechnology offers revolution to fish health management. Trends Biotechnol 24:201–205

Allam B, Paillard C, Auffret M et al (2006) Effects of the pathogenic *Vibrio tapetis* on defence factors of susceptible and non-susceptible bivalve species: II. Cellular and biochemical changes following in vivo challenge. Fish Shellfish Immunol 20:384–397

Allendorf FW, England PR, Luikart G et al (2008) Genetic effects of harvest on wild animal populations. Trends Ecol Evol 23:327–337

Anderson EC, Garza JC (2006) The power of single-nucleotide polymorphisms for large-scale parentage inference. Genetics 172:2567–2582

Andersson L, Georges M (2004) Domestic-animal genomics: deciphering the genetics of com-plex traits. Nat Rev Genet 5:202–212

Araneda C, Neira R, Iturra P (2005) Identification of a dominant SCAR marker associated with colour traits in Coho salmon (*Oncorhynchus kisutch*). Aquaculture 247:67–73

Arkush KD, Giese AR, Mendonca HL et al (2002) Resistance to three pathogens in the endangered winter-run Chinook salmon (*Oncorhynchus tshawytscha*): effects of inbreeding and major histocompatibility complex genotypes. Can J Fish Aquat Sci 59:966–975

Artamonova VS (2007) Genetic markers in population studies of Atlantic salmon *Salmo salar* L.: Analysis of DNA sequences. Russ J Genet 43:341–353

Ascenso RMT, Leite RB, Afonso R et al (2007) Suppression-subtractive hybridization: A rapid and inexpensive detection methodology for up-regulated *Perkinsus olseni* genes. Afr J Biochem Res 3:24–28

Athanassopoulou F, Billinis C, Prapas T (2004) Important disease conditions of newly cultured species in intensive freshwater farms in Greece: first incidence of nodavirus infection in *Acipenser* sp. Dis Aquat Organ 60:247–252

Audemard C, Ragone Calvo LM, Paynter KT et al (2006) Real-time PCR investigation of parasite ecology: in situ determination of oyster parasite *Perkinsus marinus* transmission dynamics in lower Chesapeake Bay. Parasitology 132:827–842

Bachère E, Gueguen Y, Gonzalez M et al (2004) Insights into the anti-microbial defense of marine invertebrates: the penaeid shrimps and the oyster *Crassostrea gigas*. Immunol Rev 198: 149–168

Badariotti F, Lelong C, Dubos MP et al (2007) Characterization of chitinase-like proteins (Cg-Clp1 and Cg-Clp2) involved in immune defence of the mollusc *Crassostrea gigas*. FEBS J 274:3646–3654

Bai J, Solberg C, Fernandes JM et al (2007) Profiling of maternal and developmental-stage specific mRNA transcripts in Atlantic halibut *Hippoglossus hippoglossus*. Gene 386:202–210

Baker CS, Cooke JG, Lavery S et al (2007) Estimating the number of whales entering trade using DNA profiling and capture-recapture analysis of market products. Mol Ecol 16:2617–2626

Barrett RDH, Rogers SM, Schluter D (2008) Natural selection on a major armor gene in threespine stickleback. Science 322:255–257

Batista FM, Arzul I, Pepin JF et al (2007) Detection of ostreid herpesvirus 1 DNA by PCR in bivalve molluscs: a critical review. J Virol Methods 139:1–11

Battershill JM (2005) Toxicogenomics: regulatory perspective on current position. Hum Exp Toxicol 24:35–40

Beaumont MA, Nichols RA (1996) Evaluating loci for use in the genetic analysis of population structure. Proc R Soc B-Biol Sci 263:1619–1626

Beaumont MA (2005) Adaptation and speciation: what can F-st tell us?. Trends Ecol Evol 20: 435–440

Bencze-Røra AM, Regost C, Lampe J (2003) Liquid holding capacity, texture and fatty acid profile of smoked fillets of Atlantic salmon fed diets containing fish oil or soybean oil. Food Res Int 36:231–239

Bernatchez L, Landry C (2003) MHC studies in nonmodel vertebrates: what have we learned about natural selection in 15 years?. J Evolution Biol 16:363–377

Bettencourt R, Roch P, Stefanni S et al (2007) Deep sea immunity: Unveiling immune constituents from the hydrothermal vent mussel *Bathymodiolus azoricus*. Mar Environ Res 64:108–127

Bierne N, Tsitrone A, David P (2000) An inbreeding model of associative overdominance during a population bottleneck. Genetics 155:1981–1990

Bílek K, Knoll A, Stratil A et al (2008) Analysis of mRNA expression of CNN3, DCN, FBN2, POSTN, SPARC and YWHAQ genes in porcine foetal and adult skeletal muscles. Czech J Anim Sci 53:181–186

Birney E, Stamatoyannopoulos JA, Dutta A et al (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature 447:799–816

Blackstone GM, Nordstrom JL, Vickery MC et al (2003) Detection of pathogenic *Vibrio parahaemolyticus* in oyster enrichments by real time PCR. J Microbiol Methods 53:149–155

Blanco G, Borrell YJ, Bernardo D (2007) The use of microsatellites for optimizing broodstocks in a hatchery of gilthead seabream (*Sparus aurata* L.). Aquaculture 272:S246

Boesen HT, Pedersen K, Larsen JL et al (1999) *Vibrio anguillarum* resistance to rainbow trout (*Oncorhynchus mykiss*) serum: role of O-antigen structure of lipopolysaccharide. Infect Immun 67:294–301

Bonin A (2008) Population genomics: a new generation of genome scans to bridge the gap with functional genomics. Mol Ecol 17:3583–3584

Borrell YJ, Alvarez J, Vazquez E et al (2004) Applying microsatellites to the management of farmed turbot stocks (*Scophthalmus maximus* L.) in hatcheries. Aquaculture 241:133–150

Bouck A, Vision T (2007) The molecular ecologist's guide to expressed sequence tags. Mol Ecol 16:907–924

Boudry P, Collet B, Cornette F et al (2002) High variance in reproductive success of the Pacific oyster (*Crassostrea gigas*, Thunberg) revealed by microsatellite-based parentage analysis of multifactorial crosses. Aquaculture 204:283–296

Boudry P, Dégremont L, Haffray P (2008) The genetic basis of summer mortality in Pacific oyster spat and potential for improving survival by selective breeding in France. In: Samain JF, McCombie H (eds) Summer mortality of Pacific oyster *Crassostrea gigas* – The morest project, Quae edn. Versailles, France

Boughman JW (2001) Divergent sexual selection enhances reproductive isolation in sticklebacks. Nature 411:944–948

Bourre JM (2005) Dietary omega-3 fatty acids and psychiatry: mood, behaviour, stress, depression, dementia and aging. J Nutr Health Aging 9:31–38

Boutet I, Tanguy A, Moraga D (2004) Response of the Pacific oyster *Crassostrea gigas* to hydrocarbon contamination under experimental conditions. Gene 329:147–157

Bouza C, Hermida M, Pardo BG et al (2007) A microsatellite genetic map of the turbot (*Scophthalmus maximus*). Genetics 177:2457–2467

Bramble L, Anderson RS (1997) Modulation of *Crassostrea virginica* hemocyte reactive oxygen species production by *Listonella anguillarum*. Dev Comp Immunol 21:337–348

Brasher CW, DePaola A, Jones DD et al (1998) Detection of microbial pathogens in shellfish with multiplex PCR. Curr Microbiol 37:101–107

Brauns LA, Hudson MC, Oliver JD (1991) Use of the polymerase chain reaction in detection of culturable and nonculturable *Vibrio vulnificus* cells. Appl Environ Microbiol 57:2651–2655

Brenner S, Williams SR, Vermaas EH et al (2000) In vitro cloning of complex mixtures of DNA on microbeads: physical separation of differentially expressed cDNAs. Proc Natl Acad Sci USA 97:1665–1670

Brul S, Schuren F, Montijn R et al (2006) The impact of functional genomics on microbiological food quality and safety. Int J Food Microbiol 112:195–199

Buchmann K, Lindenstrom T (2002) Interactions between monogenean parasites and their fish hosts. Int J Parasitol 32:309–319

Buchrieser C, Gangar VV, Murphree RL et al (1995) Multiple *Vibrio vulnificus* strains in oysters as demonstrated by clamped homogeneous electric field gel electrophoresis. Appl Environ Microbiol 61:1163–1168

Bugeon J, Lefevre F, Fauconneau B (2003) Fillet texture and muscle structure in brown trout (*Salmo trutta*) subjected to long-term exercise. Aquac Res 34:1287–1295

Bulgakov AA, Park KI, Choi KS et al (2004) Purification and characterisation of a lectin isolated from the Manila clam *Ruditapes philippinarum* in Korea. Fish Shellfish Immunol 16:487–499

Burnett KG, Bain L, Baldwin WS et al (2007) *Fundulus* as the premier teleost model in environmental biology: Opportunities for new insights using genomics. Comp Biochem Physiol D-Genomics Proteomics 2:257–286

Bustin SA, Benes V, Nolan T et al (2005) Quantitative real-time RT-PCR – a perspective. J Mol Endocrinol 34:597–601

Calder PC (2008) Polyunsaturated fatty acids, inflammatory processes and inflammatory bowel diseases. Mol Nutr Food Res 52:885–897

Camara M, Evans F, Langdson CJ (2008) Parental relatedness and survival of Pacific oysters from a naturalized population. J Shellfish Res 27:323–336

Campana SE, Thorrold SR (2001) Otoliths, increments, and elements: keys to a comprehensive understanding of fish populations? Can J Fish Aquat Sci 58:30–38

Campbell MS, Wright AC (2003) Real-time PCR analysis of *Vibrio vulnificus* from oysters. Appl Environ Microbiol 69:7137–7144

Campbell D, Bernatchez L (2004) Generic scan using AFLP markers as a means to assess the role of directional selection in the divergence of sympatric whitefish ecotypes. Mol Biol Evol 21:945–956

Canesi L, Gallo G, Gavioli M et al (2002) Bacteria-hemocyte interactions and phagocytosis in marine bivalves. Microsc Res Tech 57:469–476

Carvalho GR, Hauser L (1998) Advances in the molecular analysis of fish population structure. Ital J Zool 65:21–33

Case RAJ, Hutchinson WF, Hauser L et al (2005) Macro- and micro-geographic variation in pantophysin (Pan I) allele frequencies in NE Atlantic cod *Gadus morhua*. Mar Ecol – Prog Ser 301:267–278

Castaño-Sanches C, Fuji K, Ozaki A et al (2007) High-density linkage map of the Japanese flounder, *Paralichthys olivaceus*. Aquaculture 272:S248

Castro J, Bouza C, Presa P et al (2004) Potential sources of error in parentage assessment of turbot (*Scophthalmus maximus*) using microsatellite loci. Aquaculture 242:119–135

Castro J, Pino A, Hermida M et al (2006) A microsatellite marker tool for parentage analysis in Senegal sole (*Solea senegalensis*): Genotyping errors, null alleles and conformance to theoretical assumptions. Aquaculture 261:1194–1203

Castro J, Pino A, Hermida M et al (2007) A microsatellite marker tool for parentage assessment in gilthead seabream (*Sparus aurata*). Aquaculture 272:S210–S216

Cerda J, Mercade J, Lozano JJ et al (2008) Genomic resources for a commercial flatfish, the Senegalese sole (*Solea senegalensis*): EST sequencing, oligo microarray design, and development of the bioinformatic platform Soleamold. BMC Genomics 9:508

Charlet M, Chernysh S, Philippe H et al (1996) Innate immunity: isolation of several cysteine-rich antimicrobial peptides from the blood of a mollusc, *Mytilus edulis*. J Biol Chem 271: 21808–21813

Cheng T (1981) Bivalves. In: Ratcliffe NA, Rowley A (eds) Invertebrate blood cells. Academic Press, London

Chistiakov DA, Hellemans B, Volckaert FAM (2006) Microsatellites and their genomic distribution, evolution, function and applications: A review with special reference to fish genetics. Aquaculture 255:1–29

Choquet G, Soudant P, Lambert C et al (2003) Reduction of adhesion properties of *Ruditapes philippinarum* hemocytes exposed to *Vibrio tapetis*. Dis Aquat Organ 57:109–116

Christiansen FB, Frydenberg O (1974) Geographical patterns of 4 polymorphisms in *Zoarces viviparus* as evidence of selection. Genetics 77:765–770

Clark AG, Hubisz MJ, Bustamante CD et al (2005) Ascertainment bias in studies of human genome-wide polymorphism. Genome Res 15:1496–1502

Cnaani A, Hallerman EM, Ron M et al (2003) Detection of a chromosomal region with two quantitative trait loci, affecting cold tolerance and fish size, in an F2 tilapia hybrid. Aquaculture 223:117–128

Cnaani A, Zilberman N, Tinman S et al (2004) Genome-scan analysis for quantitative trait loci in an F-2 tilapia hybrid. Mol Genet Genomics 272:162–172

Cnaani A, Hulata G (2008) Tilapias. In: Kocher TD, Kole C (eds) Genome mapping and genomics in animals, vol 2. Springer-Heidelberg, Berlin

Cochennec N, Le Roux F, Berthe F et al (2000) Detection of *Bonamia ostreae* based on small subunit ribosomal probe. J Invertebr Pathol 76:26–32

Colborn J, Crabtree RE, Shaklee JB et al (2001) The evolutionary enigma of bonefishes (*Albula* spp.): Cryptic species and ancient separations in a globally distributed shorefish. Evolution 55:807–820

Coltman DW (2008) Molecular ecological approaches to studying the evolutionary impact of selective harvesting in wildlife. Mol Ecol 17:221–235

Cossins AR, Crawford DL (2005) Opinion – fish as models for environmental genomics. Nat Rev Genet 6:324–333

Cossins A, Fraser J, Hughes M et al (2006) Post-genomic approaches to understanding the mechanisms of environmentally induced phenotypic plasticity. J Exp Biol 209:2328–2336

Cowen RK, Paris CB, Srinivasan A (2006) Scaling of connectivity in marine populations. Science 311:522–527

Coyne JA, Orr HA (2004) Speciation. Sinauer Associates, MA

Cuesta A, Meseguer J, Esteban MA (2008) The antimicrobial peptide hepcidin exerts an important role in the innate immunity against bacteria in the bony fish gilthead seabream. Mol Immunol 45:2333–2342

Cunningham C, Hikima JI, Jenny MJ et al (2006) New resources for marine genomics: Bacterial artificial chromosome libraries for the eastern and pacific oysters (*Crassostrea virginica* and *C. gigas*). Mar Biotechnol 8:521–533

Curole JP, Hedgecock D (2005) High frequency of SNPs in the Pacific oyster genome. http://intl-pag.org/13/abstracts/PAG13_W026.html

Cushing DH (1969) Regularity of spawning season of some fishes. Journal du conseil international de l'exploitation de la mer 33:81–92

Dalvit C, De Marchi M, Cassandro M (2007) Genetic traceability of livestock products: A review. Meat Sci 77:437–449

Darawiroj D, Kondo H, Hirono I et al (2008) Immune-related gene expression profiling of yellowtail (*Seriola quinqueradiata*) kidney cells stimulated with ConA and LPS using microarray analysis. Fish Shellfish Immunol 24:260–266

Darias MJ, Zambonino-Infante JL, Hugot K et al (2008) Gene expression patterns during the larval development of European sea bass (*Dicentrarchus labrax*) by microarray analysis. Mar Biotechnol 10:416–428

David E, Boudry R, Degremont L et al (2007) Genetic polymorphism of glutamine synthetase and delta-9 desaturase in families of Pacific oyster *Crassostrea gigas* and susceptibility to summer mortality. J Exp Mar Biol Ecol 349:272–283

Davis MB, Shaw RG, Etterson JR (2005) Evolutionary responses to climate change. Ecology 86:1704–1714

Delaney JR (2007) NEPTUNE: Transforming ocean and earth sciences with distributed submarine sensor networks wired to next generation internet. In: 2007 Symposium on Underwater Technology and Workshop on Scientific Use of Submarine Cables and Related Technologies

Delgado CL, Wada N, Rosegrant MW et al (2003) Outlook for fish to 2020: meeting global demand. In: Food Policy Report – International Food Policy Research Institute WorldFish Center, Malaysia, Washington, DC

Derelle E, Ferraz C, Rombauts S et al (2006) Genome analysis of the smallest free-living eukaryote *Ostreococcus tauri* unveils many unique features. Proc Natl Acad Sci USA 103: 11647–11652

Derome N, Bernatchez L (2006) The transcriptomics of ecological convergence between two limnetic coregonine fishes (Salmonidae). Mol Biol Evol 23:2370–2378

Derome N, Duchesne P, Bernatchez L (2006) Parallelism in gene transcription among sympatric lake whitefish ecotypes (*Coregonus clupeaformis* Mitchill). Mol Ecol 15:1239–1250

De-Santis C, Jerry DR (2007) Candidate growth genes in finfish – Where should we be looking?. Aquaculture 272:22–38

Desenclos JC, Klontz KC, Wilder MH et al (1991) A multistate outbreak of hepatitis A caused by the consumption of raw oysters. Am J Public Health 81:1268–1272

Diatchenko L, Lukyanov S, Lau YF et al (1999) Suppression subtractive hybridization: a versatile method for identifying differentially expressed genes. Methods Enzymol 303:349–380

Dios S, Novoa B, Buonocore F et al (2008) Genomic resources for immunology and disease of salmonid and non-salmonid fish. Rev Fish Sci 16:119–132

Dondero F, Piacentini L, Marsano F et al (2006) Gene transcription profiling in pollutant ex-posed mussels (*Mytilus* spp.) using a new low-density oligonucleotide microarray. Gene 376:24–36

Douglas SE, Gallant JW, Bullerwell CE et al (1999) Winter flounder expressed sequence tags: Establishment of an EST database and identification of novel fish genes. Mar Biotechnol 1: 458–464

Dulvy NK, Sadovy Y, Reynolds JD (2003) Extinction vulnerability in marine populations. Fish Fish 4:25–64

Dulvy NK, Rogers SI, Jennings S et al (2008) Climate change and deepening of the North Sea fish assemblage: a biotic indicator of warming seas. J Appl Ecol 45:1029–1039

Dupray E, Caprais MP, Derrien A et al (1997) Salmonella DNA persistence in natural seawaters using PCR analysis. J Appl Microbiol 82:507–510

Duvernell DD, Lindmeier JB, Faust KE (2008) Relative influences of historical and contemporary forces shaping the distribution of genetic variation in the Atlantic killifish, *Fundulus heteroclitus*. Mol Ecol 17:1344–1360

Elandalloussi LM, Leite RM, Afonso R et al (2004) Development of a PCR-ELISA assay for diagnosis of *Perkinsus marinus* and *Perkinsus atlanticus* infections in bivalve molluscs. Mol Cell Probes 18:89–96

Ellegren H (2004) Microsatellites: simple sequences with complex evolution. Nat Rev Genet 5:435–445

Emmerich R, Weibel E (1894) Ueber eine durch bakterien erzengte seuche unter den forellen. Arch Hyg Bakteriol 21:1–21

Endler JA (1986) Natural selection in the wild. Princeton University Press, Princeton

Escoubas J-M, Briant L, Montagnani C et al (1999) Oyster IKK-like protein shares structural and functional properties with its mammalian homologues. FEBS Lett 453:293–298

Espe M, Ruohonen K, Bjørnevik M et al (2004) Interactions between ice storage time, collagen composition, gaping and textural properties in farmed salmon muscle harvested at different times of the year. Aquaculture 240:489–504

Ewart KV, Belanger JC, Williams J et al (2005) Identification of genes differentially expressed in Atlantic salmon (*Salmo salar*) in response to infection by *Aeromonas salmonicida* using cDNA microarray technology. Dev Comp Immunol 29:333–347

Faisal M, MacIntyre EA, Adham KG et al (1998) Evidence for the presence of protease inhibitors in Eastern (*Crassostrea virginica*) and Pacific (*Crassostrea gigas*) oysters. Comp Biochem Physiol B-Biochem Mol Biol 121:161–168

Faisal M, Schafhauser DY, Garreis KA et al (1999) Isolation and characterization of *Perkinsus marinus* proteases using bacitracin-sepharose affinity chromatography. Comp Biochem Physiol B-Biochem Mol Biol 123:417–426

Falch E, Rustad T, Jonsdottir R et al (2006) Geographical and seasonal differences in lipid composition and relative weight of by-products from gadiform species. J Food Compos Anal 19:727–736

Faliex E, Da Silva C, Simon G et al (2008) Dynamic expression of immune response genes in the sea bass, *Dicentrarchus labrax*, experimentally infected with the monogenean *Diplectanum aequans*. Fish Shellfish Immunol 24:759–767

Fawcett LB, Tripp MR (1994) Chemotaxis of *Mercenaria mercenaria* hemocytes to bacteria in vitro. J Invertebr Pathol 63:275–284

Feder ME, Mitchell-Olds T (2003) Evolutionary and ecological functional genomics. Nat Rev Genet 4:651–657

Fernandes JMO, Mackenzie MG, Elgar G et al (2005) A genomic approach to reveal novel genes associated with myotube formation in the model teleost, *Takifugu rubripes*. Physiol Genomics 22:327–338

Ferraresso S, Vitulo N, Mininni AN et al (2008) Development and validation of a gene expression oligo microarray for the gilthead sea bream (Sparus aurata). BMC Genomics 9:580

Fishback AG, Danzmann RG, Ferguson MM et al (2002) Estimates of genetic parameters and genotype by environment interactions for growth traits of rainbow trout (*Oncorhynchus mykiss*) as inferred using molecular pedigrees. Aquaculture 206:137–150

Ford SE, Haskin HH (1987) Infection and mortality patterns in strains of oysters *Crassostrea virginica* selected for resistance to the parasite *Haplosporidium nelsoni* (MSX). J Parasitol 73:368–376

Franch R, Louro B, Tsalavouta M et al (2006) A genetic linkage map of the hermaphrodite teleost fish *Sparus aurata*. Genetics 174:851–861

Frerichs GN, Rodger HD, Peric Z (1996) Cell culture isolation of piscine neuropathy nodavirus from juvenile sea bass, *Dicentrarchus labrax*. J Gen Virol 77:2067–2071

Fryer JL, Sanders JE (1981) Bacterial kidney disease of salmonid fish. Annu Rev Microbiol 35:273–298

Gahr SA, Vallejo RL, Weber GM et al (2008) Effects of short-term growth hormone treatment on liver and muscle transcriptomes in rainbow trout (*Oncorhynchus mykiss*). Physiol Genomics 32:380–392

Galindo HM, Olson DB, Palumbi SR (2006) Seascape genetics: A coupled oceanographic-genetic model predicts population structure of Caribbean corals. Curr Biol 16:1622–1626

Gallardo JA, Garcia X, Lhorente JP et al (2004) Inbreeding and inbreeding depression of female reproductive traits in two populations of Coho salmon selected using BLUP predictors of breeding values. Aquaculture 234:111–122

Gauthier JD, Jenkins JA, La Peyre JF (2004) Flow cytometric analysis of lectin binding to in vitro-cultured *Perkinsus marinus* surface carbohydrates. J Parasitol 90:446–454

Geisler R, Rauch GJ, Baier H et al (1999) A radiation hybrid map of the zebrafish genome. Nat Genet 23:86–89

Gerber S, Mariette S, Streiff R et al (2000) Comparison of microsatellites and amplified fragment length polymorphism markers for parentage analysis. Mol Ecol 9:1037–1048

Gerwick L, Corley-Smith G, Bayne CJ (2007) Gene transcript changes in individual rainbow trout livers following an inflammatory stimulus. Fish Shellfish Immunol 22:157–171

Gestal C, Costa M, Figueras A et al (2007) Analysis of differentially expressed genes in response to bacterial stimulation in hemocytes of the carpet-shell clam *Ruditapes decussatus*: identification of new antimicrobial peptides. Gene 406:134–143

Gharbi K, Gautier A, Danzmann RG et al (2006) A linkage map for brown trout (*Salmo trutta*): chromosome homeologies and comparative genome organization with other salmonid fish. Genetics 172:2405–2419

Gienapp P, Teplitsky C, Alho JS et al (2008) Climate change and evolution: disentangling environmental and genetic responses. Mol Ecol 17:167–178

Gil LA (2007) PCR-based methods for fish and fishery products authentication. Trends Food Sci Technol 18:558–566

Gilbey J, Verspoor E, Mo TA et al (2006) Identification of genetic markers associated with *Gyrodactylus salaris* resistance in Atlantic salmon *Salmo salar*. Dis Aquat Organ 71: 119–129

Gjedrem T (2000) Genetic improvement of cold-water fish species. Aquac Res 31:25–33

Goetsch SC, Hawke TJ, Gallardo TD et al (2003) Transcriptional profiling and regulation of the extracellular matrix during muscle regeneration. Physiol Genomics 14:261–271

Gomez-Bautista M, Ortega-Mora LM, Tabares E et al (2000) Detection of infectious *Cryptosporidium parvum* oocysts in mussels (*Mytilus galloprovincialis*) and cockles (*Cerastoderma edule*). Appl Environ Microbiol 66:1866–1870

Gong Z, Hu Z, Gong ZQ et al (1994) Bulk isolation and identification of fish genes by cDNA clone tagging. Mol Mar Biol Biotechnol 3:243–251

Gonzalez I, Garcia T, Fernandez A et al (1999) Rapid enumeration of *Escherichia coli* in oysters by a quantitative PCR-ELISA. J Appl Microbiol 86:231–236

Gonzalez M, Romestand B, Fievet J et al (2005) Evidence in oyster of a plasma extracellular superoxide dismutase which binds LPS. Biochem Biophys Res Commun 338:1089–1097

Gonzalez M, Gueguen Y, Destoumieux-Garzon D et al (2007a) Evidence of a bactericidal permeability increasing protein in an invertebrate, the *Crassostrea gigas* Cg-BPI. Proc Natl Acad Sci USA 104:17759–17764

Gonzalez M, Gueguen Y, Desserre G et al (2007b) Molecular characterization of two isoforms of defensin from hemocytes of the oyster *Crassostrea gigas*. Dev Comp Immunol 31:332–339

Gracey AY (2007) Interpreting physiological responses to environmental change through gene expression profiling. J Exp Biol 210:1584–1592

Gracey AY, Fraser EJ, Li W et al (2004) Coping with cold: An integrative, multitissue analysis of the transcriptome of a poikilothermic vertebrate. Proc Natl Acad Sci USA 101:16970–16975

Graczyk TK, Thompson RC, Fayer R et al (1999) *Giardia duodenalis* cysts of genotype A recovered from clams in the Chesapeake Bay subestuary, Rhode River. Am J Trop Med Hyg 61:526–529

Gregory SG, Sekhon M, Schein J et al (2002) A physical map of the mouse genome. Nature 418:743–750

Grimholt U, Larsen S, Nordmo R et al (2003) MHC polymorphisms and disease resistance in Atlantic salmon (*Salmo salar*); facing pathogens with single expressed major histocompatibility class I and class II loci. Immunogenetics 55:210–219

Gross R, Lulla P, Paaver T (2007) Genetic variability and differentiation of rainbow trout (*Oncorhynchus mykiss*) strains in Northern and Eastern Europe. Aquaculture 272: S139–S146

Gueguen Y, Cadoret JP, Flament D et al (2003) Immune gene discovery by expressed sequence tags generated from hemocytes of the bacteria-challenged oyster, *Crassostrea gigas*. Gene 303: 139–145

Gueguen Y, Herpin A, Aumelas A et al (2006) Characterization of a Defensin from the oyster *Crassostrea gigas*: Recombinant production, folding, solution structure, antimicrobial activities, and gene expression. J Biol Chem 281:313–323

Guelinckx J, Maes J, Geysen B et al (2008) Estuarine recruitment of a marine goby reconstructed with an isotopic clock. Oecologia 157:41–52

Guerard F, Sellos D, Le Ga Y (2005) Fish and shellfish upgrading, traceability. Mar Biotechnol 96:127–163

Guinand B, Lemaire C, Bonhomme F (2004) How to detect polymorphisms undergoing selection in marine fishes? A review of methods and case studies, including flatfishes. J Sea Res 51:167–182

Gunnarsson L, Kristiansson E, Forlin L et al (2007) Sensitive and robust gene expression changes in fish exposed to estrogen: a microarray approach. BMC Genomics 8:149

Gupta PK (2008) Ultrafast and low-cost DNA-sequencing methods for applied genomics research. Proc Natl Acad Sci USA 78:91–102

Hanchuan D, Liangqi L, Guang D (2006) Molecular cloning of the obese gene from *Cyprinus carpio* and its expression in *Escherichia coli*. Front Biol China 1:50–55

Hansen MM, Kenchington E, Nielsen EE (2001) Assigning individual fish to populations using microsatellite DNA markers: Methods and applications. Fish Fish 2:93–112

Hanski IA, Gilpin ME (1997) Metapopulation biology: ecology, genetics and evolution. Academic Press, New York

Hanski IA, Gaggiotti OE (2004) Ecology, genetics, and evolution of metapopulations. Elsevier Academic Press, San Diego

Harlizius B, van Wijk R, Merks JWM (2004) Genomics for food safety and sustainable animal production. J Biotechnol 113:33–42

Håstein T, Hill BJ, Berthe F et al (2001) Traceability of aquatic animals. Rev Sci Tech 20:564–583

Håstein T, Scarfe AD, Lund VL (2005) Science-based assessment of welfare: aquatic animals. Rev Sci Tech 24:529–547

Hauser L, Adcock GJ, Smith PJ et al (2002) Loss of microsatellite diversity and low effective population size in an overexploited population of New Zealand snapper (*Pagrus auratus*). Proc Nat Acad Sci USA 99:11742–11747

Hawke JP, McWhorter AC, Steigerwalt AG et al (1981) *Edwardsiella ictaluri* sp. nov., the causative agent of enteric septicemia of catfish. Int J Syst Bacteriol 31:396–400

Hayes B, Sonesson AK, Gjerde B (2005) Evaluation of three strategies using DNA markers for traceability in aquaculture species. Aquaculture 250:70–81

Hayes B, He J, Moen T et al (2006) Use of molecular markers to maximise diversity of flounder populations for aquaculture breeding programs. Aquaculture 255:573–578

Hayes B, Lærdahl JK, Lien S et al (2007) An extensive resource of single nucleotide polymorphism markers associated with Atlantic salmon (*Salmo salar*) expressed sequences. Aquaculture 265:82–90

He XP, Xu XW, Zhao SH et al (2008) Investigation of Lpin1 as a candidate gene for fat deposition in pigs. Mol Biol Rep DOI 10.1007/s11033-008-9294-4

Hedgecock D (1994) Does variance in reproductive success limit effective population sizes of marine organisms? In: Beaumont AR (ed) Genetics and evolution of aquatic organisms. Chapman and Hall, London

Hedgecock D, Li G, Hubert S et al (2004) Widespread null alleles and poor cross-species amplification of microsatellite DNA loci cloned from the Pacific oyster, *Crassostrea gigas*. J Shellfish Res 23:379–385

Hedgecock D, Davis J (2007) Heterosis for yield and crossbreeding of the Pacific oyster *Crassostrea gigas*. Aquaculture 272:S17–S29

Hedgecock D, Lin JZ, DeCola S (2007) Transcriptomic analysis of growth heterosis in larval Pacific oysters (*Crassostrea gigas*). Proc Natl Acad Sci USA 104:2313–2318

Hegde A, Teh HC, Lam TJ et al (2003) Nodavirus infection in freshwater ornamental fish, guppy, *Poicelia reticulate* – comparative characterization and pathogenicity studies. Arch Virol 148:575–586

Heijne WHM, Kienhuis AS, van Ommen B et al (2005) Systems toxicology: applications of toxicogenomics, transcriptomics, proteomics and metabolomics in toxicology. Expert Rev Proteomics 2:767–780

Hemmer-Hansen J, Nielsen EE, Frydenberg J et al (2007) Adaptive divergence in a high gene flow environment: Hsc70 variation in the European flounder (*Platichthys flesus* L.). Heredity 99:592–600

Hendriks IE, Duarte CM, Heip CHR (2006) Biodiversity research still grounded. Science 312:1715

Hendriks IE, Duarte CM (2008) Allocation of effort and imbalances in biodiversity research. J Exp Mar Biol Ecol 360:15–20

Herbinger CM, Doyle RW, Pitman ER (1995) DNA fingerprint based analysis of paternal and maternal effects on offspring growth and survival in communally reared rainbow trout. Aquaculture 137:245–256

Herlin M, Taggart JB, Mcandrew BJ et al (2007) Parentage allocation in a complex situation: A large commercial Atlantic cod (*Gadus morhua*) mass spawning tank. Aquaculture 272:195–203

Herlin M, Delghandi M, Wesmajervi M (2008) Analysis of the parental contribution to a group of fry from a single day of spawning from a commercial Atlantic cod (*Gadus morhua*) breeding tank. Aquaculture 274:218–224

Higuchi R, Fockler C, Dollinger G et al (1993) Kinetic PCR analysis: real-time monitoring of DNA amplification reactions. Biotechnology 11:1026–1030

Hill WE, Keasler SP, Trucksess MW et al (1991) Polymerase chain reaction identification of *Vibrio vulnificus* in artificially contaminated oysters. Appl Environ Microbiol 57:707–711

Hill JA, Kiessling A, Devlin RH (2000) Coho salmon (*Oncorhynchus kisutch*) transgenic for a growth hormone gene construct exhibit increased rates of muscle hyperplasia and detectable levels of differential gene expression. Can J Fish Aquat Sci 57:939–950

Hill BJ (2005) The need for effective disease control in international aquaculture. Dev Biol 121:3–12

Hine P (1999) The inter-relationships of bivalve haemocytes. Fish Shellfish Immunol 9:367–385

Hine PM, Cochennec-Laureau N, Berthe FC (2001) *Bonamia exitiosus* n. sp. (Haplosporidia) infecting flat oysters *Ostrea chilensis* in New Zealand. Dis Aquat Organ 47:63–72

Hitte C, Kirkness EF, Ostrander EA et al (2008) Survey sequencing and radiation hybrid mapping to construct comparative maps. Methods Mol Biol 422:65–77

Hjort J (1914) Fluctuations in the great fisheries of Northern Europe. Rapp Proc-Verb Réun Cons Int Expl Mer 20:1–228

Holloway AJ, van Laar RK, Tothill RW et al (2002) Options available – from start to finish – for obtaining data from DNA microarrays II. Nat Genet 32 Suppl:481–489

Hostetler HA, Collodi P, Devlin RH et al (2005) Improved phytate phosphorus utilization by Japanese medaka transgenic for the *Aspergillus niger* phytase gene. Zebrafish 2:19–31

Houston RD, Gheyas A, Hamilton A et al (2008) Detection and confirmation of a major QTL affecting resistance to infectious pancreatic necrosis (IPN) in Atlantic salmon (*Salmo salar*). Dev Biol 132:199–204

Howland K, Cheng TC (1982) Identification of bacterial chemoattractants for oyster (*Crassostrea virginica*) hemocytes. J Invertebr Pathol 89:123–132

Hubert F, Noel T, Roch P (1996) A member of the arthropod defensin family from edible Mediterranean mussels (*Mytilus galloprovincialis*). Eur J Biochem 240:302–306

Hubert S, Hedgecock D (2004) Linkage maps of microsatellite DNA markers for the Pacific oyster *Crassostrea gigas*. Genetics 168:351–362

Hukriede NA, Joly L, Tsang M et al (1999) Radiation hybrid mapping of the zebrafish genome. Proc Natl Acad Sci USA 6:9745–9750

Hurling R, Rodell JB, Hunt HD (1996) Fibre diameter and fish texture. J Texture Studies 27: 679–685

Huvet A, Herpin A, Degremont L et al (2004) The identification of genes from the oyster *Crassostrea gigas* that are differentially expressed in progeny exhibiting opposed susceptibility to summer mortality. Gene 343:211–220

Huvet A, Jeffroy F, Fabioux C et al (2008) Association among growth, food consumption-related traits and *amylase* gene polymorphism in the Pacific oyster *Crassostrea gigas*. Anim Genet 39:662–665

Itoh N, Takahashi KG (2008) Distribution of multiple peptidoglycan recognition proteins in the tissues of Pacific oyster, *Crassostrea gigas*. Comp Biochem Physiol B-Biochem Mol Biol 150:409–417

Izquierdo MS, Robaina L, Juárez E et al (2008) Regulation of growth, fatty acid composition and delta 6 desaturase expression by dietary lipids in gilthead seabream larvae (*Sparus aurata*). Fish Physiol Biochem 34:117–127

Janeway CA, Medzhitov R (2002) Innate immune recognition. Annu Rev Immunol 20:197–216

Jenny MJ, Warr GW, Ringwood AH et al (2006) Regulation of metallothionein genes in the American oyster (*Crassostrea virginica*): ontogeny and differential expression in response to different stressors. Gene 379:156–165

Jenny MJ, Chapman RW, Mancia A (2007) A cDNA microarray for *Crassostrea virginica* and *C. gigas*. Mar Biotechnol 9:577–591

Jeyasekaran G, Karunasagar I (1996) Incidence of *Listeria* spp. in tropical fish. Int J Food Microbiol 31:333–340

Johnston IA (1999) Muscle development and growth: potential implications for flesh quality in fish. Aquaculture 177:99–115

Johnston IA, Manthri S, Alderson R et al (2002) Effects of dietary protein level on muscle cellularity and flesh quality in Atlantic salmon with particular reference to gaping. Aquaculture 210:259–283

Johnston IA, Alderson R, Sandham C et al (2004) Muscle fibre density in relation to the colour and texture of smoked Atlantic salmon (*Salmo salar* L.). Aquaculture 189:335–349

Johnston IA (2006) Environment and plasticity of myogenesis in teleost fish. J Exp Biol 209:2249–2264

Jones GP, Planes S, Thorrold SR (2005) Coral reef fish larvae settle close to home. Curr Biol 15:1314–1318

Joost S, Bonin A, Bruford W et al (2007) A spatial analysis method (SAM) to detect candidate loci for selection: towards a landscape genomics approach to adaptation. Mol Ecol 16:3955–3969

Jordal A-EO, Torstensen BE, Tsoi S et al (2005) Dietary rapeseed oil affects the expression of genes involved in hepatic lipid metabolism in Atlantic salmon (*Salmo salar* L.). J Nutr 135:2355–2361

Jørgensen HBH, Hansen MM, Bekkevold D et al (2005) Marine landscapes and population genetic structure of herring (*Clupea harengus* L.) in the Baltic Sea. Mol Ecol 14:3219–3234

Jørgensen HBH, Pertoldi C, Hansen MM et al (2008) Genetic and environmental correlates of morphological variation in a marine fish: the case of Baltic Sea herring (*Clupea harengus*). Can J Fish Aquat Sci 65:389–400

Kadereit B, Kumar P, Wang WJ et al (2008) Evolutionarily conserved gene family important for fat storage. Proc Natl Acad Sci USA 105:84–99

Kalbe M, Kurtz J (2006) Local differences in immunocompetence reflect resistance of sticklebacks against the eye fluke *Diplostomum pseudospathaceum*. Parasitology 132:105–116

Kang YS, Kim YM, Park KI et al (2006) Analysis of EST and lectin expressions in hemocytes of Manila clams (*Ruditapes philippinarum*) (Bivalvia: Mollusca) infected with *Perkinsus olseni*. Dev Comp Immunol 30:1119–1131

Kang JH, Kim WJ, Lee WJ (2008) Genetic linkage map of olive flounder, *Paralichthys olivaceus*. Int J Biol Sci 4:143–149

Karlsson S, Renshaw MA, Rexroad CE et al (2008) PCR primers for 100 microsatellites in red drum (*Sciaenops ocellatus*). Mol Ecol Resour 8:393–398

Karsi A, Cao D, Li P et al (2002) Transcriptome analysis of channel catfish (*Ictalurus punctatus*): initial analysis of gene expression and microsatellite-containing cDNAs in the skin. Gene 285:157–168

Kaushik SJ (2005) Besoins et apport en phosphore chez les poissons. INRA Prod Anim 18: 203–208

Kawecki TJ, Ebert D (2004) Conceptual issues in local adaptation. Ecol Lett 7:1225–1241

Khoo SK, Ozaki A, Nakamura F et al (2004) Identification of a novel chromosomal region associated with infectious hematopoietic necrosis (IHN) resistance in rainbow trout. Fish Pathol 39:95–102

Kiessling A, Ruohonen K, Bjørnevik M (2006) Muscle fibre growth and quality in fish. Arch Tierz Dummerstorf 49:137–146

Kim YM, Park K-I, Choi K-S et al (2006) Lectin from the Manila clam *Ruditapes philippinarum* is induced upon infection with the protozoan parasite *Perkinsus olseni*. J Biol Chem 281: 26854–26864

Kim JY, Kim YM, Cho SK et al (2008) Noble tandem-repeat galectin of Manila clam *Ruditapes philippinarum* is induced upon infection with the protozoan parasite *Perkinsus olseni*. Dev Comp Immunol 32:1131–1141

King M (1995) Fisheries biology, assessment and management. Fishing News Books, Oxford

Kingsley DM, Zhu BL, Osoegawa KJ et al (2004) New genomic tools for molecular studies of evolutionary change in three-spined sticklebacks. Behaviour 141:1331–1344

Kjærsgård IVH, Jessen F (2003) Proteome analysis elucidating post-mortem changes in cod (*Gadus morhua*) muscle proteins. J Agric Food Chem 51:3985–3991

Kjærsgård IVH, Nørrelykke MR, Jessen F (2006a) Changes in cod muscle proteins during frozen storage revealed by proteome analysis and multivariate data analysis. Proteomics 6:1606–1618

Kjærsgård IVH, Nørrelykke MR, Baron CP et al (2006b) Identification of carbonylated protein in frozen rainbow trout (*Oncorhynchus mykiss*) fillets and development of protein oxidation during frozen storage. J Agric Food Chem 54:9437–9446

Kleppe K, Ohtsuka E, Kleppe R et al (1971) Studies on polynucleotides. XCVI. Repair replications of short synthetic DNA's as catalyzed by DNA polymerases. J Mol Biol 56:341–361

Kocher TD, Kole C (2008) Genome mapping and genomics in fishes and aquatic animals. Volume 2. Springer-Heidelberg, Berlin

Koehn RK, Milkman R, Mitton JB (1976) Population genetics of marine pelecypods. 4. Selection, migration and genetic differentiationn in blue mussels *Mytilus edulis*. Evolution 30:2–32

Koehn RK, Immermann FW (1981) Biochemical studies of aminopeptidase polymorphism in *Mytilus edulis*. 1. Dependence of enzyme activity on season, tissue, and genotype. Biochem Genet 19:1115–1142

Koljonen ML, Tahtinen J, Saisa M et al (2002) Maintenance of genetic diversity of Atlantic salmon (*Salmo salar*) by captive breeding programmes and the geographic. Aquaculture 212:69–92

Koljonen ML, Pella JJ, Masuda M (2005) Classical individual assignments versus mixture modeling to estimate stock proportions in Atlantic salmon (*Salmo salar*) catches from DNA microsatellite data. Can J Fish Aquat Sci 62:2143–2158

Kono T, Ponpompisit A, Sakai M (2003) The analysis of expressed genes in head kidney of common carp *Cyprinus carpio* L. stimulated with peptidoglycan. Aquaculture 235:37–52

Koskinen MT, Hirvonen H, Landry PA et al (2004) The benefits of increasing the number of microsatellites utilized in genetic population studies: an empirical perspective. Hereditas 141:61–67

Kudo H, Amizuka N, Araki K et al (2004) Zebrafish periostin is required for the adhesion of muscle fiber bundles to the myoseptum and for the differentiation of muscle fibers. Dev Biol 267:473–487

Kuhl H, Beck A, Wozniak G, Canario AVM et al (2010) The European sea bass *Dicentrarchus labrax* genome puzzle: comparative BAC-mapping and low coverage shotgun sequencing. BMC Genetics (in press)

Kumazawa N, Morimoto N (1992) Chemotactic activity of hemocytes derived from a brackish-water clam, *Corbicula japonica*, to *Vibrio parahaemolyticus* and *Escherichia coli* strains. J Vet Med Sci 5:851–855

Kwok C, Korn RM, Davis ME et al (1998) Characterization of whole genome radiation hybrid mapping resources for non-mammalian vertebrates. Nucleic Acids Res 26:3562–3566

La Peyre JF, Yarnall HÁ, Faisal M (1996) Contribution of *Perkinsus marinus* extracellular products in the infection of eastern oysters (*Crassostrea virginica*). J Invertebr Pathol 68:312–313

Labreuche Y, Soudant P, Gonçalves M et al (2006a) Effects of extracellular products from the pathogenic *Vibrio aestuarianus* strain 01/32 on lethality and cellular immune responses of the oyster *Crassostrea gigas*. Dev Comp Immunol 30:367–379

Labreuche Y, Lambert C, Soudant P et al (2006b) Cellular and molecular hemocyte responses of the Pacific oyster, *Crassostrea gigas*, following bacterial infection with Vibrio *aestuarianus* strain 01/32. Microbes Infect 8:2715–2724

Lallias D, Arzul I, Heurtebise S et al (2008) Bonamia-ostreae induced mortalities in one-year old European flat oysters *Ostrea edulis*: experimental infection by cohabitation challenge. Aquat Living Resour 21:423–439

Lam N, Araneda C, Díaz NF et al (2007) Physical mapping of SCAR-RAPD markers associated to spawning date and flesh color traits in cultivated Coho salmon (*Oncorhynchus kisutch*). Aquaculture 272:S282

Lambert C, Soudant P, Choquet G et al (2003) Measurement of *Crassostrea gigas* hemocyte oxidative metabolism by flow cytometry and the inhibiting capacity of pathogenic vibrios. Fish Shellfish Immunol 15:225–240

Lane E, Birkbeck TH (1999) Toxicity to bacteria towards haemocytes of *Mytilus edulis*. Aquat Living Resour 12:343–350

Lang P, Langdon CJ, Camara MD (2008) Predicting the resistance of adult pacific oysters (*Crassostrea gigas*) to summer mortality. J Shellfish Res 27:470

Langefors J, Lohm M, Grahn O (2001) Association between major histocompatibility complex class IIB alleles and resistance to *Aeromonas salmonicida* in Atlantic salmon. Proc R Soc B-Biol Sci 268:479–485

Larsen PF, Nielsen EE, Williams TD et al (2007) Adaptive differences in gene expression in European flounder (*Platichthys flesus*). Mol Ecol 16:4674–4683

Larsson LC, Laikre L, Palm S et al (2007) Concordance of allozyme and microsatellite differentiation in a marine fish, but evidence of selection at a microsatellite locus. Mol Ecol 16:1135–1147

Launey S, Barre M, Gerard A et al (2001) Population bottleneck and effective size in Bonamia ostreae-resistant populations of *Ostrea edulis* as inferred by microsatellite markers. Genet Res 78:259–270

Le Chevalier P, Le Boulay C, Paillard C (2003) Characterization by restriction fragment length polymorphism and plasmid profiling of *Vibrio tapetis* strains. J Basic Microbiol 43:414–422

Le Roux F, Audemard C, Barnaud A et al (1999) DNA probes as potential tools for the detection of *Marteilia refringens*. Mar Biotechnol 1:588–597

Leaver MJ, Bautista JM, Björnsson BT et al (2008) Towards fish lipid nutrigenomics: current state and prospects for fin-fish aquaculture. Rev Fish Sci 16:73–94

Lehrer SB, Ayuso R, Reese G (2003) Seafood allergy and allergens: a review. Mar Biotechnol 5:339–348

Lelong C, Badariotti F, Le Quéré H et al (2007) Cg-TGF-b, a TGF-b/activin homologue in the Pacific oyster *Crassostrea gigas*, is involved in immunity against Gram-negative microbial infection. Dev Comp Immunol 31:30–38

Lemaire C, Allegrucci G, Naciri M et al (2000) Do discrepancies between microsatellite and allozyme variation reveal differential selection between sea and lagoon in the sea bass (*Dicentrarchus labrax*)?. Mol Ecol 9:457–467

Lemaitre B, Kromer-Metzger E, Michaut L et al (1995) A recessive mutation, immune deficiency (imd), defines two distinct control pathways in the *Drosophila* host defense. Proc Natl Acad Sci USA 92:9465–9469

Levesque HM, Shears MA, Fletcher GL et al (2008) Myogenesis and muscle metabolism in juvenile Atlantic salmon (*Salmo salar*) made transgenic for growth hormone. J Exp Biol 211:128–137

Li G, Hubert S, Bucklin K et al (2003a) Characterization of 79 microsatellite DNA markers in the Pacific oyster *Crassostrea gigas*. Mol Ecol Notes 3:228–232

Li X, Field C, Doyle R (2003b) Estimation of additive genetic variance components in aquaculture populations selectively pedigreed by DNA fingerprinting. Biom J 45:61–72

Li Q, Kijima A (2006) Microsatellite analysis of gynogenetic families in the Pacific oyster, *Crassostrea gigas*. J Exp Mar Biol Ecol 331:1–8

Li P, Peatman E, Wang S et al (2007) Towards the ictalurid catfish transcriptome: generation and analysis of 31,215 catfish ESTs. BMC Genomics 8:177

Li C, Ni D, Song L et al (2008) Molecular cloning and characterization of a catalase gene from Zhikong scallop *Chlamys farreri*. Fish Shellfish Immunol 24:26–34

Liu N, Chen L, Wang S et al (2005a) Comparison of single-nucleotide polymorphisms and microsatellites in inference of population structure. BMC Genet 6:S26

Liu YG, Chen SL, Li BF et al (2005b) Analysis of genetic variation in selected stocks of hatchery flounder, *Paralichthys olivaceus*, using AFLP markers. Biochem Syst Ecol 33:993–1005

Liu XD, Liu X, Guo X et al (2006) A preliminary genetic linkage map of the pacific abalone *Haliotis discus hannai* Ino. Mar Biotechnol 8:386–397

Liu XD, Liu X, Zhang GF (2007) Identification of quantitative trait loci for growth-related traits in the Pacific abalone *Haliotis discus hannai Ino*. Aquac Res 38:789–797

Liu Z, Li RW, Waldbieser GC (2008) Utilization of microarray technology for functional genomics in ictalurid catfish. J Fish Biol 72:2377–2390

Lohm J, Grahn M, Langefors A et al (2002) Experimental evidence for major histocompatibility complex allele-specific resistance to a bacterial infection. Proc R Soc B-Biol Sci 269:2029–2033

Lopez-Castejon G, Sepulcre MP, Roca FJ et al (2007) The type II interleukin-1 receptor (IL-1RII) of the bony fish gilthead seabream *Sparus aurata* is strongly induced after infection and tightly regulated at transcriptional and post-transcriptional levels. Mol Immunol 44:2772–2780

Lucassen M, Koschnick N, Eckerle L et al (2006) Mitochondrial mechanisms of cold adaptation in cod (*Gadus morhua* L.) populations from different climatic zones. J Exp Biol 209:2462–2471

Luikart G, England PR, Tallmon D et al (2003) The power and promise of population genomics: From genotyping to genome typing. Nat Rev Genet 4:981–994

Lynch M, Walch B (1998) Genetics and analysis of quantitative traits. Sinauer Associates Inc, Sunderland, Massachusetts

Ma H, Mai K, Xu W et al (2005) Molecular cloning of a2-macroglobulin in sea scallop *Chlamys farreri* (Bivalvia, Mollusca). Fish Shellfish Immunol 18:345–349

MacKenzie S, Balasch JC, Novoa B et al (2008) Comparative analysis of the acute response of the trout, *O. mykiss*, head kidney to in vivo challenge with virulent and attenuated infectious hematopoietic necrosis virus and LPS-induced inflammation. BMC Genomics 9:141

Mackie IM (1993) The effects of freezing on flesh proteins. Food Rev Int 9:575–610

Maes GE, Pujolar JM, Hellemans B et al (2006) Evidence for isolation by time in the European eel (*Anguilla anguilla* L.). Mol Ecol 15:2095–2107

Mäkinen HS, Cano JM, Merilä J (2008) Identifying footprints of directional and balancing selection in marine and freshwater three-spined stickleback (*Gasterosteus aculeatus*) populations. Mol Ecol 17:3565–3582

Malek RL, Sajadi H, Abraham J et al (2004) The effects of temperature reduction on gene expression and oxidative stress in skeletal muscle from adult zebrafish. Comp Biochem Physiol C-Toxicol Pharmacol 138:363–373

Manel S, Gaggiotti OE, Waples RS (2005) Assignment methods: matching biological questions with appropriate techniques. Trends Ecol Evol 20:136–142

Margulies M, Egholm M, Altman WE et al (2005) Genome sequencing in microfabricated high-density picolitre reactors. Nature 437:376–380

Marteinsdottir G, Begg GA (2002) Essential relationships incorporating the influence of age, size and condition on variables required for estimation of reproductive potential in Atlantic cod *Gadus morhua*. Mar Ecol-Prog Ser 235:235–256

Martin SAM, Cash P, Blaney S et al (2001) Proteome analysis of rainbow trout (*Oncorhynchus mykiss*) liver proteins during short term starvation. Fish Physiol Biochem 24:259–270

Martin SAM, Vilhelmsson O, Médale F et al (2003) Proteomic sensitivity to dietary manipulations in rainbow trout. Biochim Biophys Acta 1651:17–29

Martin SAM, Blaney SC, Houlihan DF et al (2006) Transcriptome response following administration of a live bacterial vaccine in Atlantic salmon (*Salmo salar*). Mol Immunol 43:1900–1911

Martin SAM, Collet B, Mackenzie S et al (2008) Genomic tools for examining immune gene function in salmonid fish. Rev Fish Sci 16:112–118

Martinez I, Friis TJ (2004) Application of proteome analysis to seafood authentication. Proteomics 4:347–354

Martinez I, James D, Loréal H (2005) Application of modern analytical techniques to ensure seafood safety and authenticity. FAO Fish Tech Paper 455, FAO, Rome

Martinez I, Šližytė R, Daukšas E (2007) High resolution two-dimensional electrophoresis as a tool to differentiate wild from farmed cod (*Gadus morhua*) and to assess the protein composition of klipfish. Food Chem 102:504–510

Martinez V, di Giovanni S (2007) Breeding programmes of scallops: effect of self-fertilization when estimating genetic parameters. Aquaculture 272:S287

Matsuda M, Kawato N, Asakawa S et al (2001) Construction of a BAC library derived from the inbred Hd-rR strain of the teleost fish, *Oryzias latipes*. Genes Genet Syst 76:61–63

Matthews SJ, Ross NW, Lall SP et al (2006) Astaxanthin binding protein in Atlantic salmon. Comp Biochem Physiol B-Biochem Mol Biol 144:206–214

Matzinger P (2002) The danger model: A renewed sense of self. Science 296:301–305

McAvoy ES, Wood AR, Gardeur JN (2008) Development and evaluation of microsatellite markers for identification of individual Greenshell^{TM} mussels (*Perna canaliculus*) in a selective breeding programme. Aquaculture 274:41–48

McDonald JH, Verrelli BC, Geyer LB (1996) Lack of geographic variation in anonymous nu-clear polymorphisms in the American oyster, *Crassostrea virginica*. Mol Biol Evol 13:1114–1118

McDonald GJ, Danzmann RG, Ferguson MM (2004) Relatedness determination in the absence of pedigree information in three cultured strains of rainbow trout. Aquaculture 233:65–78

McKay SD, Schnabel RD, Murdoch BM et al (2007) Construction of bovine whole-genome radiation hybrid and linkage maps using high-throughput genotyping. Anim Genet 38:120–125

McKinnon JS, Rundle HD (2002) Speciation in nature: the threespine stickleback model systems. Trends Ecol Evol 17:480–488

Medzhitov R, Janeway CA Jr (2002) Decoding the patterns of self and nonself by the innate immune system. Science 296:298–300

Meyers BC, Scalabrin S, Morgante M (2004) Mapping and sequencing complex genomes: let's get physical!. Nat Rev Genet 5:578–588

Milinski M, Griffiths S, Wegner KM (2005) Mate choice decisions of stickleback females predictably modified by MHC peptide ligands. Proc Natl Acad Sci USA 102:4414–4418

Milkman R, Koehn RK (1977) Temporal variation in relationship between size, numbers, and allele frequency in a population of *Mytilus edulis*. Evolution 31:103–115

Min B, Ahn DU (2005) Mechanism of lipid peroxidation in meat and meat products: a review. Food Sci Biotechnol 14:152–163

Mitta G, Vandenbulcke F, Hubert F et al (1999a) Mussel defensins are synthesised and processed in granulocytes then released into the plasma after bacterial challenge. J Cell Sci 112:4233–4242

Mitta G, Hubert F, Noël T et al (1999b) Myticin, a novel cysteine-rich antimicrobial peptide isolated from haemocytes and plasma of the mussel *Mytilus galloprovincialis*. Eur J Biochem 265:71–78

Mitta G, Vandenbulcke F, Hubert F et al (2000a) Involvement of mytilins in mussel antimicrobial defense. J Biol Chem 275:12954–12962

Mitta G, Vandenbulcke F, Roch P (2000b) Original involvement of antimicrobial peptides in mussel innate immunity. FEBS Lett 486:185–190

Moen T, Agresti JJ, Cnaani A et al (2004) A genome scan of a four-way tilapia cross supports the existence of a quantitative trait locus for cold tolerance on linkage group 23. Aquac Res 35:893–904

Moen T, Sonesson AK, Hayes B et al (2007) Mapping of a quantitative trait locus for resistance against infectious salmon anaemia in Atlantic salmon (*Salmo Salar*): comparing survival analysis with analysis on affected/resistant data. BMC Genet 8:53

Moen T, Hayes B, Baranski M et al (2008a) A linkage map of the Atlantic salmon (*Salmo salar*) based on EST-derived SNP markers. BMC Genomics 9:223

Moen T, Hayes B, Nilsen F et al (2008b) Identification and characterisation of novel SNP markers in Atlantic cod: Evidence for directional selection. BMC Genet 9:18

Moghadam HK, Poissant J, Fotherby H et al (2007) Quantitative trait loci for body weight, condition factor and age at sexual maturation in Arctic charr (*Salvelinus alpinus*): comparative analysis with rainbow trout (*Oncorhynchus mykiss*) and Atlantic salmon (*Salmo salar*). Mol Genet Genomics 277:647–661

Mommsen TP (2001) Paradigms of growth in fish. Comp Biochem Phys B-Biochem Mol Biol 129:207–219

Montagnani C, Le Roux F, Berthe F et al (2001) Cg-TIMP, an inducible tissue inhibitor of metalloproteinase from the Pacific oyster *Crassostrea gigas* with a potential role in wound healing and defense mechanisms. FEBS Lett 500:64–70

Montagnani C, Kappler C, Reichhart JM et al (2004) Cg-Rel, the first Rel/NF-kB homolog characterized in a mollusk, the Pacific oyster *Crassostrea gigas*. FEBS Lett 561:75–82

Montagnani C, Avarre JC, de Lorgeril J et al (2007) First evidence of the activation of Cg-timp, an immune response component of Pacific oysters, through a damage-associated molecular pattern pathway. Dev Comp Immunol 31:1–11

Montagnani C, Labreuche Y, Escoubas JM (2008) Cg-IkB, a new member of the IkB protein family characterized in the Pacific oyster *Crassostrea gigas*. Dev Comp Immunol 32: 182–190

Montes JF, Durfort M, Garcia-Valero J (1995a) Cellular defense mechanisms of the clam *Tapes semidecussatus* against infection by the protozoan parasite *Perkinsus* sp. Cell Tissue Res 279:529–538

Montes JF, Durfort M, Garcia-Valero J (1995b) Characterization and localization of an Mr 225 kDa polypeptide specifically involved in the defence mechanisms of the clam *Tapes semidecussatus*. Cell Tissue Res 280:27–37

Montes JF, Dufort M, Garcia-Valero J (1996) When the venerid clam *Tapes decussatus* is parasitized by the protozoan *Perkinsus* sp. it synthesizes a defensive polypeptide that is closely related to p225. Dis Aquat Organ 26:149–157

Morin PA, Luikart G, Wayne RK et al (2003) SNPs in ecology, evolution and conservation. Trends Ecol Evol 19:208–216

Morzel M, Verrez-Bagnis V, Arendt EK et al (2000) Use of two-dimensional electrophoresis to evaluate proteolysis in salmon *Salmo salar* muscle as affected by a lactic fermentation. J Agric Food Chem 48:239–244

Morzel M, Chambon C, Lefèvre F et al (2006) Modifications of trout (*Oncorhynchus mykiss*) muscle proteins by pre-slaughter activity. J Agric Food Chem 54:2997–3001

Mullis K, Faloona F (1987) Specific synthesis of DNA in vitro via a polymerase-catalyzed chain reaction. Methods Enzymol 155:335–350

Naciri-Graven Y, Launey S, Lebayon N et al (2000) Influence of parentage upon growth in *Ostrea edulis*: evidence for inbreeding depression. Genet Res 46:159–168

Navas JI, Castillo MC, Vera P et al (1992) Principal parasites observed in clams, *Ruditapes decussatus* (L.), *Ruditapes philippinarum* (Adams et Reeve), *Venerupis pullastra* (Montagu) and *Venerupis aureus* (Gmelin), from the Huelva coast (S.W. Spain). Aquaculture 107:193–199

Ni D, Song L, Wu L et al (2007) Molecular cloning and mRNA expression of peptidoglycan recognition protein (PGRP) gene in bay scallop (*Argopecten irradians*, Lamarck 1819). Dev Comp Immunol 31:548–558

Nielsen EE, Hansen MM, Schmidt C et al (2001) Determining the population origin of individual cod in the Northeast Atlantic. Nature 413:272

Nielsen EE, Hansen MM, Ruzzante DE et al (2003) Evidence of a hybrid-zone in Atlantic cod (*Gadus morhua*) in the Baltic and the Danish Belt Sea, revealed by individual admixture analysis. Mol Ecol 12:1497–1508

Nielsen EE, Hansen MM, Meldrup D (2006) Evidence of microsatellite hitch-hiking selection in Atlantic cod (*Gadus morhua* L.): Implications for inferring population structure in non-model organisms. Mol Ecol 15:3219–3229

Nielsen EE, MacKenzie BR, Magnussen E et al (2007) Historical analysis of Pan I in Atlantic cod (*Gadus morhua*): temporal stability of allele frequencies in the southeastern part of the species distribution. Can J Fish Aquat Sci 64:1448–1455

Nielsen EE, Hansen MM (2008) Waking the dead: the value of population genetic analyses of historical samples. Fish Fish 9:450–461

Nikula R, Strelkov P, Vainola R (2008) A broad transition zone between an inner Baltic hybrid swarm and a pure North Sea subspecies of *Macoma balthica* (Mollusca, Bivalvia). Mol Ecol 17:1505–1522

Nordstrom JL, Vickery MC, Blackstone GM et al (2007) Development of a multiplex real-time PCR assay with an internal amplification control for the detection of total and pathogenic *Vibrio parahaemolyticus* bacteria in oysters. Appl Environ Microbiol 73:5840–5847

Norris AT, Bradley DG, Cunningham EP (1999) Microsatellite genetic variation between and within farmed and wild Atlantic salmon (*Salmo salar*) populations. Aquaculture 180: 247–264

Norris AT, Bradley DG, Cunningham EP (2000) Parentage and relatedness determination in farmed Atlantic salmon (*Salmo salar*) using microsatellite markers. Aquaculture 182:73–83

Nottage AS, Birkbeck TH (1990) Interactions between different strains of *Vibrio alginolyticus* and hemolymph fractions from adult *Mytilus edulis*. J Invertebr Pathol 56:15–19

Nürnberger T, Brunner F, Kemmerling B et al (2004) Innate immunity in plants and animals: striking similarities and obvious differences. Immunol Rev 198:249–266

Nygaard V, Liu F, Holden M et al (2008) Validation of oligoarrays for quantitative exploration of the transcriptome. BMC Genomics 9:258

Oleksiak MF, Churchill GA, Crawford DL (2002) Variation in gene expression within and among natural populations. Nat Genet 32:261–266

Oliver JL, Gaffney PM, Allen SK et al (2000) Protease inhibitory activity in selectively bred families of eastern oysters. J Aquat Anim Health 12:136–145

Ommen B, Groten JP (2004) Nutrigenomics in efficacy and safety evaluation of food components. World Rev Nutr Diet 93:134–152

Orr HA (2005) The genetic theory of adaptation: a brief history. Nat Rev genet 6:119–127

Ozaki A, Sakamoto T, Khoo S et al (2001) Quantitative trait loci (QTLs) associated with resistance/susceptibility to infectious pancreatic necrosis virus (IPNV) in rainbow trout (*Oncorhynchus mykiss*). Mol Genet Genomics 265:23–31

Paillard C, Le Roux F, Borrego JJ (2004) Bacterial disease in marine bivalves, a review of recente studies: Trends and evolution. Aquat Living Resour 17:477–498

Palumbi SR (2004) Marine reserves and ocean neighborhoods: The spatial scale of marine populations and their management. Annu Rev Environ Resour 29:31–68

Pampoulie C, Jorundsdottir TD, Steinarsson A et al (2006) Genetic comparison of experimental farmed strains and wild Icelandic populations of Atlantic cod (*Gadus morhua* L.). Aquaculture 261:556–564

Panicker G, Vickery MC, Bej AK (2004) Multiplex PCR detection of clinical and environmental strains of *Vibrio vulnificus* in shellfish. Can J Microbiol 50:911–922

Panserat S, Ducasse-Cabanot S, Plagnes-Juan E et al (2008a) Dietary fat level modifies the expression of hepatic genes in juvenile rainbow trout (*Oncorhynchus mykiss*) as revealed by microarray analysis. Aquaculture 275:235–241

Panserat S, Kolditz C, Richard N et al (2008b) Hepatic gene expression profiles in juvenile rainbow trout (*Oncorhynchus mykiss*) fed fishmeal or fish oil-free diets. Br J Nutr 100:953–967

Park CC, Ahn S, Bloom JS et al (2008) Fine mapping of regulatory loci for mammalian gene expression using radiation hybrids. Nat Genet 40:421–429

Pauly D, Christensen V, Dalsgaard J et al (1998) Fishing down marine food webs. Science 279:860–863

Peatman E, Baoprasertkul P, Terhune J et al (2007) Expression analysis of the acute phase response in channel catfish (*Ictalurus punctatus*) after infection with a Gram-negative bacterium. Dev Comp Immunol 31:1183–1196

Peatman E, Terhune J, Baoprasertkul P et al (2008) Microarray analysis of gene expression in the blue catfish liver reveals early activation of the MHC class I pathway after infection with *Edwardsiella ictaluri*. Mol Immunol 45:553–566

Peichel CL, Nereng KS, Ohgi KA et al (2001) The genetic architecture of divergence between threespine stickleback species. Nature 414:901–905

Pella J, Masuda M (2001) Bayesian methods for analysis of stock mixtures from genetic characters. Fish Bull 99:151–167

Penna MS, Khan M, French RA (2001) Development of a multiplex PCR for the detection of *Haplosporidium nelsoni*, *Haplosporidium costale* and *Perkinsus marinus* in the eastern oyster (*Crassostrea virginica*, Gmelin, 1971). Mol Cell Probes 15:385–390

Pieniak Z, Verbeke W, Brunsø K et al (2006) Consumer knowledge and interest in information about fish. In: Luten JB, Jacobsen C, Bekaert K, Sæbø A, Oehlenschläger J (eds) Seafood Research from Fish to Dish. Wageningen Academic Publishers, Wageningen

Place SP, O'Donnell MJ, Hofmann GE (2008) Gene expression in the intertidal mussel *Mytilus californianus*: physiological response to environmental factors on a biogeographic scale. Mar Ecol – Prog Ser 356:1–14

Pogson GH, Fevolden SE (2003) Natural selection and the genetic differentiation of coastal and Arctic populations of the Atlantic cod in northern Norway: a test involving nucleotide sequence variation at the pantophysin (PanI) locus. Mol Ecol 12:63–74

Pollock DD, Bergman A, Feldman MW et al (1998) Microsatellite behavior with range constraints: Parameter estimation and improved distances for use in phylogenetic reconstruction. Theor Popul Biol 53:256–271

Porta J, Porta JM, Martinez-Rodriguez G et al (2006) Genetic structure and genetic relatedness of a hatchery stock of Senegal sole (*Solea senegalensis*) inferred by microsatellites. Aquaculture 251:46–55

Potasman I, Paz A, Odeh M (2002) Infectious outbreaks associated with bivalve shellfish consumption: a worldwide perspective. Clin Infect Dis 35:921–928

Prieur G, Mevel G, Nicolas JL et al (1990) Interactions between bivalve molluscs and bacteria in the marine environment. Oceanogr Mar Biol Annu Rev 28:277–352

Primmer CR, Koskinen MT, Piironen J (2000) The one that did not get away: individual assignment using microsatellite data detects a case of fishing competition fraud. Proc R Soc B-Biol Sci 267:1699–1704

Prudence M, Moal J, Boudry P et al (2006) An amylase gene polymorphism is associated with growth differences in the Pacific cupped oyster *Crassostrea gigas*. Anim Genet 37:348–351

Prunet P, Cairns MT, Winberg S et al (2008) Functional genomics of stress responses in fish. Rev Fish Sci 16:157–166

Pruzzo C, Gallo G, Canesi L (2005) Persistence of vibrios in marine bivalves: the role of interactions with haemolymph components. Environ Microbiol 7:761–772

Purcell MK, Nichols KM, Winton JR et al (2006) Comprehensive gene expression profiling following DNA vaccination of rainbow trout against infectious hematopoietic necrosis virus. Mol Immunol 43:2089–2106

Qiu L, Song L, Xu W et al (2007a) Molecular cloning and expression of a Toll receptor gene homologue from Zhikong Scallop, *Chlamys farreri*. Fish Shellfish Immunol 22: 451–466

Qiu L, Song L, Xu W et al (2007b) Identification and characterization of a myeloid differentiation factor 88 (MyD88) cDNA from Zhikong scallop *Chlamys farreri*. Fish Shellfish Immunol 23:614–623

Raeymaekers JAM, Van Houdt JKJ, Larmuseau MHD et al (2007) Divergent selection as revealed by PST and QTL-based FST in three-spined stickleback (*Gasterosteus aculeatus*) populations along a coastal-inland gradient. Mol Ecol 16:891–905

Raida MK, Buchmann K (2007) Temperature-dependent expression of immune-relevant genes in rainbow trout following *Yersinia ruckeri* vaccination. Dis Aquat Organ 77:41–52

Rasmussen RS, Morrissey MT (2008) DNA-based methods for the identification of commercial fish and seafood species. Compr Rev Food Sci Food Saf 7:280–295

Raymond M, Vaanto RL, Thomas F et al (1997) Heterozygote deficiency in the mussel *Mytilus edulis* species complex revisited. Mar Ecol – Prog Ser 156:225–237

Reece KS, Bushek D, Graves JE (1997) Molecular markers for population genetic analysis of *Perkinsus marinus*. Mol Mar Biol Biotechnol 6:197–206

Reese G, Viebranz J, Leong-Kee SM et al (2005) Reduced allergenic potency of VR9-1, a mutant of the major shrimp allergen Pen a1 (Tropomyosin). J Immunol 175:8354–8364

Reid DP, Szanto A, Glebe B et al (2005) QTL for body weight and condition factor in Atlantic salmon (*Salmo salar*): Comparative analysis with rainbow trout (*Oncorhynchus mykiss*) and Arctic charr (*Salvelinus alpinus*). Heredity 94:166–172

Reid DP, Smith CA, Rommens M et al (2007) A genetic linkage map of Atlantic halibut (*Hippoglossus hippoglossus* L.). Genetics 77:1193–1205

Rengmark AH, Slettan A, Skaala O et al (2006) Genetic variability in wild and farmed Atlantic salmon (*Salmo salar*) strains estimated by SNP and microsatellites. Aquaculture 253: 229–237

Rescan PY, Montfort J, Ralliere C et al (2007) Dynamic gene expression in fish muscle during recovery growth induced by a fasting-refeeding schedule. BMC Genomics 8:438

Reusch TBH, Wood TE (2007) Molecular ecology of global change. Mol Ecol 19:3973–3992

Rexroad CE, Rodriguez MF, Coulibaly I et al (2005) Comparative mapping of expressed sequence tags containing microsatellites in rainbow trout (*Oncorhynchus mykiss*). BMC Genomics 6:54

Rise ML, Jones SR, Brown GD et al (2004) Microarray analyses identify molecular biomarkers of Atlantic salmon macrophage and hematopoietic kidney response to *Piscirickettsia salmonis* infection. Physiol Genomics 20:21–35

Rise ML, Douglas SE, Sakhrani D et al (2006) Multiple microarray platforms utilized for hepatic gene expression profiling of GH transgenic Coho salmon with and without ration restriction. J Mol Endocrinol 37:259–282

Roberge C, Einum S, Guderley H et al (2006) Rapid parallel evolutionary changes of gene transcription profiles in farmed Atlantic salmon. Mol Ecol 15:9–20

Roberge C, Normandeau E, Einum S et al (2008) Genetic consequences of interbreeding between farmed and wild Atlantic salmon: insights from the transcriptome. Mol Ecol 17:314–324

Rodriguez F, Navas JI (1995) A comparison of gill and hemolymph assays for the thioglycolate diagnosis of *Perkinsus atlanticus* (Apicomplexa, Perkinsea) in clams, *Ruditapes decussatus*, (L.) and *Ruditapes philipinarum* (Adams et Reeve). Aquaculture 132:145–152

Rodriguez MF, LaPatra S, Williams S et al (2004) Genetic markers associated with resistance to infectious hematopoietic necrosis in rainbow trout and steelhead trout (*Oncorhynchus mykiss*) backcrosses. Aquaculture 241:93–115

Roessig JM, Woodley CM, Cech JJ et al (2004) Effects of global climate change on marine and estuarine fishes and fisheries. Rev Fish Biol Fisher 14:251–275

Rogers SM, Bernatchez L (2005) Integrating QTL mapping and genomic scans towards the characterization of candidate loci under parallel directional selection in the lake whitefish (*Coregonus clupeaformis*). Mol Ecol 14:351–361

Rogers SM, Bernatchez L (2007) The genetic architecture of ecological speciation and the association with signatures of selection in natural lake whitefish (*Coregonus* sp. Salmonidae) species pairs. Mol Biol Evol 24:1423–1438

Rogers SM, Isabel N, Bernatchez L (2007) Linkage maps of the dwarf and normal lake whitefish (*Coregonus clupeaformis*) species complex and their hybrids reveal the genetic architecture of population divergence. Genetics 175:1–24

Rooker JR, Bremer JRA, Block BA et al (2007) Life history and stock structure of Atlantic bluefin tuna (*Thunnus thynnus*). Rev Fish Sci 15:263–310

Rucker RR (1966) Redmouth disease of rainbow trout (*Salmo gairdneri*). Bull Off Int Epizoot 65:825–830

Ruzzante DE, Mariani S, Bekkevold D et al (2006) Biocomplexity in a highly migratory pelagic marine fish, Atlantic herring. Proc R Soc B-Biol Sci 273:1459–1464

Ryynanen HJ, Primmer CR (2004) Distribution of genetic variation in the growth hormone 1 gene in Atlantic salmon (*Salmo salar*) populations from Europe and North America. Mol Ecol 13:3857–3869

Saeed S, Howell NK (2002) Effect of lipid oxidation and frozen storage on muscle proteins of Atlantic mackerel (*Scomber scombrus*). J Sci Food Agric 82:579–586

Sagristà E, Durfort M, Azevedo C (1995) *Perkinsus* sp. (Phylum Apicomplexa) in Mediterranean clam *Ruditapes semidecussatus*: ultrastructural observations of the cellular response of the host. Aquaculture 132:153–160

Saha MR (2005) The role of muscle proteins in the retention of carotenoid in Atlantic salmon flesh. PhD Thesis Dalhousie University, Canada

Saito M, Higuichi T, Suzuki H et al (2006) Post-mortem changes in gene expression of the muscle tissue of rainbow trout, *Oncorhynchus mykiss*. J Agric Food Chem 54:9417–9421

Sakai M, Kono T, Savan R (2005) Identification of expressed genes in carp (*Cyprinus carpio*) head kidney cells after in vitro treatment with immunostimulants. Dev Biol 121:45–51

Salem M, Kenney PB, Rexroad IIICE et al (2006) Microarray gene expression analysis in atrophying rainbow trout muscle: a unique nonmammalian muscle degradation model. Physiol Genomics 28:33–45

Salem M, Silverstein J, Rexroad IIICE et al (2007) Effect of starvation on global gene expression and proteolysis in rainbow trout (*Oncorhynchus mykiss*). BMC Genomics 8:328

Samuelsen OB, Nerland AH, Jorgensen T et al (2006) Viral and bacterial diseases of Atlantic cod *Gadus morhua*, their prophylaxis and treatment: a review. Dis Aquat Organ 71:239–254

Sanetra M, Meyer A (2008) A microsatellite-based genetic linkage map of the cichlid fish, *Astatotilapia burtoni* and a comparison of genetic architectures among rapidly speciating cichlids. Genetics doi: 10.1534/genetics.108.089367

Sarropoulou E, Power DM, Magoulas A et al (2005a) Comparative analysis and characterization of expressed sequence tags in gilthead sea bream (*Sparus aurata*) liver and embryos. Aquaculture 243:69–81

Sarropoulou E, Kotoulas G, Power DM et al (2005b) Gene expression profiling of gilthead sea bream during early development and detection of stress-related genes by the application of cDNA microarray technology. Physiol Genomics 23:182–191

Sarropoulou E, Franch R, Louro B et al (2007) A gene-based radiation hybrid map of the gilthead sea bream *Sparus aurata* refines and exploits conserved synteny with *Tetraodon nigroviridis*. BMC Genomics 8:44

Sarropoulou E, Nousdili D, Magoulas A et al (2008) Linking the genomes of nonmodel teleosts through comparative genomics. Mar Biotechnol 10:227–233

Sarropoulou E, Sepulcre P, Poisa-Beiro L et al (2009) Profiling of infection specific mRNA transcripts of the European sea bass *Dicentrarchus labrax*. BMC Genomics 10:157

Sauvage C, Bierne N, Lapègue S et al (2007) Single-nucleotide polymorphisms and their relationship to codon usage bias in the Pacific oyster *Crassostrea gigas*. Gene 406:13–22

Scheffer M, Carpenter S, de Young B (2005) Cascading effects of overfishing marine systems. Trends Ecol Evol 20:579–581

Schena M, Shalon D, Davis RW et al (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science 270:467–470
Schiavone R, Zilli L, Storelli C et al (2008) Identification by proteome analysis of muscle proteins in sea bream (*Sparus aurata*). Eur Food Res Technol 227:1403–1410
Schlötterer C (2002) A microsatellite-based multilocus screen for the identification of local selective sweeps. Genetics 160:753–763
Schlötterer C (2004) The evolution of molecular markers – just a matter of fashion?. Nat Rev Genet 5:63–69
Schluter D (1995) Adaptive radiation in sticklebacks: Trade-offs in feeding performance and growth. Ecology 76:82–90
Schwab KJ, Neill FH, Le Guyader F et al (2001) Development of a reverse transcription-PCR-DNA enzyme immunoassay for detection of Norwalk-like viruses and hepatitis A virus in stool and shellfish. Appl Environ Microbiol 67:742–749
Sekino M, Hara M, Taniguchi N (2002) Loss of microsatellite and mitochondrial DNA variation in hatchery strains of Japanese flounder *Paralichthys olivaceus*. Aquaculture 213:101–122
Sekino M, Hara M (2007) Linkage maps for the Pacific abalone (genus *Haliotis*) based on microsatellite DNA markers. Genetics 175:945–958
Senger F, Priat C, Hitte C et al (2006) The first radiation hybrid map of a perch-like fish: the gilthead seabream (*Sparus aurata* L). Genomics 87:793–800
Seo J-K, Crawford JM, Stone KL et al (2005) Purification of a novel arthropod defensin from the American oyster, *Crassostrea virginica*. Biochem Biophys Res Commun 338:1998–2004
Sepulcre MP, Sarropoulou E, Kotoulas G et al (2007) *Vibrio anguillarum* evades the immune response of the bony fish sea bass (*Dicentrarchus labrax* L.) through the inhibition of leukocyte respiratory burst and down-regulation of apoptotic caspases. Mol Immunol 44:3751–3757
Serapion J, Kucuktas H, Feng J et al (2004a) Bioinformatic mining of type I microsatellites from expressed sequence tags of channel catfish (*Ictalurus punctatus*). Mar Biotechnol 6:364–377
Serapion J, Waldbieser GC, Wolters W et al (2004b) Development of type I markers in channel catfish through intron sequencing. Anim Genet 35:463–466
Shangkuan YH, Show YS, Wang TM (1995) Multiplex polymerase chain reaction to detect toxigenic *Vibrio cholerae* and to biotype *Vibrio cholerae* O1. J Appl Bacteriol 79:264–273
Shapiro MD, Bell MA, Kingsley DM (2006) Parallel genetic origins of pelvic reduction in vertebrates. Proc Natl Acad Sci USA 103:13753–13758
Sick K (1965) Haemoglobin polymorphism of cod in the Baltic and the Danish Belt Sea. Hereditas 54:19–48
Sigholt T, Erikson U, Rustad T et al (1997) Handling stress and storage temperature affect meat quality of farmed-raised Atlantic salmon (*Salmo salar*). J Food Sci 62:898–905
Silverman N, Maniatis T (2001) NF-kB signaling pathways in mammalian and insect innate immunity. Genes Dev 15:2321–2342
Sinclair M (1988) Marine populations: An essay on population regulation and speciation. University of Washington Press, Seattle
Smith CT, Antonovich A, Templin WD et al (2007) Impacts of marker class bias relative to locus-specific variability on population inferences in Chinook salmon: A comparison of single-nucleotide polymorphisms with short tandem repeats and allozymes. Trans Am Fish Soc 136:1674–1687
Smith CT, Seeb LW (2008) Number of alleles as a predictor of the relative assignment accuracy of short tandem repeat (STR) and single-nucleotide-polymorphism (SNP) baselines for chum salmon. Trans Am Fish Soc 137:751–762
Solé-Cava AM, Thorpe JP (1991) High levels of genetic variation in natural populations of ma-rine lower invertebrates. Biol J Linnean Soc 44:65–80
Somorjai IML, Danzmann RG, Ferguson MM (2003) Distribution of temperature tolerance quantitative trait loci in Arctic charr (*Salvelinus alpinus*) and inferred homologies in rainbow trout (*Oncorhynchus mykiss*). Genetics 165:1443–1456
Sonesson AK (2005) A combination of walk-back and optimum contribution selection in fish: a simulation study. Genet Sel Evol 37:587–599

Sørensen JG, Kristensen TN, Loeschcke V (2003) The evolutionary and ecological role of heat shock proteins. Ecol Lett 6:1025–1037

St-Cyr J, Derome N, Bernatchez L (2008) The transcriptomics of life-history trade-offs in whitefish species pairs (*Coregonus* sp.). Mol Ecol 17:1850–1870

Stear MJ, Bishop SC, Mallard BA et al (2001) The sustainability, feasibility and desirability of breeding livestock for disease resistance. Res Vet Sci 71:1–7

Steinberg CEW, Stürzenbaum SR, Menzel R (2008) Genes and environment – striking the fine balance between sophisticated biomonitoring and true functional environmental genomics. Sci Total Environ 400:142–161

Stemshorn KC, Nolte AW, Tautz D (2005) A genetic map of *Cottus gobio* (Pisces, Teleostei) based on microsatellites can be linked to the physical map of *Tetraodon nigroviridis*. J Evol Biol 18:1619–1624

Storz JF (2005) Using genome scans of DNA polymorphism to infer adaptive population divergence. Mol Ecol 14:671–688

Su J, Ni D, Song L et al (2007) Molecular cloning and characterization of a short type peptidoglycan recognition protein (CfPGRP-S1) cDNA from Zhikong scallop *Chlamys farreri*. Fish Shellfish Immunol 23:646–656

Sultan M, Schulz MH, Richard H et al (2008) A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. Science 321:956–960

Swoboda I, Bugajska-Schretter A, Verdino P et al (2002) Recombinant carp parvalbumin, the major cross-reactive fish allergen: a tool for diagnosis and therapy of fish allergy. J Immunol 168:4576–4584

Symonds JE, Bowman S (2007) Atlantic cod genomics and broodstock development in Canada. Aquaculture 272:S313

Takezaki N, Nei M (1996) Genetic distances and reconstruction of phylogenetic trees from microsatellite DNA. Genetics 144:389–399

Tall BD, La Peyre JF, Bier JW et al (1999) *Perkinsus marinus* extracellular protease modulates survival of *Vibrio vulnificus* in Eastern oyster (*Crassostrea virginica*) hemocytes. Appl Environ Microbiol 65:4261–4263

Tanguy A, Guo X, Ford SE (2004) Discovery of genes expressed in response to *Perkinsus marinus* challenge in Eastern (*Crassostrea virginica*) and Pacific (*C. gigas*) oysters. Gene 338:121–131

Tanguy A, Boutet I, Laroche J et al (2005) Molecular identification and expression study of differentially regulated genes in the Pacific oyster *Crassostrea gigas* in response to pesticide exposure. FEBS J 272:390–403

Tanguy A, Bierne N, Saavedra C et al (2008) Increasing genomic information in bivalves through new EST collections in four species: development of new genetic markers for environmental studies and genome evolution. Gene 408:27–36

Tao WJ, Boulding EG (2003) Associations between single-nucleotide polymorphisms in candidate genes and growth rate in Arctic charr (*Salvelinus alpinus* L.). Heredity 91:60–69

Taris N, Ernande B, McCombie H et al (2006) Phenotypic and genetic consequences of size selection at the larval stage in the Pacific oyster (*Crassostrea gigas*). J Exp Mar Biol Ecol 333:147–158

Tasumi S, Vasta GR (2007) A galectin of unique domain organization from hemocytes of the Eastern oyster (*Crassostrea virginica*) is a receptor for the protistan parasite *Perkinsus marinus*. J Immunol 179:3086–3098

Tilton SC, Gerwick LG, Hendricks JD et al (2005) Use of a rainbow trout oligonucleotide microarray to determine transcriptional patterns in aflatoxin B1-induced hepatocellular carcinoma compared to adjacent liver. Toxicol Sci 88:319–330

Tirape A, Bacque C, Brizard R et al (2007) Expression of immune-related genes in the oyster *Crassostrea gigas* during ontogenesis. Dev Comp Immunol 31:859–873

Tocher DR (2003) Metabolism and functions of lipids and fatty acids in teleost fish. Rev Fish Sci 11:107–184

Tocher DR, Zheng XZ, Schlechtriem C et al (2006) Highly unsaturated fatty acid synthesis in marine fish: Cloning, functional characterization, and nutritional regulation of fatty acyl delta-6 desaturase of Atlantic cod (*Gadus morhua* L.). Lipids 41:1003–1016

Toranzo A, Barreiro S, Casa JF et al (1991) *Pasteurellosis* in cultured gilthead seabream (*Sparus aurata*): first report in Spain. Aquaculture 99:1–15

Trudel M, Tremblay A, Schetagne R et al (2001) Why are dwarf fish so small? An energetic analysis of polymorphism in lake whitefish (*Coregonus clupeaformis*). Can J Fish Aquat Sci 58:394–405

Vandeputte M, Kocour M, Mauger S et al (2004) Heritability estimates for growth-related traits using microsatellite parentage assignment in juvenile common carp (*Cyprinus carpio* L.). Aquaculture 235:223–236

Vandeputte M, Mauger S, Dupont-Nivet M (2006) An evaluation of allowing for mismatches as a way to manage genotyping errors in parentage assignment by exclusion. Mol Ecol Notes 6:265–267

Vasemägi A, Nilsson J, Primmer CR (2005) Expressed sequence tag-linked microsatellites as a source of gene-associated polymorphisms for detecting signatures of divergent selection in Atlantic salmon (*Salmo salar* L.). Mol Biol Evol 22:1067–1076

Velculescu VE, Zhang L, Vogelstein B et al (1995) Serial analysis of gene expression. Science 270:484–487

Venier P, De Pittà C, Pallavicini A et al (2006) Development of mussel mRNA profiling: can gene expression trends reveal coastal water pollution?. Mutat Res 602:121–134

Verrez-Bagnis V, Ladrat C, Morzel M et al (2001) Protein changes in post-mortem sea bass *D. labrax* muscle monitored by one- and two-dimensional electrophoresis. Electrophoresis 22:1539–1544

Vielma J, Makinen T, Ekholm P et al (2000) Influence of dietary soy and phytase levels on performance and body composition of large rainbow trout (*Oncorhynchus mykiss*) and algal availability of phosphorus load. Aquaculture 183:349–362

Vilhelmsson OT, Martin SAM, Médale F et al (2004) Dietary plant-protein substitution affects hepatic metabolism in rainbow trout (*Oncorhynchus mykiss*). Br J Nutr 92:71–80

Volckaert FAM, Batargias C, Canário A et al (2008) European sea bass. In: Kocher TD, Kole C (eds) Genome mapping and genomics in animals Volume 2. Springer-Heidelberg, Berlin

Walter MA, Spillett DJ, Thomas P et al (1994) A method for constructing radiation hybrid maps of whole genomes. Nat Genet 7:22–28

Wang CM, Zhu ZY, Lo LC et al (2007) A microsatellite linkage map of Barramundi, *Lates calcarifer*. Genetics 175:907–915

Wang H, Song L, Li C et al (2007) Cloning and characterization of a novel C-type lectin from Zhikong scallop *Chlamys farreri*. Mol Immunol 44:722–731

Wang CM, Lo LC, Feng F et al (2008) Construction of a BAC library and mapping BAC clones to the linkage map of Barramundi, *Lates calcarifer*. BMC Genomics 9:139

Waples RS (1998) Separating the wheat from the chaff: Patterns of genetic differentiation in high gene flow species. J Hered 89:438–450

Waples RS, Gaggiotti O (2006) What is a population? An empirical evaluation of some genetic methods for identifying the number of gene pools and their degree of connectivity. Mol Ecol 15:1419–1439

Ward RD, Woodwark M, Skibinski DOF (1994) A comparison of genetic diversity levels in marine, freshwater, and anadromous fishes. J Fish Biol 44:213–232

Was A, Wenne R (2002) Genetic differentiation in hatchery and wild sea trout (*Salmo trutta*) in the Southern baltic at microsatellite loci. Aquaculture 204:493–506

Watabe S (2001) Myogenic regulatory factors. In: Johnston A (ed) Muscle development and growth. Academic Press, San Diego

Wegner KM, Kalbe M, Reusch TBH (2007) Innate versus adaptive immunity in sticklebacks: evidence for trade-offs from a selection experiment. Evol Ecol 21:473–483

Wenne R, Boudry P, Hemmer-Hansen J et al (2007) What role for genomics in fisheries management and aquaculture?. Aquat Living Resour 20:241–255
Werbeke W, Vermeir I, Brunsø K (2007) Consumer evaluation of fish quality as basis for fish market segmentation. Food Qual Prefer 18:651–661
Whitaker HA, McAndrew BJ, Taggart JB (2006) Construction and characterization of a BAC library for the European sea bass *Dicentrarchus labrax*. Anim Genet 37:526
Williams TD, Gensberg K, Minchin SD et al (2003) A DNA expression array to detect toxic stress response in European flounder (*Platichthys flesus*). Aquat Toxicol 65:141–157
Withler RE, Candy JR, Beacham TD et al (2004) Forensic DNA analysis of Pacific salmonid samples for species and stock identification. Environ Biol Fishes 69:275–285
Wu X, Xiong X, Xie L et al (2007) Pf-Rel, a Rel/nuclear factor-kappaB homolog identified from the pearl oyster, *Pinctada fucata*. Acta Biochim Biophys Sin 39:533–539
Xue Q-G, Waldrop GL, Schey KL et al (2006) A novel slow-tight binding serine protease inhibitor from Eastern oyster (*Crassostrea virginica*) plasma inhibits perkinsin, the major extracellular protease of the oyster protozoan parasite *Perkinsus marinus*. Comp Biochem Physiol B-Biochem Mol Biol 145:16–26
Yamaura K, Takahashi KG, Suzuki T (2008) Identification and tissue expression analysis of C-type lectin and galectin in the Pacific oyster, *Crassostrea gigas*. Comp Biochem Physiol B-Biochem Mol Biol 149:168–175
Yooseph S, Sutton G, Rusch DB et al (2007) The Sorcerer II global ocean sampling expedition: Expanding the universe of protein families. PLoS Biol 5:432–466
Yu ZN, Guo XM (2006) Identification and mapping of disease-resistance QTLs in the Eastern oyster, *Crassostrea virginica* Gmelin. Aquaculture 254:160–170
Yu H, Li QI (2008) Exploiting EST databases for the development and characterization of EST-SSRs in the Pacific oyster (*Crassostrea gigas*). J Hered 99:208–214
Zane L, Bargelloni L, Patarnello T (2002) Strategies for microsatellite isolation: a review. Mol Ecol 11:1–16
Zhang H, Song L, Li C et al (2008a) A novel C1q-domain-containing protein from Zhikong scallop *Chlamys farreri* with lipopolysaccharide binding activity. Fish Shellfish Immunol 25:281–289
Zhang Y, Zhang XJ, Scheuring CF et al (2008b) Construction and characterization of two bacterial artificial chromosome libraries of Zhikong scallop, *Chlamys farreri* Jones et Preston, and identification of BAC clones containing the genes involved in its innate immune system. Mar Biotechnol 10:358–365
Zhao J, Song L, Li C et al (2007) Molecular cloning, expression of a big defensin gene from bay scallop *Argopecten irradians* and the antimicrobial activity of its recombinant protein. Mol Immunol 44:360–368
Zheng XZ, Tocher DR, Dickson CA et al (2005) Highly unsaturated fatty acid synthesis in vertebrates: New insights with the cloning and characterization of a delta 6 desaturase of Atlantic salmon. Lipids 40:13–24
Zhu L, Song L, Chang Y et al (2006) Molecular cloning, characterization and expression of a novel serine proteinase inhibitor gene in bay scallops (*Argopecten irradians*, Lamarck 1819). Fish Shellfish Immunol 20:320–331
Zhu L, Song L, Zhao J et al (2007) Molecular cloning, characterization and expression of a serine protease with clip-domain homologue from scallop *Chlamys farreri*. Fish Shellfish Immunol 22:556–566
Zinkernagel RM, Bachmann MF, Kundig TM et al (1996) On immunological memory. Annu Rev Immunol 14:333–367

Chapter 8
Marine Biotechnology

Joel Querellou, Jean-Paul Cadoret, Michael J. Allen, and Jonas Collén

Abstract Biotechnology based upon genes from the marine environment (sometimes referred to as "blue-biotechnology") has a considerable, if hitherto relatively unused, potential because of the enormous phylogenetic diversity of marine organisms and the potential for novel undiscovered biological mechanisms, including biochemical pathways. The increasing knowledge of marine genomics has started to have a major impact on the field of marine biotechnology. The advent of the sequenced genome and the development of important metagenomic resources is providing new access to the metabolic diversity of the oceans and is thereby greatly facilitating the development of new products derived from marine biotechnology. This chapter is a brief description of the field of marine biotechnology describing some of the products that have been realised and an analysis of how new genomic resources are being acquired and how they will change the landscape of future marine biotechnology.

8.1 A Brief Description of the Field of Marine Biotechnology

A simple definition of marine biotechnology is: the use of marine organisms or their components to provide goods or services. Within this chapter we will review several aspects of marine biotechnology with an emphasis on domains other than food production. Rather than providing a generic introduction to marine biotechnology, we will aim to highlight areas where, in our opinion, the interaction between genomics and biotechnology represents a powerful approach to address future challenges.

J. Collén (✉)
Station Biologique de Roscoff, UMR 7139, BP 74, 29682, Roscoff cedex, France
e-mail: collen@sb-roscoff.fr

J.M. Cock et al. (eds.), *Introduction to Marine Genomics*,
Advances in Marine Genomics 1, DOI 10.1007/978-90-481-8639-6_8,

There are several reasons for believing that marine biotechnology will be an exciting, and more importantly fruitful, domain in the coming years. Life started in water and has existed in the sea for approximately 3.8 billion years. In addition, there is significantly more phylogenetic diversity in the sea than on land; for example, of the 36 animal phyla, 14 are found only in the marine environment and, in comparison, one is endemic to the terrestrial environment (Gray 1997). One consequence of the combination of phylogenetic diversity and long evolutionary history is an enormous diversity with regards to metabolic processes. The marine environment is also extremely variable and has certain characteristics and properties that further serve to increase metabolic diversity; for example, hydrothermal vents represent high temperature environments with very special chemical composition and the intertidal area the dynamic meeting point between the terrestrial and marine environments. Seawater also contains high concentrations of halides, including bromide and iodide ions, which are used by many marine organisms in their metabolisms. The metabolic diversity of marine organisms is poorly characterised and relatively little effort has been invested so far in the marine area compared to the terrestrial domain. We believe that genomics is a powerful tool to explore this diversity and discover new marine biotechnology applications.

Marine biotechnology is a relatively recent concept, drug discovery begun with the identification of nucleosides in the sponge *Tethya crypta*, which served as models for the development of antiviral drugs in the fifties and which were later important for the development of drugs such as AZT and Acyclovir (Newman and Cragg 2004, Leary et al. 2009); the first marine antibiotic [2,3,4-tribromo-5(1′ hydroxy, 2′,4′–dibromo phenyl)pyrrole] derived from the marine bacterium *Pseudomonas bromoutilis* was described in the sixties (Burkholder et al. 1966). Yet, marine biotechnology remains at an early stage of development. Despite the overwhelming distribution of marine habitats on Earth and the promise of novel products from the organisms living there, the derived revenues have been modest. In 2006 the sales of pharmaceutical products were approximately 650 billion USD with less than half of a percent of this being derived from marine natural products despite the fact that 27% of all products were of biological origin. The enzyme market represented approximately 50 billion USD and showed a similar ratio (Leary et al. 2009). Furthermore, of the 15,000 registered marine products, presently only two registered drugs exist, Prialt® and Yondelis®, with an additional 50 (approximately) under various phases of development (Newman and Cragg 2004, www.marinebiotech.org/pipeline.html). A more effective use of genomics could change this situation, providing new enzymes for biotechnology, opening new avenues for the treatment of disease and monitoring health, increase the efficiency of aquaculture, and develop new resources for industrial materials and processes. This could be, for example, a compound from a coral used as an anti-inflammatory drug (Mayer et al. 1998), new anticancer drugs derived from marine algae (Fuller et al. 1992, 1994) and other marine sources (Amador et al. 2003) or bacteria that digests up oil spills (Head et al. 2006). One factor that will, in our opinion, increase the interest in marine biotechnology is REACH, the new European Community Regulation on chemicals and their safe

use (http://ec.europa.eu/environment/chemicals/reach/reach_intro.htm) which will increase the emphasis of finding new enzymes and processes that will allow for a more efficient and less polluting chemical industry.

8.2 How Genomics Impacts on the Various Fields of Marine Biotechnology

Until now marine genetic resources have been scarcely valorised with only 135 relevant patents found in an "indicative" survey between 1973 and 2007 (Leary et al. 2009). One reason for this limited number of patents is that hitherto, research for new compounds from the marine environment has been either fortuitous, large scale or guided by physiological or ecological knowledge, an example of the latter being searches for bioactive compounds in organisms that are rarely predated, epiphytised, or grazed (Sennett 2001). An increased knowledge in genomics will provide us with a powerful guide to the genetic basis of the complex metabolic activities in the marine environment and will allow for a multitude of discoveries and the development of commercial products.

Many marine microorganisms are difficult to keep in culture, making it difficult to exploit them for the production of new products. Metagenomics on organisms or communities that are presently difficult or impossible to culture can give insights into physiological mechanisms that are difficult to obtain with other methods (see below). Indeed, the coupling of metagenomics with an appropriate screen can prove to be a powerful tool in modern day biotechnology. The high demand for biocatalytic enzymes has generated a wide range of specific and useful screens that can be easily exploited using libraries of marine origin, one of the most common being a screen for esterase function. Using marine metagenomic libraries from the South China Sea and Arctic sediment four novel esterases have been identified and characterised in the last year alone (Jeon et al. 2009, Chu et al. 2008). Hydrolytic enzymes, such as esterases, are useful biocatalysts because of their extensive versatility for industrial application: they generally have broad substrate specificity and are stereoselective.

Increased genomic knowledge can facilitate gene discovery, making it easier to go from protein to sequence (and vice versa). If, for example, an interesting biological activity has been found and the corresponding protein or enzyme has been purified it is much simpler, with a known genome, to find the relevant gene. Comparison of a gene of known function with sequences in the same or other organisms allows similar genes to be identified. This provides access to a diversity of genes encoding a protein of interest. For example, recently a new family B DNA polymerase from *Thermococcus thioreducens* (an archaeon isolated from the Rainbow hydrothermal vent field) was cloned, expressed, purified and characterised. This novel DNA polymerase was shown to perform well under a range of PCR conditions, being faster, more stable and more accurate than many commonly used enzymes (Marsic et al. 2008). Conversely, the absence of a gene in an organism

could also be a reason to not look for an activity or, if the activity is vital, a reason to look for alternative enzymes with similar activities.

8.3 Expanding Gene Resources Through Microbial-Community Genomic Projects, Complete Genomes of Isolated Organisms and Data Mining

Gene resources for biotechnology have been increasing exponentially during the last decade owing to improved sequencing technology and the resulting increase in the number of genome projects. However, the databases are filled with human-centric genomic information with model species, (including human, rat, mouse, rice) as well as their associated pathogens (both human and agricultural) constituting the bulk of genome projects. Nevertheless, several marine species and marine communities have directly benefited from the genomic revolution either because their position in the tree of life is strategic or for their potential in biotechnology.

The first genome-scale information was obtained directly from isolated and cultured species of interest, already available in microbial culture collections. However, the vast majority of archaeal and bacterial taxa remain uncultivated (Amann et al. 1995, Rappé and Giovannoni 2003). In addition, analyses of marine viromes have demonstrated that most marine virus genes are unrelated to those in the current databases (Breitbart et al. 2002, Angly et al. 2006) and of course, they are massively unexploited. Gaining access to the almost unlimited microbial gene resources from various marine environments is an exciting task. It relies mainly on two complementary approaches: metagenomics and single-cell genomics. A third approach aiming at developing high throughput protocols for isolation and culture of previously uncultured (the vast majority) microorganisms has largely been ignored during the last ten years, despite its obvious interest and some pioneering work in this area (Giovannoni and Stingl 2007).

Many metagenomic projects are based on the sequencing of large DNA libraries constructed from environmental DNA as exemplified by the Sargasso Sea project (Venter et al. 2004). In order to obtain a more complete picture beyond random examples, the exploration of microbial diversity needs to be carried out in a rational manner. The development of appropriate methods and strategies to estimate the sampling effort necessary for a particular marine environment is in progress (Quince et al. 2008). Despite the obvious limits of metagenomics, this approach has aroused a considerable interest and it has been suggested that it might be possible in the future to gain access to most of the gene resources of the biosphere, including those of rare species, by applying high-throughput sequencing methodology. A single metagenomic project, the Sorcerer II Global Ocean Survey, nearly doubled the number of known proteins in the databases (Rusch et al. 2007). One major consequence of metagenome projects is the radical change in our understanding of microbial diversity in samples of marine environments like sediments and deep-sea vents (Sogin et al. 2006, Huber et al. 2007, Quince et al. 2008). Strategies to fully

exploit the huge reservoir of genes in these types of environments have still to be refined, in particular to gain access to rare phylotypes or species.

In spite of considerable efforts to develop novel bioinformatic methods, the assembly and analysis of discrete microbial genomes from complex communities is still a difficult task. This has triggered the development of alternative approaches among which single-cell genomics is becoming increasingly popular. This method involves a first step based on single cell isolation followed by whole-genome amplification using φ29 DNA polymerase in a reaction called multiple displacement amplification (Dean et al. 2002). It opens interesting perspectives in various cases, notably when the targeted species are still uncultured or when the amount of DNA available is extremely limited.

The overall result of sequencing projects dedicated to both specific genomes and to metagenomes is the unprecedented access to unique gene sequences. To put this novel gene resource to good use is the major challenge of modern day genomics. The traditional combination of *in silico* analysis (data mining) and functional screening is still the cornerstone for discovery in biotechnological applications. Yet with the increasingly large and complex datasets being generated, we are at a stage where we will soon be drowning in data whilst still thirsting for knowledge. In the future, novel multidisciplinary approaches will need to be developed to efficiently exploit our increasing genomic knowledge.

8.3.1 Complete Genomes

For a long time it was thought that the sequencing of the complete genome of a type species would give access to the majority of the gene pool of this species and that a few dozen representative genomes from the bacterial and archaeal domains would describe most of the gene pool of the microbial world. This is clearly not the case. The genomic variation within a single species was inferred from comparative genomics of pathogens including *E. coli* and *Streptococcus agalactiae* (Bielaszewska et al. 2007). These studies demonstrated that it is not possible to characterize a species from a single genome sequence and that the number of genome sequences (pan-genome) needed to describe a species may vary depending on the species. As a consequence, for some important pathogens, the number of strains undergoing genome sequencing is increasing rapidly (for example: 2 complete and 14 incomplete sequences for *Vibrio cholerae*, and probably more than 50 within 2 years). Until now, all genome sequences of marine species of biotechnological interest have been obtained from single clonal cultures established from single cell. In the process, part (possibly, most!) of the variability has been lost and the corresponding set of genes may not be easily retrieved by genome analysis and data mining. This is why functional screening is still needed, in parallel with data mining, to explore strain variability. In the near future, as with pathogens, projects dedicated to genome sequencing of several strains belonging to a single species of biotechnological interest, are likely to become widespread. This is already the case for the fungi

Penicillium chrysogenum, with two genome sequences currently being completed (www.genomesonline.org/gold.cgi).

The microbes whose complete genome have been sequenced up to date (August 2008) are mostly from the Bacteria (684 in NCBI Entrez Genome Database; 696 in Genomesonline Database), followed by Archaea (52–53). Fungi, despite being present in many marine habitats, are underrepresented in the list of complete genomes. The ratio of marine to terrestrial species varies considerably depending on the domain of life (Table 8.1).

Table 8.1 Status of genome sequencing projects in the three domains of life and in marine species (compiled from genomesonline.org, and modified, August 2008)

	Bacteria			Archaea			Eukarya		
Genome status	All	Marine	Ratio (%)	All	Marine	Ratio (%)	All	Marine	Ratio (%)
Complete and published	690	74	11	50	19	38	94	6	6.4
Complete	13	1	7.7	1	1	–	28	1	–
Incomplete	1573	184	11.7	69	10	14.5	547	2	–
Targeted	302	8	2.6	22	6	27	6	?	–
ALL	2578	267	10.4	142	36	25	675	9	1.3

From these data, it is clear that eukaryotic marine species are under-represented. The picture is to some extent more satisfactory for bacteria, with 10% of the sequencing projects dedicated to marine species. Finally, the best figure is observed for Archaea. The total number of sequencing projects for Archaea is rather low, but the proportion of marine species is as high as 38% for the published projects and 27% for the targeted projects. Globally, the trends for marine species are currently low despite the increase in the number of genomes available in absolute terms. This does not reflect the huge interest of marine species both for basic research (notably for evolution and development) and for biotechnology. In addition, no more that 4% of sequencing projects deal with extremophiles that have industrial applications.

Among the various completely sequenced genomes with already proven or potential interest for biotechnology different clusters emerge; the archaeal group of *Pyrococci* and *Thermococci* is significantly overrepresented. Numerous species of these two genera combine interesting DNA processing enzymes (DNA polymerases, DNA ligases) with a rich repertoire of thermostable hydrolases displaying interesting properties (reviewed by Egorova and Antranikian 2007) with respect to the degradation of starch, cellulose and chitin, in biofuel production, and in the degradation of complex or recalcitrant proteinaceous substrates, including keratin and prion protein (Tsiroulnikov et al. 2004). Up until the end of the nineties, all the thermostable enzymes from these species were obtained by biochemical methods based on functional screenings. Following the subsequent publication of several complete genomes (Kawarabayasi et al. 1998, Maeder et al. 1999, Cohen et al.

2003), functional screens were combined with genome data mining. In some cases, the initial research step was an *in silico* analysis and searches for specific targets were followed, when successful, by biochemical characterization. The characterization of *Pyrococcus abyssi* thermostable nitrilase is a good example of this procedure applied to enzyme discovery (Mueller et al. 2006).

The recent oil crisis emphasized the need for the development of alternatives to fossil fuels. A recent review of "bioenergy genomes" (Rubin 2008) listed the species already known as biomass degraders or fuel producers whose genomes have been completely sequenced or for which projects are under way. This list highlights the small contribution of marine genomes to this huge project. This does not mean that marine species are inappropriate in this context, we have already underlined the presence of thermostable amylases, cellulases, xylanases, etc. in hyperthermophilic archaea (Table 8.2), but it is likely that a strong bias toward terrestrial species underlies these choices (see also below).

8.3.2 The Growing Contribution of Metagenomes

The title of a review published in the Biotechnological Journal and entitled "Metagenomics: an inexhaustible access to nature's diversity" summarizes a common opinion about the potential contribution of metagenomes to biotechnology (Langer et al. 2006). Marine metagenomics is mostly based on techniques and methods developed from the study of soil metagenomics. It emerged as the combination of extraction and digestion of bacterial DNA from soil (Torsvik 1980), the generation of gene libraries from environmental DNA (Pace et al. 1986), the conception and generation of marine plankton environmental DNA libraries (Schmidt et al. 1991). The term metagenome was proposed by Handelsman to describe the entire set of sequences of organisms living in a defined habitat (Handelsman et al. 1998). The biotechnological potential of this novel approach was soon recognised (Short et al. 1997). The initial goal of marine metagenomics was the inventory of microbial diversity in the oceans, with the aim of providing access to the functions of previously hidden key players and their role in the main geochemical cycles. However, it was only at the beginning of the twenty-first century, following the introduction of large sequencing facilities, that this approach was used on a large scale by C. Venter and his team during the Sargasso Sea metagenome investigation (Venter et al. 2004) and the Sorcerer II/GOS expedition (Rusch et al. 2007).

The principal advantage of metagenomics is that it allows the current limits of culture-dependent methods to be bypassed. This is especially interesting when microbial samples are recovered from highly complex communities and from "metaorganisms" (usually one eukaryote and its associated microbiome), like sponges where the vast majority of the microbial community remains uncultured.

Marine environments are extremely diverse on Earth and their exploration and exploitation offer untapped gene resources for biotechnology. Metagenomic approaches, combined with heterologous expression, appropriate high throughput

Table 8.2 Marine microbial genomes of biotechnological interest

	Organism	Size(kb)	Orf s	NCBI Ref	Institution	Targets, type of enzymes, type of applications References (genomes; biotech) Archae
Archaea	*Aciduliprofundum boonei T*469	2973		NZ ABSD00000000	JCVI, Univ Portland	Incomplete (proteases) (unpublished)
	*Aeropyrum pernix K*1	1669	1700	NC_000854	NITE	Proteases (Kawarabayasi et al. 1999)
	Hyperthermus butylicus DSM 5456	1667	1602	NC_008818	Eipdauros Biotech.	Peptidases (Brügger et al. 2007)
	*Nitrosopumilus maritimus SCM*1	1645	1795	NC_010085	JGI	* (unpublished)
	*Pyrobaculum aerophilum IM*2	2222	2605	NC_003364	UCLA, Caltech	Glycoside-hydrolases, proteases (Fitz-Gibbon et al. 2002)
	*Pyrococcus abyssi GE*5	1765	1896	NC_000868	Genoscope	DNA processing; starch, cellulose, xylan degrading; proteases (Cohen et al. 2003)
	Pyrococcus furiosus JCM 8422	1908	2125	NC_003413	Univ. Utah/Maryland	DNA processing; starch, cellulose, chitin degrading; proteases (Vanfossen et al. 2008, Jenney and Adams 2008)
	*Pyrococcus horikoshii OT*3	1738	1955	NC_000961	NITE, Univ Tokyo	DNA processing; starch degrading; proteases, hydrogenases (Kawarabayasi et al. 1998)
	Pyrodictium abyssi DSM 6198	*	*	*	JCVI, GBM-MGSP	Starch, xylan degrading (unpublished);
	Pyrolobus fumarii	1850	2000	*	Celera, Diversa	Thersmostability/pressure; organic solutes (unpublished); (Gonçalves et al. 2008)
	*Staphylothermus marinus F*1	1570	1570	NC_009033	JGI	Starch processing, protease, DNA processing (unpublished)
	*Sulfolobus solfataricus P*2	2292	2977	NC_002754	CBR	Starch, cellulose degrading, proteases (She et al. 2001)
	*Thermococcus kodakaraensis KOD*1	2088	2306	NC_006624	Kwansei Gakuin Univ	DNA processing; starch degrading enzymes; hydrohenases (Fukui et al. 2005, Kanai et al. 2005)
	Thermococcus barophilus MP	2059*	*	NZ ABSF00000000	JCVI, Prokarya	Proteases, starch degrading, pressure (unpublished)
	*Thermococcus onnurineus NA*1	*	*	*	KORDI	DNA polymerase, Proteases, starch processing (unpublished); (Lim et al. 2007)
	*Thermococcus gammatolerans EJ*3	*	*	*	IGM	DNA processing, proteases (unpublished)

Table 8.2 (continued)

	Organism	Size(kb)	Orf s	NCBI Ref	Institution	Targets, type of enzymes, type of applications References (genomes; biotech) Archae
Bacteria	*Alcanivorax borkumensis SK*2	3120	2755	NC_008260	Bielefeld Univ	Oil hydrocarbon degradation (Schneiker et al. 2006)
	Alteromonas macleodii DSMZ 17117	4236	4163	NZ_AAOD00000000	JCVI	Exopolysaccharide production (Ivars-Martinez et al. 2008)
	*Aquifex aeolicus VF*5	1551	1529	NC_000918	Diversa, Univ Illinois	Glucan branching, molecular biology (Deckert et al. 1998)
	Bacillus halodurans C-125	4202	4066	NC 002570	JAMSTEC JCVI	Starch, pullulan and xylan degrading (Takami et al. 2000)
	Dehalobium chlorocoercia DF-1	*	*		Epidauros Biotech.	PCB dechlorination (unpublished)
	*Desulfotalea psychrophila ESV*54	3523	3116	NC_006138	JGI	Psychrophilic enzymes (Rabus et al. 2004)
	*Marinoboacter hydrocarbonoclasticus VT*8	4326	3858	NC_008740	JVI	Recalcitrant hydrocarbon sources degrading (unpublished)
	*Marinitoga piezophila KA*3	2000	*	*	JGI	Piezophilic thermostable enzymes (unpublished)
	*Marinitoga camini MV*1075	*	*	*	JAMSTEC	Thermostable proteases (unpublished)
	*Nitratiruptor sp SB*155-2	1877	1843	NC_009662	MPI, MPI-MM	Virulence factors (Nakagawa et al. 2007)
	*Rhodhopirellula baltica SH*1	7145	7325	NC_005027	JCVI	Sulfatases (Glockner et al. 2003)
	*Thermotoga maritima MSB*8	1860	1858	NC_000853	JCVI	Thermostable glycosides hydrolases (Nelson et al. 1999, Vanfossen et al. 2008)
	*Vibrio fischeri ES*114	4478*	4061	NZ_ABIH00000000	CNRS-SBR	Luminescence (Ruby et al. 2005)
	Zobellia galactanovorans Dsij	*	*	*		Sulfatases, sulfotransferases, food industry (unpublished)

* indicates unpublished

screening procedures and directed evolution have been applied with success by several research groups and SMEs as well as larger corporations (Diversa, Genencor, Degussa, Henkel, etc.) to different targets, enzymes and/or novel natural products. Table 8.3 gives an overview of the most significant metagenomic contributions during the recent years.

Though limited to ocean surface microbes, the Sorcerer II/GOS expedition represented a milestone for metagenomics and the team predicted more than six million proteins in the GOS data, approximately twice the number of proteins present at that time in the databases (Yooseph et al. 2007). The practical utility of the Sorcerer II results for marine biotechnology is hard to estimate and their exploitation for that purpose will take many more years to come to fruition. One of the primary goals of the Sorcerer II studies was to understand the rate of discoveries of protein families with the increasing number of protein predictions. From the relationship observed, they concluded that their sampling of the protein universe is far away from saturation, meaning that many more protein families remain to be discovered. This conclusion is subject to controversy and Koonin (Koonin 2007) suggested comparing not only the increase in the number of protein clusters with the number of sequences observed in the Sorcerer II data but also adding the rate of increase of orthologs. The virome (viral fraction of the microbiome in a given habitat) of several marine regions has been studied and this has also revealed a completely novel diversity of proteins (Breitbart et al. 2002, 2004, Angly et al. 2006). The results from genomics projects dedicated to Archaea viral proteins with unknown functions have shown that marine viruses represent a reservoir of interest for the discovery of new folds and novel structures and therefore novel protein families (Vestergaard et al. 2008).

With the advent of metagenomic approaches, the microbial diversity is increasingly accessible and we might expect an exponential growth in the number of biocatalysts potentially useful for the industry, as already demonstrated by some recent work on nitrilases, lipases and esterases (Robertson et al. 2004, Bertram et al. 2008), at least when functional screens are efficient. Note, however, that in contrast to global marine metagenome surveys, these results were obtained in direct collaboration with industry, with clearly defined biotechnological goals.

There are two main obstacles to the biotechnological exploitation of marine extremophiles. Screening methods (sequence-based and activity-based screens) represent the first limitation. The second is that, in some extreme environments, cell densities are so low that the amount of DNA available for cloning is extremely small, implying additional steps like whole genome amplification. Sequence-based screens are limited to already known gene families and cannot identify completely new genes frequent in metagenome databases like CAMERA (Seshadri et al. 2007, http://camera.calit2.net). Moreover, many genes in metagenome libraries can not be expressed successfully in organisms such as *E. coli* and therefore, cannot be detected by activity-based enzymatic screens. Attempts have been made recently to circumvent these limitations and these have led to the development of innovative methods like substrate-induced gene expression screening (SIGEX) (Uchiyama et al. 2005).

Table 8.3 Contributions to marine metagenomics and biotechnology

	Authors	Origin of sample(s)	Screening method	Main targets	Industry	Reference
Methods	Rondon et al. (2000)	Soils	Function based	Enzymes, antibacterial, hemolytic activity	– –	
	Henne et al. (2000)	Soils (meadow, field)	Function based	Lipases, esterases	Patent –	
	Uchiyama et al. (2005)	Groundwater	SIGEX	Method, aromatic hydrocarbon degradation		
	Kalyuzhanya et al. (2008)	Lake sediment	C_1 substrates	Group diversity/function		
Diversity	Breitbart et al. (2002)	Surface seawater	–	Viral diversity	–	
	Venter et al. (2004)	Sargasso Sea surface	–	Microbial diversity	–	
	Rusch et al. (2007)	Atlantic, Pacific Oceans	–	Microbial diversity	–	
	Yooseph et al. (2007)	Atlantic, Pacific Oceans	–	Microbial diversity		

Table 8.3 (continued)

	Authors	Origin of sample(s)	Screening method	Main targets	Industry	Reference
Biotechnology	Schirmer et al. (2005)	Sponge *Discodermia*	Sequenced/PCR	Polyketide synthases (PKS)	Kosan[a]	
	Fieseler et al. (2007)	20 sponge species	Sequenced/PCR	Identification of pharma relevant PKS genes	Kosan	
	Chu et al. (2008)	Surface seawater, China	Function based	Lipases, esterases	?	
	Short et al. (1997)	Diverse incl. plankton	Function based	Hydrolases (notably thermostables)	Diversa[b]	WO9704077
	Short (1999)	Whale bone, picoplankton	Function based	Hydrolases	Diversa	US5958672
	Weiner et al. (2007)	Pacific Ocean picoplankton	Function based	Oxidative enzymes (epoxidases, P450) :	Diversa	US2007231820
	Lee et al. (2008)	Deep-sea sediment	Function based	Fibrinolytic Metalloprotease	Patent	WO2008056840
	Robertson et al. (2004)	>600 biotopes	Function based	Nitrilases (137 novel enzymes)	Diversa	
	Bertram et al. (2008)	Information missing	Function based	Lipases, esterases (350 novel enzymes)	Verenium	

[a]Kosan Biosciences is now a subsidiary of Bristol-Myers;
[b]Diversa is currently a branch of Verenium. Patent data were compiled from the European Patent Office (http://ep.espacenet.com/). In August 2008, a query with metagenome as keyword led to 8 entries, most of them dedicated to soil metagenomes and cow rumen.

SIGEX is designed to select the clones that harbour catabolic genes induced by various substrates coupled with GFP and sorted by fluorescence-activated cell sorting (FACS). This high throughput method was successfully applied to isolate aromatic-hydrocarbon-induced genes from a metagenomic library. In spite of its advantages, it suffers from several limitations (such as the orientation of genes in the cloning vector or genes with terminators and consequently poor yield with large inserts) (De Lorenzo 2005, Yun and Ryu 2005).

The search for various biocatalysts from natural sources can be facilitated by the addition of various substrates to the environment to enrich the microbial community in strains able to degrade these substrates. In the same manner, an approach that might considerably reduce the sequencing effort required by metagenomics would be to sort the DNA of interest from the bulk DNA. Considerable progress has been done in that direction in the last years using stable isotope probing (SIP). Applied to lake sediments, targeting specifically methylothrophs (metabolism of organic compounds containing no carbon-carbon bonds), through ^{13}C-labeled C_1 compounds, it was possible to extract DNA from the microcosms and separate the labelled fraction containing the genes of interest (Kalyuzhnaya et al. 2008). This method, though developed mainly for analysis of the functions of the different components of a microbial community, can be used in metagenomics of complex communities to enrich DNA fractions to be sequenced in a define gene profile and therefore, contribute to biotech research.

8.4 Contribution of Marine Biotechnology to the Discovery of Natural Products, Novel Pharmaceuticals and White Technology

Numerous natural products and enzymes have been discovered in marine organism; some examples are given here from bacteria, archaea, marine fungi, metazoans, virus and algae. This is by no means a complete listing of all products found in the marine environment, only a few products are indicated from each group to demonstrate the potential of the different taxa. The absence of certain phylogenetic groups should be seen as an indication that an important unexplored metabolic diversity may be found, for example, in the amoebozoa, rhizaria and among the excavates.

8.4.1 Viruses

Marine viruses represent perhaps the greatest untapped biotechnological resource on the planet. With an average of 10^6 viruses/ml of sea water and an estimated 10^{30} in our oceans, this is a massively under-studied and under-utilised resource (Suttle 2007). Historically, enzyme and natural product discovery has depended upon culture dependent approaches, approaches that are more complicated for viruses which require a suitable host for propagation. Thus, the advent of the metagenomic

era and its associated culture independent approaches has opened up the field of marine virology for exploitation. No one gene is common to all viruses and random sequencing of the viral fraction of the ocean has revealed a plethora of genetic diversity that has so far been inaccessible (Angly et al. 2006).

In the past, viruses were thought of as simple, self propagating bags of genes. The discovery of giant viruses in particular has turned this notion on its head and has shown that viruses can often have the ability to manipulate and control complex metabolic pathways using enzymes encoded on their own genomes (Raoult et al. 2004, Wilson et al. 2005, la Scola et al. 2008). For example, proteins of viral origin feature heavily in the molecular biologists armoury. Proteins from bacteriophage T4 (which infects *E. coli*) such as its DNA ligase, polynucleotide kinase, DNA polymerase are all commercially available for cloning purposes. RNA polymerases from T7, phi6 and SP6 are in common usage. Reverse transcriptase (RNA-directed DNA polymerase, a function only found in viruses thus far) from Avian Myeloblastosis Virus and Moloney Murine Leukemia Virus are standard enzymes for making cDNA from RNA templates. Whilst none of these viruses are marine in origin, these examples demonstrate that viruses harbour useful, efficient and exploitable enzymes. Few marine viruses have been studied intensively, yet the handful that have been have revealed considerable biotechnological potential (Allen and Wilson 2008). For example, the coccolithovirus EhV-86 has a genome that appears to encode a near complete ceramide synthesis pathway (a transferase, elongation protein, phosphatase and three desaturases) (Wilson et al. 2005, Han et al. 2006). Ceramides are components of the plasma membrane and has been made use of as anti-aging component commonly found in cosmetics. This newly discovered viral pathway is being investigated not only for its academic relevance but also for potential commercial exploitation. Also found on this genome are a lipase and esterase (as well as numerous nucleases and proteases) which may have biocatalytic applications (Allen et al. 2006b). Yet, even with the briefest of glimpses at viral genetic diversity, we have realised how little we know about how these organisms function: typically 80% of newly derived sequence of viral origin can be unique and have no known function associated with it (Suttle 2005). Clearly, the vast majority of these novel genes must be of some use to viruses and yet we have no clues as to their function and relevance (Yin and Fischer 2008). As the functions of these genes are unravelled, we predict new and exciting biotechnological applications exploiting novel biochemical pathways and reactions.

Inteins have gained increasing attention since their discovery (Gogarten and Hilario 2006). An intein is a self splicing segment of a protein which has the ability to excise itself and rejoin the adjacent segments of the protein with a peptide bond. They have important biotechnological applications for protein expression, synthesis, purification and labelling (Perler 2002). Inteins have been found throughout all kingdoms of life, but are now being found to be increasingly common in marine viruses. Inteins have been found in *Heterosigma akashiwo* viruses (Nagasaki et al. 2005), in some strains of coccolithoviruses (Allen et al. 2006a; Goodwin et al. 2006), as well as in the mimivirus (Ogata et al. 2005), relatives of which (with intein fragments) have been found in the Sargasso Sea metagenome dataset (Ghedin and Claverie 2005).

8.4.2 Archaea and Bacteria

Bacteria and more recently archaea are important organisms in biotechnology including the prokaryotes of marine origin. Complete genomes of hyperthermophilic bacteria and archaea have opened novel avenues for enzyme discovery. Examples include enzymes used in research such as thermostable enzymes from deep-sea hyperthermophilic Archaea, some of them already in common use in molecular biology. Potential applications range from biomass processing in order to produce ethanol and biofuel to organic chemistry and biohydrogen production.

Examples of the exploration of bacteria include *Rhodopirellula baltica* and *Zobellia galactanovorans*. The first one was known to be densely associated with "marine snow" in coastal waters and presumed to be an efficient degrader of organic compounds. The second is frequently associated with macroalgae and potentially involved in cell-wall degradation. Complete genome sequencing identified unprecedentedly high numbers of enzymes belonging to the sulfatase (Glockner et al. 2003) and sulfotransferase families in both species (unpublished data). Based on these results, functional genomics projects emerged in order to characterize both the various substrates and the corresponding enzymes.

8.4.3 Algae

The marine algae, including cyanobacteria, is an extremely divergent group with members from archaeplastida, chromalveolates and eubacteria (see Algal chapter) and thus with an extensive metabolic diversity. Despite this, examples of genes from marine algae that are used in biotechnology are sparse, but one success story is the hexose oxidase gene from the red alga *Chondrus crispus* which has been over-expressed in heterologous systems for the production of a commercial product. When added to food products hexose oxidase limits the Maillard reaction and thereby limits coloration and can strengthen the gluten network (Hansen and Stougaard 1997). Polyunsaturated fatty acids are also a good example of the role of marine biotechnology and hence marine genomics since there is no significant source of these important lipids outside of the marine world. Phytoplankton is considered as one of the major source for e.g. EPA and DHA. Algal genes involved in fatty acid synthesis have indeed attracted some interest. Genes from the diatom *Thalassiosira* have already contributed to research and commercial efforts aimed at producing very long chain polyunsaturated fatty acids in transgenic crop plants (Tonon et al. 2004a, b, 2005). An example of the use of algal products is the use of purified laminarin from *Laminaria digitata* to elicit defence reactions in tobacco (Klarzynski et al. 2000). This was the base for the development of Iodus by Goëmar as a partial replacement product of fungicides in commercial cultivation of cereals.

Indeed, one area that has attracted a lot of interest recently is algal biofuels and marine genomics can potentially make an important contribution in this domain. The world investment in biofuels research has risen dramatically in the last few years,

and is predicted to reach the sum of over a billion USD. Microalgae show some advantages compared to terrestrial plants: unicellular algae can under favourable conditions achieve photosynthesis rates greatly exceeding that of higher plants. Marine microalgae offer two other main advantages: their exploitation does not need to be in conflict with fresh water or food supply. Compared to higher plants, the culture in controlled ponds or photobioreactors allows a fine regulation of the metabolism by adjusting on line the level of entering nutrients. In addition, culture of these photosynthetic organisms requires the supply of CO_2 and nutrients such as nitrogen and phosphorous which further explains the rising number of programs dealing with gas abatement and wastewater management. It is important to distinguish the production in controlled photobioreactors with illuminations or natural light, which allows for very high productivities with high costs from productions in natural environment in open ponds for which costs are lower. Indeed, extensive cultivation in outdoor ponds offers the best competitiveness with the major drawback that these cultures are more "natural" and thus more difficult to control. With the exception of cultures in extreme media such as high salt or high pH, the pollution with local endemic species is very likely to occur as well as the contamination with various grazers. However, where the usual terrestrial crops yield around a gram per square meters per day, the average yield for microalgae lies between 10 and 30 g/m^2/day. Algal cultivations have thus the potential to produce substantial amounts of biofuels, without competing with the food production industry in terms of use of arable land and fresh water. There are, however, several technological challenges that need to be overcome in order to produce algal biofuels at competitive prices (Cadoret and Bernard 2008).

The four main suggested alternatives for biofuel production from algae are: production of biodiesel, hydrogen and ethanol as well as biomass for fermentation (Chisti 2007).

8.4.4 Algae for Biodiesel Production

Some algae and especially microalgae stock lipids as energy reserve. Among 3,000 contenders studied by the National Renewable Energy laboratory, NREL (Sheehan et al. 1998), candidates were identified with an oil content of up to 75% of dry weight (Schenk et al. 2008). The normal yields are usually around 40% for most of the best candidates (Rodolfi et al. 2009). To produce biodiesel algae are grown under conditions that induce lipid production; the algae are harvested; the lipids are extracted and converted to a useful fuel. It has been calculated that the production of more than 100 m^3 biodiesel per hectare per year is theoretical feasible and 10–50 m^3/ha/year, beyond the 6 m^3/ha/year of the oil palm (Chisti 2007). To achieve high biodiesel production, it will be necessary to improve our knowledge of the biochemistry of lipid synthesis in different algae as well as increasing our understanding of the physiology of lipid metabolism. The diatom *Phaeodactylum tricornutum* can contain up to 31% of dry weight as lipids (Sheehan et al. 1998) and its genome has

been completely sequenced (Bowler et al. 2008). In this species, therefore, genomic approaches can give important insights into the biology and the biochemistry of a high lipid producing algae.

8.4.5 Algae for Ethanol Production

An alternative approach for biofuel production is to use the photosynthetic products of cyanobacteria to produce ethanol in vivo. The feasibility of the approach was demonstrated by Deng and Coleman (1999) who used *Synechococcus* sp. PCC 7942 transformed with pyruvate decarboxylase and alcohol dehydrogenase from the bacterium *Zymomonas mobilis*. The transformed cyanobacterium synthesised ethanol from photosynthetates, which diffused into the culture medium. The system is claimed to be ready for commercial scale-up and the companies Algenol and BioFields have committed 100 million USD for the construction of a full scale factory (www.algenolbiofuels.com). The company Algenol now also claims that the technique has been adapted for cyanobacteria growing in seawater. This is an interesting example of one potential future for marine genomics in which marine organisms and "terrestrial genes" (and vice versa) are combined to produce promising products.

8.4.6 Algae for Hydrogen Gas Production

Many cyanobacteria and some green algae contain enzymes capable of producing hydrogen gas (Tamagnini et al. 2002, Schütz et al. 2004, Melis and Happe 2001). In this case the photosynthetic energy is converted directly to a useful high energy fuel without the need for extraction since the hydrogen gasses off and can be collected. Research on green algae has so far concentrated on *Chlamydomonas reinhardtii*, which has an efficient but oxygen sensitive hydrogenase (Hankamer et al. 2007). This constraint was partially overcome by Melis et al. (2000) by cycling between a state of active photosynthesis and an anaerobic hydrogen producing state induced by sulphur starvation. Strains have been developed with potential for increased hydrogen production for example in a work by (Surzycki et al. 2007) or with the selection of strains with smaller light harvesting complexes (Polle et al. 2003) or a strain with high starch content (Kruse et al. 2005). In cyanobacteria hydrogen can be formed as a by-product of nitrogen fixation by nitrogenase. The hydrogen is normally oxidized by an uptake hydrogenase (for reviews see Tamagnini et al. 2002, Sakurai and Masukawa 2007). As with hydrogen production in green algae, production by cyanobacterial is inhibited by oxygen since the nitrogenase is oxygen sensitive.

8.4.7 Algae for Biomass Fermentation

It is also possible to grow or harvest microalgae and seaweeds for fermentation of biomass. This fermentation could be either for ethanol or for biogas. In this case it would be important to either achieve high growth rates or to increase the fractions

of easily fermentable compounds, such as starch and lipids. One example of an approach that could use algae in this way is the exploitation, of high starch strains of *Chlamydomonas reinhardtii* (Kruse et al. 2005).

8.4.8 Marine Genomics and Algal Biofuels

What can marine genomics contribute to biofuel production? The genomes of a number of algae have been fully sequenced. These include eukaryotes like the green algae *Chlamydomonas reinhardtii, Ostreococcus lucimarinus,* and *O. tauri,* the red alga *Cyanidioschyzon merolae,* and the diatoms *Thalassiosira pseudonana* and *Phaeodactylum tricornutum* (Bowler et al. 2008) and several cyanobacteria, for example, *Synechococcus* spp (Six et al. 2007). These genomes can provide important insights into different areas relevant to biofuel production. For example, *Ostreococcus tauri* is the world's smallest free-living autotrophic eukaryote has already contributed to our knowledge about small genomes and will undoubtedly in the future increase our knowledge about organisms with minimalistic genomes and thus possibilities of efficient growth since fewer resources are used for non-essential purposes. Another example is *Cyanidioschyzon merolae,* a red algal extremophile which also has a small genome but is adapted to high temperature and low pH. The genomes of different strains and species of *Synechococcus* include strains that are adapted to factors such as different light intensities and these have already provided insights into antennae structure and function (Six et al. 2007).

The idea of using microalgae as a source of biofuel is relatively new, the strains that are used are mostly wild-type strains (Sheehan et al. 1998). There is therefore a considerable room for improvement of the strains used. Strain improvement could be done both by classical approaches and by genetic modification of the organisms. In both cases an improved knowledge of their genes and genomes would be a major advantage in order to direct classical genetics and to optimise genetic modifications. Below (Table 8.4) are some more general areas that we believe can greatly benefit from genomics.

8.4.9 Algae as a Cell Factory

Algal transformation to produce so-called "green cell factories" represents a booming area of research with broad implications not only for biofuels, but also speciality chemicals, high value compounds, food additives and bioremediation (Rosenberg et al. 2008). Natural substance extraction still constitutes the primary source for a great number of pharmaceutical molecules. However, as it is possible to identify genes responsible for the development of a protein, these can then be introduced into cells in culture, as part of "cell factories" which manufacture on demand, the desired products. This strategy – the expression of molecules with high added value

Table 8.4 Research areas where marine genomics can play a key role

Research area	Comment
Increasing growth rates	Growth rates vary enormously between algal species and strains
Reducing photo-synthetic antenna size	Reducing antennae size would reduce internal shading and diminish photoinhibition allowing higher densities in culture
More efficient inorganic carbon uptake	Allowing for high inorganic carbon concentration in cultures will be a technological challenge; therefore traits such as C4 like mechanisms, high activity of carbonic anhydrase, or bicarbonate pumps could be utilized
Wider range of tolerance to CO_2	It is important to tolerate variable CO_2 concentrations and pH. Extremophiles such as *Galderia sulfuraria* and *Cyanidioschyzon merolae* can provide insights
Increasing resistance to photoinhibition	High rates of photosynthesis need to be achieved even at high light intensities. Different strains of cyanobacteria with different light adaptation can guide the search for relevant genes
Increasing tolerance to oxidative stress	Intensive cultures causing increased concentrations of O_2, inducing photorespiration and pseudocyclic photophosphorylation that need to be compensated by efficient antioxidative systems
Increasing thermotolerance	Temperature can be a problem in cultures; *C. merolae* and *G. sulfuraria* could be important thermophilic models
Effective channelling of photosynthetates	Important lessons can be learned from "minimal organisms" such as *Ostreococcus tauri*

in controlled cellular systems – offers extraordinary perspectives in a very promising biotechnology market of several tens of billions dollars according to different sources (Gasdaska et al. 2003, Schmidt 2004). The production systems available today are bacteria, yeasts, animal cells or terrestrial plants genetically modified to ensure the production of insulin, growth hormones, antibodies monoclonal and other therapeutic proteins. Each system has its advantages and disadvantages, among which the costs, the safety of production, the extraction facility, purification and the degree of complexity of the produced molecules. (Leon-Banares et al. 2004; Walker et al. 2005, Cadoret et al. 2008). Restricted to mostly non-marine strains such as *Chlamydomonas* it is only a matter of time before stable transformation systems are developed for marine strains of microalgae. Currently, the stable manipulation of the algal system remains the major limitation but as our genetic knowledge and understanding of the systems increases this will soon be solved. There are examples of successful transformation of green, red and heterokont algae (Cadoret et al. 2008). Stable expression of transgenic proteins in green algae has been shown in the freshwater species *Chlamydomonas reinhardtii, Volvox carteri* (Rosenberg et al. 2008), *Chlorella sp* (Dawson et al. 1997) in the marine species *Dunaliella salina* (Geng et al. 2003), *Haematococcus pluvialis* (Teng et al. 2002), *Ostreococcus tauri*

(pers comm). Transformation with homologous recombination has been demonstrated in the red microalga *Cyanidioschyzon merolae* (Minoda et al. 2004). Among the heterokonts, transformation has been reported in, for example, the diatoms *Thalassiosira pseudonana* and *Phaeodactylum tricornutum* (Poulsen et al. 2006, Kroth 2007) and in the brown kelp *Saccharina (Laminaria) japonica* (Jiang et al. 2002, 2003).

8.4.10 Marine Fungi

Terrestrial fungi are important sources of biomolecules such as antibiotics and there is also a growing interest in marine fungi as a rich source of anticancer, antibacterial, anti-parasite, anti-inflammatory and antiviral agents (see reviews by Bhadury et al. 2006 and Raghukumar 2008). Examples of products from marine fungi include alkaloids with possible anticancer compounds from *Penicillium citrinum* (Tsuda et al. 2004), *Fusarium* sp (Ebel 2006) and *Apiospora montagnei* (Klemke et al. 2004) all isolated from seaweeds. Another example is the polyketide ascosalipyrrolidinone-A isolated from the *Ascochyta salicorniae* (Osterhage et al. 2000) with potential antimalarial activity. One marine fungi, also present in cheese and dairy products, has been sequenced, the marine yeast *Debaryomyces hansenii* (Dujon et al. 2004). This yeast has, for example, been studied for its xylose metabolism (e.g. Sampaio et al. 2004).

8.4.11 Metazoans

One outstanding example of the possibilities of marine biotechnology is the discovery, description and development of fluorescent proteins as important tools in cell biology research. Osamu Shimomura, Martin Chalfie and Roger Y. Tsien were awarded the Nobel Price in chemistry in 2008 for their research on the green fluorescent protein. The first fluorescent protein was discovered in the jellyfish *Aequorea victoria* (Shimomura et al. 1962); the gene for this protein was later cloned by Prasher et al. (1985) and used by Chalfie et al. (1994) to study gene expression in *E. coli* and *C. elegans*. Fluorescent proteins were later found in other marine organisms such as hydrozoans, anthozoans and copepods. These fluorescent proteins have been modified to generate numerous commercial products with different spectral characteristics (Mocz 2007).

Sponges and their associated microorganisms produce numerous compounds with bioactive potential; however supply of biological material has been a limiting factor. The use of an approach of heterologous expression of genes found with metagenomics approach, the cultivation of associated microorganism in the absence of sponges, and the establishment of sponge cell lines have the potential to solve the problem of supply (Wijffels 2008).

Antimicrobial peptides from marine invertebrates have lately attracted attention, for example peptides from *Tachypleus tridentatus* (Japanese horse shoe crab) showed activity against *Leishmania braziliensis* and *Trypanosoma cruzi* (Löfgren et al. 2008), peptides from *Ciona intestinalis* showed antibacterial activity (Fedders et al. 2008), as well as peptides from mussels (Mitta et al. 2000).

8.4.12 In Closing

Clearly, the marine environment offers great potential for biotechnological exploitation. As this review shows, this under-utilised resource has been tapped in to infrequently, but with highly successful results. The recent explosion of interest in marine genomics combined with the societal shift in opinion towards a greener living, reducing carbon and environmental footprints, and increased technological efficiency indicates a bright future for marine biotechnology. With 70% of the planet covered in water, it is obvious that the marine environment will hold the long term answers to our increasing demands on the Earth's resources. The potential has already been shown, what is needed is increased funding for research and a concerted effort by marine scientists to exploit opportunities as they arise. Nature has taught us to expect the unexpected, and there is no doubt that the marine environment will continue to amaze us with its diversity of function. As biotechnologists, we must endeavour to capitalise on this wherever and whenever we can.

References

Allen MJ, Schroeder DC, Donkin A et al (2006a) Genome comparison of two Coccolithoviruses. Virol J 3:15

Allen MJ, Schroeder DC, Holden MT et al (2006b) Evolutionary history of the Coccolithoviridae. Mol Biol Evol 23:86–92

Allen MJ, Wilson WH (2008) Aquatic virus diversity accessed through omic techniques: a route map to function. Curr Opin Microbiol 11:226–232

Amador ML, Jimeno J, Paz-Ares L et al (2003) Progress in the development and acquisition of anticancer agents from marine sources. Ann Oncol 14:1607–1615

Amann RI, Ludwig W, Schleifer K-H (1995) Phylogenetic identification and in situ detection of individual microbial cells without cultivation. Microbiol Rev 59:143–169

Angly FE, Felts B, Breitbart M et al (2006) The marine viromes of four oceanic regions. PLoS Biol 4:e368

Bertram M, Hildebrandt P, Weiner D et al (2008) Characterization of lipases and esterases from metagenomes for lipid modification. J Am Oil Chem Soc 85:47–53

Bhadury P, Mohammad BT, Wright PC (2006) The current status of natural products from marine fungi and their potential as anti-infective agents. J Ind Microbiol Biotechnol 33:325–337

Bielaszewska M, Dobrindt U, Gärtner J et al (2007) Aspects of genome plasticity in pathogenic *Escherichia coli*. Int J Med Microbiol 297:625–639

Bowler C, Allen AE, Badger JH et al (2008) The *Phaeodactylum* genome reveals the evolutionary history of diatom genomes. Nature 456:239–244

Breitbart M, Felts B, Kelley S et al (2004) Diversity and population structure of a near-shore marine-sediment viral community. Proc Biol Sci 271:565–574

Breitbart M, Salamon P, Andresen B et al (2002) Genomic analysis of uncultured marine viral communities. Proc Natl Acad Sci USA 99:14250–14255
Brügger K, Chen L, Stark M et al (2007) The genome of *Hyperthermus butylicus*: a sulfur-reducing, peptide fermenting, neutrophilic Crenarchaeote growing up to 108 degrees C. Archaea 2: 127–135
Burkholder PR, Pfister RM, Leitz FH (1966) Production of a pyrrole antibiotic by a marine bacterium. Appl Environ Microbiol 14:649–653
Cadoret J-P, Bardor M, Lerouge P et al (2008) Les microalgues: Usines cellulaires productrices de molécules commerciales recombinants. Med Sci 24:375–382
Cadoret J-P, Bernard O (2008) La production de biocarburant lipidique avec des microalgues : promesses et défis. J Soc Biol 202:201–211
Chalfie M, Tu Y, Euskirchen G et al (1994) Green fluorescent protein as a marker for gene expression. Science 263:802–805
Chisti Y (2007) Biodiesel from microalgae. Biotechnol Adv 25:294–306
Chu X, He H, Guo C et al (2008) Identification of two novel esterases from a marine metagenomic library derived from South China Sea. Appl Microbiol Biotechnol 80:615–625
Cohen GN, Barbe V, Flament D et al (2003) An integrated analysis of the genome of the hyperthermophilic archaeon *Pyrococcus abyssi*. Mol Microbiol 47:1495–1512
Dawson HN, Burlingame R, Cannons AC (1997) Stable transformation of *Chlorella*: Rescue of nitrate reductase-deficient mutants with the nitrate reductase gene. Curr Microbiol 35:356–362
De Lorenzo V (2005) Problems with metagenomic screenings. Nat Biotech 23:1045–1046
Dean FB, Hosono S, Fang L et al (2002) Comprehensive human genome amplification using multiple displacement amplification. Proc Natl Acad Sci USA 99:5261–5266
Deckert G, Warren PV, Gaasterland T et al (1998) The complete genome of the hyperthermophilic bacterium *Aquifex aeolicus*. Nature 392:353–358
Deng M-D, Coleman JR (1999) Ethanol synthesis by genetic engineering in Cyanobacteria. Appl Environ Microbiol 65:523–528
Dujon B, Sherman D, Fischer G et al (2004) Genome evolution in yeasts. Nature 430:35–44
Ebel R (2006) Secondary metabolites from marine derived fungi. In: Proksch P, Müller WEG (eds) Frontiers in marine biotechnology. Horizon Bioscience, England, pp 73–143
Egorova K, Antranikian G (2007) Biotechnology. In: Garrett RA, Klenk HP (eds) Archaea: evolution, physiology, and molecular biology. Blackwell, Malden
Fedders H, Michalek M, Grötzinger J et al (2008) An exceptional salt-tolerant antimicrobial peptide derived from a novel gene family of haemocytes of the marine invertebrate *Ciona intestinalis*. Biochem J 416:65–75
Fieseler L, Hentschel U, Grozdanov L et al (2007) Widespread occurrence and genomic context of unusually small polyketide synthase genes in microbial consortia associated with marine sponges. Appl Environ Microbiol 73:2144–2155
Fitz-Gibbon ST, Ladner H, Kim UJ et al (2002) Genome sequence of the hyperthermophilic crenarchaeon *Pyrobaculum aerophilum*. Proc Natl Acad Sci USA 99:984–989
Fukui T, Atomi H, Kanai T et al (2005) Complete genome sequence of the hyperthermophilic archaeon *Thermococcus kodakaraensis* KOD1 and comparison with *Pyrococcus* genomes. Genome Res 15:352–363
Fuller RW, Cardellina JH II, Jurek J et al (1994) Isolation and structure/activity features of halomon-related antitumor monoterpenes from the red alga *Portieria hornemannii*. J Med Chem 37:4407–4411
Fuller RW, Cardellina II JH, Kato Y et al (1992) A pentahalogenated monoterpene from the red alga *Portieria hornemannii* produces a novel cytotoxicity profile against a diverse panel of human tumor cell lines. J Med Chem 35:3007–3011
Gasdaska JR, Spencer D, Dickey L (2003) Advantages of therapeutic protein production in the aquatic plant *Lemna*. Bioprocess J Mar/Apr
Geng DG, Wang YQ, Wang P et al (2003) Stable expression of hepatitis B surface antigen gene in *Dunaliella salina* (Chlorophyta). J Appl Phycol 15:451–456
Ghedin E, Claverie JM (2005) Mimivirus relatives in the Sargasso sea. Virol J 2:62

Giovannoni S, Stingl U (2007) The importance of culturing bacterioplankton in the 'omics' age. Nat Rev Microbiol 5:820–826

Glockner FO, Kube M, Bauer M et al (2003) Complete genome sequence of the marine planctomycete *Pirellula sp*. strain 1. Proc Natl Acad Sci USA 100:8298–8303

Gogarten JP, Hilario E. (2006) Inteins, introns, and homing endonucleases: recent revelations about the life cycle of parasitic genetic elements. BMC Evol Biol 6:94

Gonçalves LG, Lamosa P, Huber R et al (2008) Di-myo-inositol phosphate and novel UDP-sugars accumulate in the extreme hyperthermophile *Pyrolobus fumarii*. Extremophiles 12:383–389

Goodwin TJ, Butler MI, Poulter RT (2006) Multiple, non-allelic, intein-coding sequences in eukaryotic RNA polymerase genes. BMC Biol 4:38

Gray JS (1997) Marine biodiversity: patterns, threats and conservation needs. Biodiv Conserv 6:153–175

Han G, Gable K, Yan L et al (2006) Expression of a novel marine viral single-chain serine palmitoyltransferase and construction of yeast and mammalian single-chain chimera. J Biol Chem 281:39935–39942

Handelsman J, Rondon MR, Brady SF et al (1998) Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. Chem Biol 5:R245–R249

Hankamer B, Lehr F, Rupprecht J et al (2007) Photosynthetic biomass and H_2 production by green algae: from bioengineering to bioreactor scale-up. Physiol Plant 131:10–21

Hansen OC, Stougaard P (1997) Hexose oxidase from the red alga *Chondrus crispus*: purification, molecular cloning, and expression in *Pichia pastoris*. J Biol Chem 272:11581–11587

Head IM, Jones DM, Röling WFM (2006) Marine microorganisms make a meal of oil. Nat Rev Microbiol 4:173–182

Henne A, Schmitz RA, Bomeke M et al (2000) Screening of environmental DNA libraries for the presence of genes conferring lipolytic activity on *Escherichia coli*. Appl Environ Microbiol 66:3113–3116

Huber JA, Mark WDB, Morrison HG et al (2007) Microbial population structures in the deep marine biosphere. Science 318:97–100

Ivars-Martinez E, Martin-Cuadrado A-B, D'Auria G et al (2008) Comparative genomics of two ecotypes of the marine planktonic copiotroph *Alteromonas macleodii* suggests alternative lifestyles associated with different kinds of particulate organic matter. ISME J 2:1194–1212

Jeon JH, Kim JT, Kang SG et al (2009) Characterization and its potential application of two esterases derived from the arctic sediment metagenome. Mar Biotechnol 11:307–311

Jiang P, Qin S, Tseng CK (2002) Expression of hepatitis B surface antigen gene (HBsAg) in *Laminaria japonica* (Laminariales Phaeophyta). Chin Sci Bull 47:1438–1440

Jiang P, Qin S, Tseng CK (2003) Expression of the lacZ reporter gene in sporophytes of the seaweed *Laminaria japonica* (Phaeophyceae) by gametophyte-targeted transformation. Plant Cell Rep 21:1211–1216

Kalyuzhnaya MG, Lapidus A, Ivanova N et al (2008) High-resolution metagenomics targets specific functional types in complex microbial communities. Nat Biotechnol 26:1029–1034

Kanai T, Imanaka H, Nakajima A et al (2005) Continuous hydrogen production by the hyperthermophilic archaeon, *Thermococcus kodakaraensis* KOD1. J Biotechnol 116:271–282

Kawarabayasi Y, Hino Y, Horikawa H et al (1999) Complete genome sequence of an aerobic hyperthermophilic crenarchaeon, *Aeropyrum pernix* K1. DNA Res 6:145–152

Kawarabayasi Y, Sawada M, Horikawa H et al (1998) Complete sequence and gene organization of the genome of a hyper- thermophilic archaebacterium, *Pyrococcus horikoshii* OT3. DNA Res 5:147–155

Klarzynski O, Plesse B, Joubert J-M et al (2000) Linear β-1,3 glucans are elicitors of defense responses in tobacco. Plant Physiol 124:1027–1038

Klemke C, Kehraus S, Wright AD et al (2004) New secondary metabolites from the endophytic fungus *Apiospora montagnei*. J Nat Prod 67:1058–1063

Koonin EV (2007) Metagenomic sorcery and the expanding protein universe. Nat Biotechnol 25:540–542

Kroth PG (2007) Genetic transformation: a tool to study protein targeting in diatoms. Methods Mol Biol 390:257–267

Kruse O, Rupprecht J, Bader K-P et al (2005) Improved photobiological H_2 production in engineered green algal cells. J Biol Chem 280:34170–34177

La Scola B, Desnues C, Pagnier I et al (2008) The virophage as a unique parasite of the giant mimivirus. Nature 455:100–104

Langer M, Gabor EM, Liebeton K et al (2006) Metagenomics: an inexhaustible access to nature's diversity. Biotechnol J 1:815–821

Leary D, Vierros M, Hamon G et al (2009) Marine genetic resources: A review of scientific and commercial interest. Mar Policy 33:183–194

Lee S-H, Lee D-G, Jeon J-H et al (2008) Fibrinolytic metalloprotease and composition comprising the same. WO2008056840

Leon-Banares R, Gonzalez-Ballester D, Galvan A et al (2004) Transgenic microalgae as green cell-factories. Trends Biotechnol 22:45–52

Lim JK, Lee HS, Kim YJ et al (2007) Critical factors to high thermostability of an alpha-amylase from hyperthermophilic archaeon *Thermococcus onnurineus* NA1. J Microbiol Biotechnol 17:1242–1248

Löfgren SE, Milettib LC, Steindel M et al (2008) Trypanocidal and leishmanicidal activities of different antimicrobial peptides (AMPs) isolated from aquatic animals. Exp Parasitol 118: 197–202

Maeder DL, Weiss RB, Dunn DM et al (1999) Divergence of the hyperthermophilic archaea *Pyrococcus furiosus* and *P. horikoshii* inferred from complete genomic sequences. Genetics 152:1299–1305

Marsic D, Flaman JM, Ng JD (2008) New DNA polymerase from the hyperthermophilic marine archaeon *Thermococcus thioreducens*. Extremophiles 12:775–788

Mayer AMS, Jacobson PB, Fenical W et al (1998) Pharmacological characterization of the pseudopterosins: novel anti-inflammatory natural products isolated from the Caribbean soft coral, *Pseudopterogorgia elisabethae*. Life Sci 62:PL401–PL407

Melis A, Zhang L, Forestier M et al (2000) Sustained photobiological hydrogen gas production upon reversible inactivation of oxygen evolution in the green alga *Chlamydomonas reinhardtii.* Plant Physiol 122:127–136

Melis A, Happe T (2001) Hydrogen production. Green algae as a source of energy. Plant Physiol 127:740–748

Minoda A, Rei Sakagami R, Yagisawa F et al (2004) Improvement of ulture conditions and evidence for nuclear transformation by homologous recombination in a red alga, *Cyanidioschyzon merolae* 10D. Plant Cell Physiol 45:667–671

Mitta G, Vandenbulcke F, Roch P (2000) Original involvement of antimicrobial peptides in mussel innate immunity. FEBS Lett 486:185–190

Mocz G (2007) Fluorescent proteins and their use in marine biosciences, biotechnology, and proteomics. Mar Biotech 9:305–328

Mueller P, Egorova K, Vorgias CE et al (2006) Cloning, overexpression, and characterization of a thermoactive nitrilase from the hyperthermophilic archaeon *Pyrococcus abyssi*. Protein Exp Purif 47:672–681

Nagasaki K, Shirai Y, Tomaru Y et al (2005) Algal viruses with distinct intraspecies host specificities include identical intein elements. Appl Environ Microbiol 71:3599–3607

Nakagawa S, Takaki Y, Shimamura S et al (2007) Deep-sea vent epsilon-proteobacterial genomes provide insights into emergence of pathogens. Proc Natl Acad Sci USA 104:12146–12150

Nelson KE, Clayton RA, Gill SR, et al (1999) Evidence for lateral gene transfer between Archaea and bacteria from genome sequence of *Thermotoga maritima*. Nature 399:323–329

Newman DJ, Cragg GM (2004) Marine natural products and related compounds in clinical and advanced preclinical trials. J Nat Prod 67:1216–1238

Ogata H, Raoult D, Claverie JM (2005) A new example of viral intein in Mimivirus. Virol J 2:8

Osterhage C, Kaminsky R, Konig GM et al (2000) Ascosalipyrrolidonone A, an antimicrobial alkaloid from the obligate marine fungus *Ascochyta salicorniae*. J Org Chem 65:6412–6417

Pace NR, Stahl DA, Lane DJ et al (1986) The analysis of natural microbial populations by ribosomal RNA. Adv Microbiol Ecol 9:1–55

Perler FB (2002) InBase: the intein database. Nucleic Acids Res 30:383–384

Polle JEW, Kanakagiri SD, Melis A (2003) tla1, a DNA insertional transformant of the green alga *Chlamydomonas reinhardtii* with a truncated light-harvesting chlorophyll antenna size. Planta 217:49–59

Poulsen N, Chesley PM, Kröger N (2006) Molecular genetic manipulation of the diatom *Thalassiosira pseudonana* (Bacillariophyceae). J Phycol 42:1059–1065

Prasher D, McCann RO, Cormier MJ (1985) Cloning and expression of the cDNA coding foe aequorin, a bioluminescent calcium-binding protein. Biochem Biophys Res Com 126: 1259–1268

Quince C, Curtis TP, Sloan WT (2008) The rational exploration of microbial diversity. ISME J 2:997–1006

Rabus R, Ruepp A, Frickey T et al (2004) The genome of *Desulfotalea psychrophila*, a sulfate-reducing bacterium from permanently cold Arctic sediments. Environ Microbiol 6:887–902

Raghukumar C (2008) Marine fungal biotechnology: an ecological perspective. Fungal Divers 31:19–35

Raoult D, Audic S, Robert C et al (2004) The 1.2-megabase genome sequence of Mimivirus. Science 306:1344–1350

Rappé MS, Giovannoni SJ (2003) The uncultured microbial majority. Annu Rev Microbiol 57: 369–394

Robertson DE, Chaplin JA, DeSantis G et al (2004) Exploring nitrilase sequence space for enantioselective catalysis. Appl Environ Microbiol 70:2429–2436

Rodolfi L, Zittelli GC, Bassi N et al (2009) Microalgae for oil: strain selection, induction of lipid synthesis and outdoor mass cultivation in a low-cost photobioreactor. Biotechnol Bioeng 102:100–112

Rondon MR, August PR, Bettermann AD et al (2000) Cloning the soil metagenome: a strategy for accessing the genetic and functional diversity of uncultured microorganisms. Appl Environ Microbiol 66:2541–2547

Rosenberg JN, Oyler GA, Wilkinson L et al (2008) A green light for engineered algae: redirecting metabolism to fuel a biotechnology revolution. Curr Opin Biotechnol 19:430–436

Rubin EM (2008) Genomics of cellulosic biofuels. Nature 454:841–845

Ruby EG, Urbanowski M, Campbell J et al (2005) Complete genome sequence of *Vibrio fischeri*: a symbiotic bacterium with pathogenic congeners. Proc Natl Acad Sci USA 102: 3004–3009

Rusch DB, Halpern AL, Sutton G et al (2007) The Sorcerer II global ocean sampling expedition: northwest Atlantic through eastern tropical Pacific. PLoS Biol 5:e77

Sakurai H, Masukawa H (2007) Promoting R & D in photobiological hydrogen production utilizing mariculture-raised cyanobacteria. Mar Biotechnol 9:128–145

Sampaio FC, Torre P, Passos FML et al (2004) Xylose metabolism in *Debaryomyces hansenii* UFV-170. Effect of the specific oxygen uptake rate. Biotechnol Prog 20:1641–1650

Schenk P, Thomas-Hall S, Stephens E et al (2008) Second generation biofuels: high-efficiency microalgae for biodiesel production. BioEnerg Res 1:20–43

Schirmer A, Gadkari R, Reeves CD et al (2005) Metagenomic analysis reveals diverse polyketide synthase gene clusters in microorganisms associated with the marine sponge *Discodermia dissoluta*. Appl Environ Microbiol 71:4840–4849

Schmidt FR (2004) Recombinant expression systems in the pharmaceutical industry. Microbiol Biotechnol 65:363–372

Schmidt TM, DeLong EF, Pace NR (1991) Analysis of a marine picoplankton community by 16S rRNA gene cloning and sequencing. J Bacteriol 173:4371–4378

Schneiker S, Martins SVA, Bartels D et al (2006) Genome sequence of the ubiquitous hydrocarbon-degrading marine bacterium *Alcanivorax borkumensis*. Nat Biotechnol 24: 997–1004

Schütz K, Happe T, Troshina O et al (2004) Cyanobacterial H_2 production – a comparative analysis. Planta 218:350–359

Sennett SH (2001) Marine chemical ecology: application in marine biomedical prospecting. In: McClintock JB, Baker BJ (eds) Marine chemical ecology. CRC Press, Boca Ratton, FL, pp 523–542

Seshadri R, Kravitz SA, Smarr L et al (2007) CAMERA: a community resource for metagenomics. PLoS Biol 5:S18–S21

She Q, Singh RK, Confalonieri F et al (2001) The complete genome of the crenarchaeon *Sulfolobus solfataricus* P2. Proc Natl Acad Sci USA 98:7835–7840

Sheehan J, Dunahay T, Benemann J et al (1998) A look back at the US Department of Energy's aquatic species program: Biodiesel from Algae. US Report NREL/TP-580-24190 Golden, US Department of Energy: 323

Shimomura O, Johnson FH, Saiga Y (1962) Extraction, purification and properties of aequorin, a bioluminescent protein from the luminous hydromedusan *Aequorea*. J Cell Comp Physiol 59:223–239

Short JM, Marss B, Stein JL (1997) Screening methods for enzymes and enzyme kits. WO9704077 (A1)

Short JM (1999) Protein activity screening of clones having DNA from uncultivated microorganisms. US5958672

Six C, Thomas J-C, Garczarek L et al (2007) Diversity and evolution of phycobilisomes in marine *Synechococcus* spp: a comparative genomics study. Genome Biol 8:R259

Sogin ML, Morrison HG, Huber JA et al (2006) Microbial diversity in the deep sea and the underexplored "rare biosphere". Proc Natl Acad Sci USA 103:12115–12120

Surzycki R, Cournac L, Peltier G et al (2007) Potential for hydrogen production with inducible chloroplast gene expression in *Chlamydomonas*. Proc Natl Acad Sci USA 104:17548–17553

Suttle CA (2005) Viruses in the sea. Nature 437:356–361

Suttle CA (2007) Marine viruses- major players in the global ecosystem. Nat Rev Microbiol 5: 801–812

Takami H, Nakasone K, Takaki Y et al (2000) Complete genome sequence of the alkaliphilic bacterium *Bacillus halodurans* and genomic sequence comparison with *Bacillus subtilis*. Nucleic Acids Res 28:4317–4331

Tamagnini P, Axelsson R, Lindberg P et al (2002) Hydrogenases and hydrogen metabolism of Cyanobacteria. Microbiol Mol Biol Rev 66:1–20

Teng C, Qin S, Liu J et al (2002) Transient expression of lacZ in bombarded unicellular green alga *Haematococcus pluvialis*. J Appl Phycol 14:497–500

Tonon T, Harvey D, Qing R et al (2004a) Identification of a fatty acid Δ11-desaturase from the microalga *Thalassiosira pseudonana*. FEBS Lett 563:28–34

Tonon T, Qing R, Harvey D et al (2005) Identification of a long-chain polyunsaturated fatty acid acyl-coenzyme A synthetase from the diatom *Thalassiosira pseudonana*. Plant Physiol 138:402–408

Tonon T, Sayanova O, Michaelson LV et al (2004b) Fatty acid desaturases from the microalga *Thalassiosira pseudonana*. FEBS J 272:3401–3412

Torsvik V (1980) Isolation of bacterial DNA from soil. Soil Biol Biochem 12:15–21

Tsiroulnikov K, Rezai H, Bonch-Osmolovskaya E et al (2004) Hydrolysis of the amyloid prion protein and nonpathogenic meat and bone meal by anaerobic thermophilic prokaryotes and streptomyces subspecies. J Agric Food Chem 52:6353–6360

Tsuda M, Kasai Y, Komatsu K et al (2004) Citrinadin A, a novel pentacyclic alkaloid from marine-derived fungus *Penicillium citrinum*. Org Lett 6:3087–3089

Uchiyama T, Abe T, Ikemura T et al (2005) Substrate-induced gene-expression screening of environmental metagenome libraries for isolation of catabolic genes. Nat Biotech 23: 88–93

VanFossen AL, Lewis DL, Nichols JD et al (2008) Polysaccharide degradation and synthesis by extremely thermophilic anaerobes. Ann NY Acad Sci 1125:322–337

Venter JC, Remington K, Heidelberg JF et al (2004) Environmental genome shotgun sequencing of the Sargasso Sea. Science 304:66–74

Vestergaard G, Aramayo R, Basta T et al (2008) Structure of the *Acidianus* filamentous virus 3 and comparative genomics of related archaeal lipothrixviruses. J Virol 82:371–381

Walker TL, Collet C, Purton S (2005) Algal transgenics in the genomic era. J Phycol 41:1077–1093

Weiner D, Short JM, Hitchman T et al (2007) P450 enzymes, nucleic acids encoding them and methods of making and using them. US2007231820(A1)

Wijffels RH (2008) Potential of sponges and microalgae for marine biotechnology. Trends Biotechnol 26:26–31

Wilson WH, Schroeder DC, Allen MJ et al (2005) Complete genome sequence and lytic phase transcription profile of a Coccolithovirus. Science 309:1090–1092

Yin Y, Fischer D (2008) Identification and investigation of ORFans in the viral world. BMC Genomics 9:24

Yooseph S, Sutton G, Rusch DB et al (2007) The Sorcerer II Global Ocean Sampling expedition: expanding the universe of protein families. PLoS Biol 5:S56–S90

Yun J, Ryu S (2005) Screening for novel enzymes from metagenome and SIGEX, as a way to improve it. Microb Cell Fact 4:8

Chapter 9
Practical Guide: Genomic Techniques and How to Apply Them to Marine Questions

Virginie Mittard-Runte, Thomas Bekel, Jochen Blom, Michael Dondrup, Kolja Henckel, Sebastian Jaenicke, Lutz Krause, Burkhard Linke, Heiko Neuweger, Susanne Schneiker-Bekel, and Alexander Goesmann

Abstract In recent years, modern high-throughput techniques in genome and post-genome research have made a marked impact on the marine sciences. Today, massively parallel DNA sequencing and hybridization approaches allow the identification of not only the gene repertoire but also the gene regulatory networks that function within an organism. The huge amounts of data acquired from such experiments can only be handled with intensive bioinformatics support that has to provide an adequate infrastructure for storing and analysing these data. Bioinformatics has to deliver efficient data analysis algorithms, user-friendly tools and software applications, as well as extensive hardware infrastructure to deal with these genome-scale analyses.

The following chapter briefly introduces not only the most relevant topics of bioinformatics for functional and structural genomics but also addresses the practical aspects of other steps of a genome project such as sequencing or data management issues. The chapter will take the reader through the different technical approaches that can be applied in marine genomics projects.

In the first part, we will mainly focus on data generation, introducing classical genome sequencing approaches such as the Sanger method and the shotgun technique. Moreover, a short overview of the current status of the next generation of sequencing techniques will be given. In the second part, we briefly introduce the concept of data management for bioinformatics applications. In the third part, we describe the basic principles of genome sequence analysis and address topics like EST clustering and genome assembly, gene prediction, gene function assignment and classification as well as whole genome annotation. In the fourth part of this chapter, we present an overview of transcriptome data analysis using microarray hybridization technology. After a brief introduction to microarray technology we describe state-of-the-art methods for image processing, data normalization, significance testing and cluster analysis.

V. Mittard-Runte (✉)
BRF/Computational Genomics group, CeBiTec, Universität Bielefeld,
D33594 Bielefeld, Germany
e-mail: vrunte@cebitec.uni-bielefeld.de

J.M. Cock et al. (eds.), *Introduction to Marine Genomics*,
Advances in Marine Genomics 1, DOI 10.1007/978-90-481-8639-6_9,

9.1 Sequence Data Generation

In the early days of genetics, scientists did not have the resources to look at more than a few genes at a time. Nowadays microbial genome analysis is in a phase of enormous growth. Since the publication of the first completely sequenced bacterial genome *Haemophilus influenzae* (Fleischmann et al. 1995) in 1995 the genomes of hundreds of bacteria have been sequenced. By January 2008 about 700 complete genomes had been published and 3,250 ongoing genome projects were listed in the Genomes OnLine Database (GOLD) [http://www.genomesonline.org]. Regarding the latest GOLD reference from September 2007 (Liolios et al. 2008), the total number of recorded archaeal and bacterial projects was 1950 projects, while the advent of new sequencing technology platforms such as pyrosequencing has significantly contributed to the increase in the number of new microbial sequencing projects. The GOLD site reports 134 projects using the 454 technology platform as part of a Whole Genome Shotgun (WGS) sequencing project (Liolios et al. 2008).

9.1.1 Classical Genome Sequencing Approaches

Although several different sequencing techniques are available, the DNA sequencing method using chain termination inhibitors reported by Sanger et al. (1977) has remained the basis for genome sequencing for more than 25 years.

9.1.1.1 The Sanger Method

Classical DNA sequencing is accomplished by the chain termination inhibitor method (Sanger et al. 1977). This method essentially involves copying one strand of a piece of DNA using a short DNA primer, which is complementary to the DNA strand you want to sequence, an enzyme called DNA polymerase and the four nucleotides. The Sanger method employs a mixture of normal nucleotides dNTPs and special dideoxy-nucleotides (ddNTPs). These ddNTPs lack a hydroxyl-group at the 3′ carbon of the ribose sugar. This prevents new nucleotides from being added to a DNA strand after a ddNTP has been incorporated. Thus, once a ddNTP is inserted into a growing DNA strand, synthesis of that strand is stopped. In addition, different fluorescent tags are attached to the four types of ddNTPs providing a means of identifying which ddNTP nucleotide has been incorporated (Prober et al. 1987). After many repeated cycles of synthesis all the possible lengths of DNA are represented, each piece of synthesized DNA containing a fluorescent label at its terminus.

Amplified DNA can then be separated according to size using gel electrophoresis. Early Sanger sequencing technology used vertical polyacrylamide gels to carry out this electrophoresis step. These gels need to be prepared manually. Nowadays capillary-electrophoretic sequencers, in which the fragments migrate through a resin within an individual capillary, carry out this step of the process. The machine uses an ultraviolet laser to detect the fluorescent labels that have been incorporated during the polymerisation reaction. The most recent machines, such as the ABI3730XL,

can analyse as many as 96 samples at a time. The signal generated by the laser detection system constitutes a sequence trace file in which the intensities of fluorescence of the four dyes corresponding to the four bases are plotted as a function of speed of migration (which is equivalent to fragment size and therefore provides a position within the sequenced region).

9.1.1.2 Shotgun Technique

The Sanger sequencing method generates sequence data for regions of up to about one kilobase. A strategy is therefore needed to apply this method to the sequencing of an entire genome. For bacterial genomes and small eukaryote genomes whole genome shotgun sequencing is the method of choice. This technique involves fragmentation of the genomic DNA into pieces of a defined length, cloning them into sequencing vectors, and introducing them into an *E. coli* host strain. A random selection of the resulting recombinant clones is then sequenced and the sequence reads are assembled into contiguous regions (contigs) based on sequence overlaps. For larger genomes the hierarchical shotgun approach provides an alternative method (Green 2002). In this approach, a genome is decomposed into larger fragments, for example large fragments cloned into Bacterial Artificial Chromosomes (BACs). The BACs are then ordered into a minimal tiling path using Polymerase Chain Reaction (PCR) or labour-intensive hybridization techniques. The selected subset of clones (the minimum tiling path) are then individually sequenced using a shotgun approach for each piece of DNA (Kaiser et al. 2003).

9.1.1.3 Bacterial Genome Assembly and Finishing

Several bioinformatics tools were developed at the same time as the first genome sequencing projects were carried out. Examples include the basecalling and sequence trace file quality clipping program PHRED (Ewing et al. 1998), the DNA sequence assembly programs PHRAP (Green, 1996) and CAP3 (Huang and Madan, 1999) and the genome sequence finishing program Consed (Gordon et al. 1998). In 2003 the Bioinformatics Resource Facility at the university of Bielefeld developed an optimized approach for whole genome shotgun sequencing (Kaiser et al. 2003), which combined the advantages of fast high throughput shotgun sequence data generation with a sequence finishing and validation phase driven by large insert BAC or fosmid clone libraries. The objective of this approach was to produce high quality bacterial genome sequences in a time- and cost-effective manner. This sequencing strategy consists of two main steps: (i) high-throughput generation of shotgun reads using Sanger sequencing (up to an eight-fold coverage is performed) and (ii) a manually driven sequencing and linking phase, using end sequences of large insert BAC- or fosmid- libraries and finishing reads generated by primer walking.

To monitor the fist step, the high-throughput shotgun sequencing phase, a new bioinformatics tool called SAMS (Sequence Analysis and Management System) was developed (Bekel et al. 2009). SAMS processes raw sequence data (e.g. .scf files) as follows: first a normalizing step involving basecalling and quality clipping

(see also Section 9.3.1) is carried out using PHRED (Ewing et al. 1998), followed by BLAST-based (Altschul et al. 1990) vector clipping (see also Section 9.3.1). To determine the overall progress of the project, subsets of the given sequence data are assembled. Lander-Waterman-like (Lander and Waterman 1988) statistical analysis graphs are created, which plot the number of gaps over the number of sequencing reads using data generated by the CAP3 and PHRAP assembly tools.

For the second sequencing and linking step, involving sequence editing and manual assembly inspection, the Consed software package (Gordon et al. 1998) is used. To link and polish the existing contigs, a tool called Autofinish (Gordon et al. 2001) is applied. The tool BACCardI is used to guide the linking of contigs (Bartels et al. 2005). This latter program is able to automatically generate BAC or fosmid maps.

This optimized approach for whole genome shotgun assembly guided by a bioinformatics pipelines was successfully applied to a variety of complete bacterial genome projects such as those of the Competence Network "Genome research on bacteria relevant for agriculture, environment and biotechnology", which included the genomes of the oil-degrading marine bacterium *Alcanivorax borkumensis* (Schneiker et al. 2006), the plant pathogens *Clavibacter michiganensis* (Gartemann et al. 2008), *Xanthomonas campestris* pv. vesicatoria (Thieme et al. 2005) and pv. campestris B100 (Vorhölter et al. 2008), the plant growth promoting bacterium *Azoarcus* (Krause et al. 2006) and the biotechnologically relevant 13Mbp bacterium *Sorangium cellulosum* (Schneiker et al. 2007).

The high-throughput shotgun sequencing phase is now being superseded by massively parallel sequencing techniques but sequence assembly and genome finishing remain an important issue for genome sequencing projects.

9.1.2 Next Generation of Genome Sequencing

Sanger-based sequencing technology cannot be improved indefinitely. Cloning bias and difficulties with sequencing regions of a genome that exhibit strong secondary structures ("hard stops") limit the quality of the assemblies obtained with the Sanger method. Some of the most promising new sequencing technologies are based on massive parallelization. The emulsion PCR method for in vitro clonal amplification is used in the pyrosequencing technology published by Margulies et al. in 2005 (commercialized by 454 Life Sciences, acquired by Roche), the polony sequencing method (Shendure et al. 2005) and the SOLiD system (developed by Agencourt, and later acquired by Applied Biosystems). Another method called "bridge PCR" is used in the single base extension system distributed by Illumina. The 454 GS20 sequencing platform from 454/Roche has been proven to be efficient for the sequencing of whole bacterial genomes (Goldberg et al. 2006) as well as for environmental samples (Edwards et al. 2006) by eliminating many of the cloning problems that are associated with metagenomics. At present three platforms for massively parallel DNA sequencing are in use. These are the Roche/454 GS FLX system (standard and titanium series) offering longer reads than the GS20 machine, the Illumina/Solexa

Table 9.1 Genome sequencing technologies comparison (next generation technologies according to Millar et al. 2008)

Genome sequencing technologies				
	Second or Next generation			First generation
Company	Roche®	Illumina®	Applied Biosystems®	Applied Biosystems®
Machine	Genome Sequencer FLX (standard series)	Genome analyser	SOLiD gene sequencer	3730xl DNA analyser
	Emulsion PCR of bead anchored oligos and pyrosequencing using light emission	Solid-phase -anchored oligo bridge amplification and sequencing with reversible dNTP terminators	Paired-end oligo cloning. Emulsion PCR of bead-anchored oligos. Fluorescent oligo ligation and detection	Sanger dideoxy chemistry and capillary array electrophoresis-based DNA analyser
Read length	~250 bp	~50 bp	~35 bp	~900 bp
Number of reads/run	400,000	40,000,000	85,000,000 (mate-pair run)	96
Raw data	100 MB/run/7.5 h	1 GB/run/ 67–91 h	1 GB/run/4 days for fragment library, 8 days for paired library	1 MB/day and system
Future prospects	500 bases reads; 1 GB per run (Titanium series, since October 2008 on the market)	> 6 GB per run	50 bp single read; 6 Gb/run	Up to 1,100

Genome Analyser and the Applied Biosystems SOLiD™ system. Table 9.1 provides a comparison of these three massively parallel sequencing technologies with first generation sequencing using the Sanger sequencing technology and capillary electrophoresis as described in Section 9.1.1.

Massively parallel DNA sequencing techniques have helped to reduce the cost and time associated with sequencing a genome. Each method will be presented briefly in the following paragraphs.

Readers who would like to have further information on next generation sequencing approaches are referred to recent reviews such as Mardis (2008) and Millar et al. (2008), which provide a detailed overview of the emerging field of sequencing technologies.

9.1.2.1 Pyrosequencing or 454 Sequencing

The 454 Life Sciences® Corporation, USA has developed a scalable, highly parallel sequencing system. The 454 pyrosequencing method was described in detail by Margulies and colleagues (2005). Briefly, the initial version of the apparatus (GS20) used a novel fibre-optic slide of individual wells and was able to sequence 25 million bases, at 99% or better accuracy, in one 4-h run. The machine used an emulsion method for DNA amplification (called emulsion polymerase chain reaction – short emPCR) (Nakano et al. 2003) of bead-anchored oligonucleotides (Dressman et al. 2003). The technologies applied included pyrosequencing and an instrument for sequencing by synthesis using a pyrosequencing protocol (Ronaghi et al. 1998) optimized for solid support and picoliter-scale volumes (Margulies et al. 2005). In pyrosequencing, each incorporation of a nucleotide by DNA polymerase results in the release of pyrophosphate, which initiates a series of downstream reactions that ultimately produce light by the firefly enzyme luciferase. The amount of light produced is proportional to the number of nucleotides incorporated (Mardis 2008).

For more details regarding the utility, throughput, accuracy and robustness of the 454 system refer to the Margulies et al. publication (2005). The 454 pyrosequencing method has experienced rapid growth since its partnership with Roche Diagnostics and release of its GS20 sequencing machine in 2005 and the Genome Sequencer FLX machine in 2007 (Table 9.1). The new GS FLX system Titanium series generates up to one Million reads and 1 GB of data per one 10 h sequencing run.

For more detailed information and recent developments regarding the 454 sequencing technology and the Roche® GS FLX machine refer to the following website http://www.454.com

9.1.2.2 Illumina® Sequencing Technology

Illumina® sequencing technology is based on massively parallel sequencing of millions of fragments using a reversible terminator-based sequencing chemistry (Ju et al. 2006). The technology relies on the attachment of randomly fragmented genomic DNA to a planar, optically transparent surface, and solid phase amplification also called bridge amplification (Adams et al. 1997, Fedurco et al. 2006) to create an ultra-high density sequencing flow cell with >50 million clusters, each containing ~1,000 copies of the same template. These dense clusters of dsDNA are sequenced using reversible fluorescent dNTP terminators with removable fluorophores yielding about 1 GB of raw data (Table 9.1). This novel approach ensures high accuracy and true base-by-base sequencing, eliminating sequence-context specific errors and enabling sequencing through homopolymers and short repetitive sequences.

After completion of the first read, the templates can be regenerated in situ using the Paired-End Module to enable a second > 36 bp read from the opposite end of the fragment. This paired-end methodology leads to an increase of the yield to > 3 GB of paired-end data.

For more detailed information regarding the Illumina® sequencing technology and the Illumina® Genome Analyser and recent developments please refer to the Illumina® website http://www.illumina.com/. The Illumina sequencing platform includes merged technology from Solexa, Lynx and Manteia SA.

9.1.2.3 SOLiD™ System

Applied Biosystems® (ABI) have developed the SOLiD™ System (Supported Oligonucleotide Ligation and Detection System), which is a genetic analysis platform that enables massively parallel sequencing of clonally amplified DNA fragments linked to beads. The SOLiD™ sequencing methodology is based on sequential ligation with dye-labelled oligonucleotides. The company claims that this technology provides unmatched accuracy, ultra-high throughput and application flexibility. The SOLiD™ System features di-base encoding, a proprietary mechanism that interrogates each base twice. The SOLiD™ System generates more than one gigabase of mapable data per run (Table 9.1). Increased bead yields, improved bead enrichment methods, higher density bead packing, and improved feature-finding software will allow for rapid gains in throughput from the same system. The di-base encoding algorithms filter out errors in the raw data following sequencing, providing built-in error correction.

For more detailed information and recent developments regarding the SOLiD™ technology and the ABI SOLiD™ system refer to the Applied Biosystems® website https://products.appliedbiosystems.com/index.cfm or http://solid.appliedbiosystems.com

9.1.3 Other New Advanced Approaches to DNA Sequencing

In August 2007 the National Institutes of Health (NIH) issued a news release announcing several new grant awards to boost efforts towards next generation sequencing technologies (http://www.genome.gov). Shendure et al. (2008) have provided a recent overview of DNA sequencing strategies, including approaches such as microelectrophoresis and mass spectrometry. Here a short overview is presented of three distinct approaches to DNA sequencing based on two articles from Shendure et al. (2004, 2008): polony sequencing, sequencing-by-hybridization, and nanopore sequencing.

9.1.3.1 Open-Source "Polony Sequencing" System

Using the "polymerase colony (polony) Cyclic Sequencing by Synthesis" method it was possible to resequence an evolved strain of *Escherichia coli* with less than one error per million consensus bases (Shendure et al. 2005). The polony resequencing method involves the amplification of short DNA fragments on one-micrometer magnetic beads using the emulsion polymerase chain reaction (see Section 9.1.2, pyrosequencing part). The beads are then immobilized in a polyacrylamide gel and

subjected to another enzymatic method of sequencing called "sequencing by ligation". The wavelength of fluorescence emission from each bead is detected by a four-colour (e.g., red: adenine, green: cytosine, blue: guanine, yellow: thymine) epifluorescence microscope, which follows each cycle of ligation. The "polony sequencing" method is available as an open-source platform and has also been licensed for further development to Agencourt/Applied Biosystems.

9.1.3.2 Sequencing-by-Hybridization

In contrast to the other methods described so far in this section, sequencing by hybridization is a non-enzymatic method using the differential hybridization of target DNA to an array of immobilized oligonucleotide probes. This technique has recently been used successfully in resequencing approaches.

Affymetrix® arrays were used for resequencing two *Saccharomyces cerevisae* strains (Gresham et al. 2006) allowing the detection of approximately 30,000 known single-nucleotide polymorphisms (representing more than 90%) between the two strains. Another study of five *Escherichia coli* strains was carried out using NimbleGen resequencing arrays (Herring et al. 2006) in order to monitor spontaneous mutations that conveyed a selective growth advantage during adaptation to a glycerol-based growth medium.

9.1.3.3 Nanopore Sequencing

Both Agilent and several academic research groups are developing nanopore sequencing. The aim with this method is to sequence a single molecule of DNA with no need for amplification by driving it through a small channel or nanopore using an applied electric field. This technology is at the development stage and does not yet seem to be capable of measuring single bases accurately.

9.1.4 Conclusion

The field of DNA sequencing is changing rapidly. We have therefore attempted to provide a snapshot of the techniques now available. We have been concentrated on established and widely used technologies, but nonetheless we have mentioned emerging technologies even if these methods are probably far from becoming "standard" sequencing technologies within the next years.

Whatever sequencing technology is used, the development of bioinformatic tools which allow the analysis of the huge amount of sequencing data that will be generated will continue to be a critical factor in the coming years.

The available budget for a particular study and the type of study that will be carried out (de novo, resequencing, metagenomics, gene expression via sequence tags, etc) will also dictate which DNA sequencing technology will be used. A hybrid strategy combining high coverage provided by the new generation of sequencing

technologies, together with a lower coverage using Sanger sequencing is often optimal. This provides on the one hand sequence depth and on the other hand a scaffold for assembly and finishing steps.

9.2 Data Management for Bioinformatics Applications

The ongoing developments of new sequencing technologies result in large amounts of sequence data being generated. Together with the derived data that is created from those sequences, processing of sequence data has become a task that not only requires considerable computing resources, but also a large amount of data storage capacity.

This section will provide an overview of the most important aspects that need to be considered when planning genomic sequencing projects.

9.2.1 Data Modelling and Storage

A structured and well-organized data storage system is essential for data that are frequently accessed. This involves not only the development of a formal data description model, associated metadata and existing entity relationships that have to be stored, but also an estimation of the amount of data that will be generated and the allocation of required storage resources.

In a typical scenario, nucleotide or protein sequences can easily be saved to "flat" files that contain only the sequence data arranged sequentially. Metadata like indices are often created to allow faster access to individual records. But this approach, while perfectly valid for this type of data, is generally unfeasible for other information like annotation data for genomes. When numerous relations exist within the data, a relational database management system (Codd 1990) or another type of structured storage is likely to be more appropriate.

For data that are frequently updated and changed, a centralized storage system is preferred over keeping distributed copies of the data locally on each system. However, such a centralized storage system might easily become a bottleneck when stored data have to be accessed by a large number of systems (e.g. in a cluster environment with many computers accessing a central sequence database). So it is necessary to consider all the potential downsides and pitfalls that any solution might have.

For data that are infrequently accessed, it might even be easier to recreate them on demand and not to store them at all. Another option is to extract the relevant details and to only store a subset of the original data for further processing.

Another important aspect that has to be considered is the availability of the data. This includes not only implementing an access control scheme, which defines by whom and how the data may be accessed, but also making a decision about the

retention time for the data. It is crucial to determine how long the storage of the data will be required and what will happen once that period expires. The possible options range from deletion of the data to the transfer to another medium for long term archiving.

9.2.2 Data Access

Access to stored data can happen by various means. The data might only be processed by locally installed applications, or it could be made available to non-local users through a web site or some kind of application programming interface (API) such as Web Services.

Typically, data access is provided by standalone applications installed on computer systems that process or extract the desired information using locally stored data such as sequence databases or other data sources like microarray data.

While straightforward, this approach has several disadvantages: both the software itself and the associated data impose a certain maintenance cost on the user, e.g. when newer versions of the software become available and updates need to be deployed, or with software relying on frequently changing databases. This maintenance imposes an increasing problem when such tools have to be used in environments involving different system architectures and operating systems.

In such a typical environment with different operating systems and software being used web-based applications allow for platform-independent collaboration among researchers. This allows them to work on common projects with the data being stored in a centralized location. Furthermore, web interfaces are commonly used to make newly developed tools available to the public in situations where publishing the tool itself is not desirable (e.g. due to license restrictions) or impractical (tools operating on large or frequently updated data).

Web Services provide a standardized method to access information or perform computations over a network via the exchange of XML-based messages. As a simple means of accessing remote resources, Web Services are becoming increasingly popular in the bioinformatics field, allowing computationally expensive calculations to be executed without having to provide the necessary hardware infrastructure, or providing access to e.g. large sequence databases without the necessity of having to store the data locally (Fig. 9.1). In contrast to web-based interfaces, Web Services can easily be integrated into other applications, thus extending their functionality. Since Web Services can be accessed from almost any system with a network connection, they combine the advantages of web-based applications with a lower maintenance cost in comparison with locally installed tools. The large amount of different publicly offered Web Services gives researchers easy access to a wide range of tools and information, helping them to analyse and evaluate their data. The BioMOBY system (Wilkinson and Links 2002) for example, integrates a wide range of data sources and data analysis tools that can be accessed as web services.

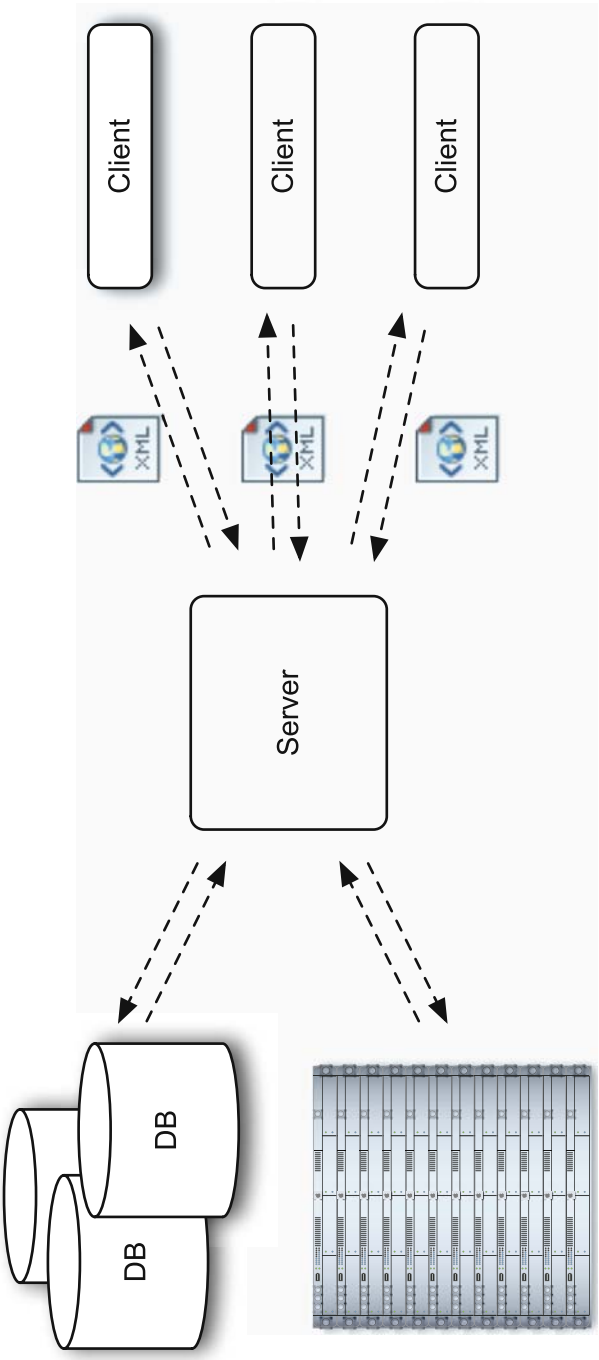

Fig. 9.1 The Web Services server provides access to stored data or compute resources. Clients can access these resources by exchanging XML-based messages with the Web Services server

9.2.3 Common File Formats

Various different file formats have been developed in order to fulfil the requirements of bioinformaticians and to provide a practical opportunity to share data between different parties.

The FASTA format, originally specified by the National Center for Biotechnology Information (NCBI), has become the de-facto standard for the exchange of raw sequence data. A FASTA file contains one or several sequences, with each sequence being preceded by a single description line initiated with the ">" character. The subsequent lines contain the actual sequence data in human-readable format; each nucleotide or amino acid is represented by a single letter following the IUBMB/IUPAC standard code (http://www.chem.qmul.ac.uk/iupac/jcbn/).

Other commonly used file formats include the EMBL and GenBank format. Both can contain more than simply the sequence data and are typically used to store e.g. nucleotide or protein sequence data of a genome together with relevant annotation information (see Section 9.3.5.1).

Several frameworks such as BioJava or BioPerl (Mangalam 2002) have been developed in the bioinformatics field for the most common programming languages, which offer an easy and convenient way to access and manipulate data contained in these file formats.

9.3 DNA Sequence Analysis

In this section, we will present some bioinformatics tools available to marine biologists that want to use genomic approaches. Firstly we will consider eukaryotic EST (Expressed Sequence Tag) processing as bacterial genome sequencing has already been presented in the sequence data generation chapter. Then we will focus on gene prediction, one of the first steps of the annotation of newly sequenced genomes. The next step in working with genetic sequences is the prediction and analysis of the functions and roles a gene or assembled EST might have. We use multiple methods and information resources for sequence annotation such as InterPro. We will then continue with an introduction to comparative genomics and functional classification, where the goal is to identify functions of particular regions of sequence data. We will close this section by presenting the major public sequences databases and other resources, which aim to provide a comprehensive coverage of sequences and annotation available to the scientific community.

9.3.1 EST Processing

Many genome projects generate thousands of Expressed Sequence Tags (ESTs) or shotgun reads. ESTs are generated by reverse transcribing mRNA into complementary DNA (cDNA), which is subsequently sequenced. ESTs provide a fast and inexpensive manner to identify segments of DNA (a few hundred nucleotides) that

code for proteins and are expressed in a certain cell state or cell type. Several steps are necessary to provide high quality sequences, as well as to get an overview of their content. The analysis of EST sequences consists of the following steps:

- Pre-processing
- Clustering and assembly
- Functional analysis

(a) General Concept

EST pre-processing: After the sequencing step, the raw EST sequences are stored as sequence trace files. These files are converted into FASTA files. The trimming procedure removes low quality sequences at both ends of a read (quality clipping).

After quality clipping, the remaining parts of the cloning vector have to be removed from the sequences, so that only the transcribed regions of the gene are left. Therefore a BLAST search against a special database (containing vector sequences) is performed and the vector sequences are removed (vector clipping).

EST clustering and assembly: To reduce redundancy, the EST sequences are grouped (clustered) based on comparisons carried out at the DNA sequence level using a clustering tool (e.g., TGICL – a clustering tool from The Institute for Genome Research – TIGR) (Pertea et al. 2003). Different parameters have to be selected for the clustering, including the length of the high-scoring segment pair (HSP), the unmatched overhang length of the sequences, and a value for the identity of the HSPs (see Fig. 9.2).

All EST sequences are searched for homology against all other ESTs using MegaBLAST (Zhang et al. 2000) to determine which ESTs belong to each cluster. The clusters are then assembled into Tentative Consensus sequences (TCs) (assembly). This procedure is carried out by a bioinformatics tool like CAP3 (Huang and Madan 1999). It is possible to cluster and assemble ESTs from more than one EST library together, so that genes occurring in different libraries (and therefore expressed in different tissues/under different conditions) are assembled into one TC. Figure 9.3 shows a schema of the clustering and assembling of ESTs to TCs. The resulting TCs can be analysed functionally subsequently.

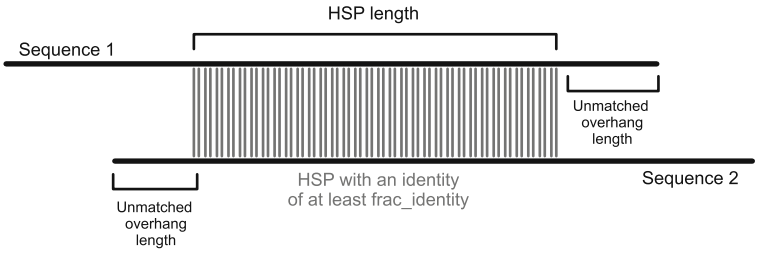

Fig. 9.2 This figure shows the overlap of two EST sequences. For the clustering procedure, the high-scoring segment pair (HSP) length, the length of the unmatched overhang and the percentage of the identity of the sequences have to be defined

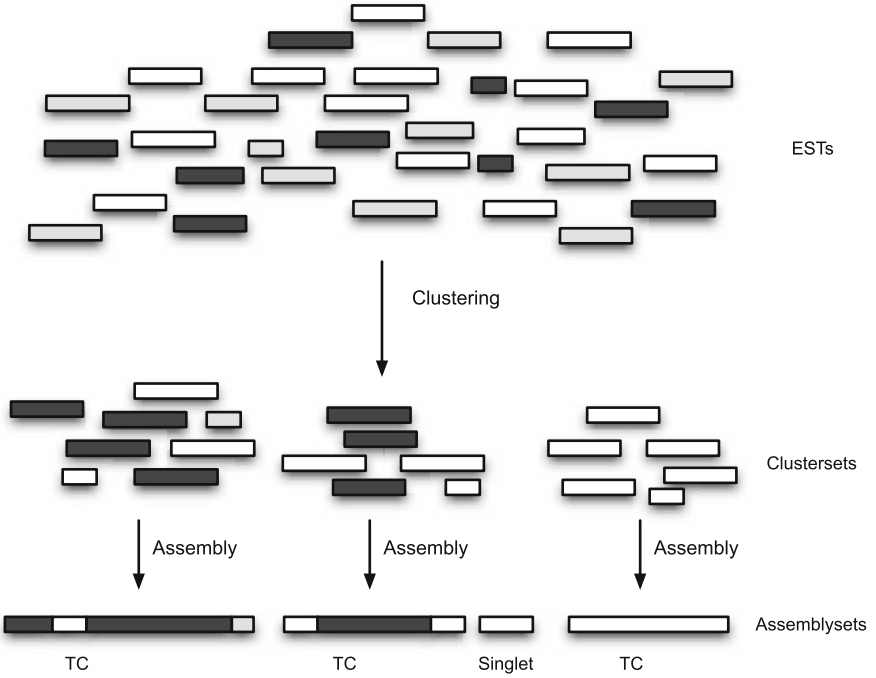

Fig. 9.3 This figure shows the processing steps to create TCs from the ESTs. The ESTs are clustered according to the similarity of their nucleotide sequence. The clusters are assembled afterwards to produce TCs plus singletons. The different colours indicate the different EST libraries

Functional analysis: Functional gene analysis is a technique to predict the function of a certain gene. The tools for the prediction are mostly homology based (e.g. Blast or InterProScan) and search within existing databases (like SwissProt or NR, the Non-redundant protein database at NCBI) for genes that have similar structures. The genes are then functionally annotated, assuming that similar genes have similar functions.

(b) SAMS: An EST Application Tool

SAMS (Sequence Analysis and Management System) is an annotation system for sequence data. It provides an environment that allows a variety of tools to be run on each read/EST and presents the results on a web interface (Bekel et al. 2009). The SAMS EST pipeline covers the four essential parts of EST analysis: preprocessing, clustering, assembly and functional analysis. The parameters of the pipeline tools can be adjusted and new tools can be integrated. As the functional analysis is the essential part, several graphical and interactive options are integrated into SAMS, such as KEGG-maps (Kyoto Encyclopedia of Genes and Genomes, see Section 9.3.4.4) displaying the genes found in the assembly, or an expression analysis tool (SteN – Statistical electronic Northern Blot) for differentially

expressed genes. The sequence and annotation data can be exported in various file formats. SAMS uses an object-oriented backend and a relational database management system MySQL, connected by an object relational mapping developed by the Bioinformatics Resource Facility at Bielefeld University.

Users need to have a username and password. Access to the project can be restricted to any number of users, so that the data are kept confidential.

SAMS can be accessed via a web browser (http://www.cebitec.uni-bielefeld.de/groups/brf/software/sams/). Passwords can be requested via the login page.

Using SAMS for the analysis of EST sequences is possible in an easy and comfortable way. Firstly, a library needs to be selected for the sequences. The user creates or specifies a library in which to store the sequence data. Then the sequences are imported using the import function. This involves several consecutive steps. First, the user selects the locally stored file to upload to the SAMS server. Several formats are supported but raw data files are recommended. In the second step, raw chromatogram files or trace files are uploaded and the program PHRED is used for the quality clipping. The user can select the quality that is required (e.g., PHRED 13 gives a base-calling accuracy of 95%, the probability that the base called is wrong is 1 in 20; PHRED 20 gives you 99% and a probability of 1 in 100). Finally, if the sequences still contain remaining vector sequences, the user has the possibility to remove them using the integrated vector clipping option.

Note that sequences shorter than 50 bp are removed after these procedures. Once the import is finished, the sequences will be listed as ESTs in the SAMS library. The user can then start a clustering procedure to reduce redundancy in the sequence data. For that, standard parameters are suggested, but can be changed if required. The standard clustering parameter set is the TIGR default parameter set (HSP length: 40, identity: 0.95, unmatched overhang length: 20). SAMS uses a clustering component, which is a TGICL-like implementation of the TIGR clustering approach. After clustering, the sequence data is assembled into Tentative Consensus sequences (TCs) using the program CAP3. In SAMS it is possible to cluster and assemble the same sequence data more than once to test different parameter settings. When the user is satisfied with the results of the clustering, the assembly task needs to be performed in order to visualise the TCs and singletons, which are stored in the database.

Functional analysis in SAMS: The goal of this part of the analysis is to achieve the annotation of TCs. An automatic annotation pipeline called Metanor (Goesmann et al. 2005) can be used for this, after the assembly procedure has been completed. This pipeline consists of several bioinformatics tools that are run for each TC, EST and singleton. The tools use BLAST (Basic Local Alignment Search Tool) to search for homology between the sequences and sequences contained in different databases, like the NT (Non-redundant nucleotide database from NCBI), NR (Non-redundant protein database from NCBI), KEGG, KOG/COG (clusters of euKaryotic Orthologous Groups or Clusters of Orthologous Groups, see Section 9.3.4.4), and SwissProt databases (see Section 9.3.5.2). In addition, InterProScan (see Section 9.3.3.3) is performed for all six reading frames. The results are stored as “observations” in the database for further usage.

Once all the tools have been run, an automatic consensus annotation is generated, providing the EC number, gene name, KOG/COG groups and putative gene functions. TCs, ESTs and singletons can be manually annotated using an annotation dialog window, which displays all the results of the previously calculated tools. Backups of all annotations are kept and previous annotations can be restored within seconds. All sequences, annotations and tool results can be exported via the SAMS web interface in various file formats.

SteN – Statistical electronic Northern Blot: SteN is a tool to analyse gene expression using EST data. It compares the number of ESTs from each of several different libraries that have been incorporated into TCs. Using SteN it is possible to filter TCs according to the composition of ESTs from different libraries. In this way it is possible to identify TCs for the genes that are expressed specifically in certain libraries. EST collections from several cDNA libraries corresponding to different time-points and/or tissues are needed for this analysis. After the TC filtering procedure, a statistical evaluation of the expression data is carried out and a list of results is presented.

An alternative to SAMS: ESTexplorer: An alternative to SAMS is the ESTexplorer application (Nagaraj et al. 2007), freely available for academic use at the following website: http://estexplorer.els.mq.edu.au/estexplorer/main_page.php.

Once the data have been uploaded, vector clipping, repeat masking, and assembly procedures are performed with the following tools: SeqClean (Chen et al. 2007), RepeatMasker (http://www.repeatmasker.org), and CAP3. Several bioinformatics tools for annotation can be run using BLAST and Blast2GO (Conesa et al. 2005) at the gene level, ESTSCAN (Iseli et al. 1999), InterProScan (see Section 9.3.3.3) and KOBAS (Wu et al. 2006) at the protein level. The results are stored and once the pipeline is completed, the results can be downloaded for 1 week, all files being deleted after this time. There is no user management system to identify the user via a login and password. Instead each user gets an id to access his files (like John_123). It should be noted that there is a danger with such a system that, if a user guesses the id of another user, the complete dataset can be accessed. The quantity of data that can be uploaded and analysed is limited. A comparison between SAMS (Bekel et al. 2009) and ESTexplorer (Nagaraj et al. 2007) can be found in Table 9.2.

Table 9.2 Comparison SAMS vs. ESTexplorer

SAMS vs. ESTexplorer		
Function\system	SAMS	ESTexplorer
User authentification	Yes	No
Permanent data storage	Yes	No
Clustering and assembly	Yes	Yes
Automatic annotation	Yes	Yes
Manual annotation	Yes	No
Export of sequences and annotations	Yes/yes	Yes/no
Expression analysis	Yes	No

Within Marine Genomics Europe (MGE), SAMS was successfully used to analyse 44 EST projects, including ESTs from the following species: *Fucus serratus* and *Fucus vesiculosus*, *Dicentrarchus labrax* (Sea bass), *Emiliania huxleyi*, or *Balanus amphitrite*.

9.3.2 Gene Prediction

The prediction of tRNA, rRNA, and protein encoding genes from raw genomic sequences is one of the first essential steps during the annotation of newly sequenced genomes. This section focuses on the computational strategies and existing software for the automated identification of protein encoding genes (coding sequences, CDSs) in genome sequences. For gene prediction, two different approaches are applied: intrinsic and similarity based methods. Intrinsic methods analyse sequence properties of genomes to discriminate between coding and non-coding regions. These methods exploit the different compositional properties of coding and non-coding sequences, mainly caused by a bias in codon usage in CDSs, which optimizes the translation efficiency in protein biosynthesis (Gouy and Gautier 1982).

Intrinsic gene finding methods frequently employ a statistical model representing the frequencies of short oligonucleotides[1] in coding and non-coding sequences. Generative models with Markov properties (Durbin et al. 1998), such as fixed-order Markov chains on nucleotides (Delcher et al. 1999, Larsen and Krogh 2003) or codons (Badger and Olsen 1999), are often applied to represent the sequence composition. Additional features, such as ribosome binding sites or overlaps between adjacent genes, can also be integrated into a probabilistic framework, if hidden Markov models (HMMs) are used to describe the context of a gene (Delcher et al. 1999).

Similarity based methods predict genes by searching for stretches of DNA sharing a significant similarity to reference sequences, including genomes from close relatives, phylogenetically related proteins, or gene transcripts. Several similarity based approaches discriminate conserved coding sequences from conserved non-coding regions based on their synonymous substitution rate (Badger and Olsen 1999, Moore and Lake 2003, Nekrutenko et al. 2003). This is based on the fact that, to maintain the amino acid sequence of an encoded protein, coding sequences show a much higher number of synonymous mutations, i.e. mutations that do not modify the encoded amino acid, than conserved non-coding regions.

9.3.2.1 Gene Finding in Prokaryotes

In prokaryotes (Bacteria and Archaea), protein encoding genes are open reading frames (ORFs), which are sequences of codons beginning with a start codon, ending with a stop codon, and without an internal stop codon. The task of predicting protein

[1] Usually frequencies of oligonucleotides of length between 3 and 12 bp are modelled.

encoding genes in prokaryotic genomes can therefore be regarded as a two class classification problem: Discriminating coding sequences from the majority of non-coding ORFs, which correspond to genomic regions that are not transcribed and translated in vivo.

For prokaryotes, many different gene finding software tools have become available that use either intrinsic or a combination of intrinsic and similarity based methods to automatically discover genes from raw genomic sequences. While these methods in general achieve impressive accuracy values, the correct identification of short genes, genes with atypical sequences properties and translation start sites are still challenging tasks. Short genes are more difficult to identify because their sequences carry less information that can be evaluated for identification (Skovgaard et al. 2001, Larsen and Krogh 2003, Ou et al. 2004). Moreover, genes from the same organism may exhibit divergent sequence properties, owing to expression dependent codon usage (McHardy et al. 2004b), leading/lagging strand-related biases (Lafay et al. 1999), or the transfer of genes between different bacterial species, called horizontal gene transfer (Smith et al. 1992). In such cases, intrinsic methods may have difficulties in identifying the different gene classes. This issue has been addressed by the inclusion of an additional model for genes with "atypical" sequence composition or by the unsupervised discovery of CDS classes prior to the prediction phase (Lukashin and Borodovsky 1998, Krause et al. 2007).

Among the most widely used intrinsic gene finding programs are Glimmer (Delcher et al. 1999, 2007) and Genemark (Lukashin and Borodovsky 1998). Glimmer is fast, easy to install locally, and detects about 99% of "certain" genes with known functions on average. However, Glimmer-2 has been reported to produce a rather high number of false positive predictions (McHardy et al. 2004a, Krause et al. 2007). In the latest version (Glimmer-3) this has been addressed by selecting a set of highest-scoring predictions consistent with the maximal allowed overlap during a post-processing phase. Genemark, which uses a hidden Markov model (Besemer et al. 2001, Besemer and Borodovsky 2005) also achieves a high prediction accuracy, with a sensitivity of up to 99% and a specificity[2] of about 93% (Delcher et al. 2007). The program can easily be executed via a web-interface.

Similarity based methods in general are much slower and more difficult to install than intrinsic approaches but more reliable predictions are obtained. Moreover, in a combined approach, similarity supported predictions provide a reliable initial training set to train a genome-specific intrinsic model. For example, by combining a Pfam-based search for conserved protein family members with an intrinsic approach, the gene finder GISMO is able to identify up to 99% of genes with known functions (Krause et al. 2007). GISMO also produces highly reliable predictions with a specificity of more than 94% and achieves high accuracy levels for the identification of short genes and for finding genes in GC-rich genomes. For intrinsic

[2]Specificity measures the reliability of the predictions. It is defined as the fraction of correct gene predictions, i.e. the fraction of predicted genes that corresponds to known genes.

gene identification, GISMO employs a Support Vector Machine, which is able to learn sequence properties of different classes of genes in an unsupervised manner.

Several programs have been devised for the prediction of translation start sites, including GS-Finder (Ou et al. 2004), RBS-Finder (Suzek et al. 2001), and Tico (Tech and Meinicke 2006), which have all achieved accuracy values of more than 90%. However, owing to the limited number of experimentally confirmed translation initiation sites that were available for a performance evaluation, these accuracy values cannot be generalized.

The online resources REGANOR (Linke et al. 2006) and RAST (Aziz et al. 2008) provide easy means to automatically predict tRNA, rRNA and protein encoding genes in prokaryotic genomes via a web-interface. REGANOR achieves sensitivity values of up to 98% for "certain" genes with known gene function and a specificity of 95% (McHardy et al. 2004a) by combining the gene finders Glimmer-2 (Delcher et al. 1999) and CRITICA (Badger and Olsen 1999). The RAST server on the other hand combines evidence from different sources to identify protein-encoding genes, including Glimmer-2 predictions, homology searches for conserved protein families (FIGfams), and a BLAST search versus a large protein database. Additionally, RAST provides an automated functional annotation of all genes that are detected. Both the REGANOR and RAST server employ the programs tRNAscan-SE (Lowe and Eddy 1997) and SearchForRNAs (Niels Larsen et al., unpublished), which allow the automated identification of tRNA and rRNA genes in raw genomic sequences. The REGANOR server can be accessed at https://www.cebitec.uni-bielefeld.de/groups/brf/software/reganor/, the RAST server at http://rast.nmpdr.org/.

In summary, the gene finding web-servers REGANOR and RAST provide high quality gene predictions and are easy to use. The public gene finders GLIMMER and GISMO on the other hand can be installed locally, enabling their integration into existing genome annotation pipelines. While Glimmer is easy to use and time-effective, GISMO is slower and more difficult to maintain, but provides highly accurate gene calls. tRNAscan-SE (Lowe and Eddy 1997), SearchForRNAs and RNAmmer (Lagesen et al. 2007) allow the automated identification of tRNA and rRNA genes (Table 9.3).

9.3.2.2 Gene Finding in Eukaryotes

Although improvements in eukaryotic gene finding programs have been made during recent years, prediction of gene structure in eukaryotic genomes is still a highly difficult task. The main challenges are the complex structure of eukaryotic genes and the low fraction of chromosomes corresponding to protein encoding exons (Mathe et al. 2002, Zhang 2002, Brent 2007). While about 90% of prokaryotic genomes encode proteins, eukaryotic exons are embedded in vast amounts of non-coding DNA. The task is further complicated by the fact that eukaryotic genes may have alternative splice and polyadenylation sites, as well as alternative transcription and translation initiation sites.

Table 9.3 Existing bacterial gene finding software. The implemented strategy is depicted as I (intrinsic) and S (similarity based), the availability as D (available for download) or O (accessible online via web-interface)

Gene finding programs					
Program	I	S	Comments	A	URL (in 12.2008)
Critica (Badger and Olsen 1999)	+	+	Identifies genes based on synonymous substitution rate	D	http://www.ttaxus.com/software.html
EasyGene (Larsen and Krogh 2003)	+	+	Employs HMMs, training set derived by BLAST search	O	http://www.cbs.dtu.dk/services/EasyGene/
GeneMark.hmm/S (Lukashin and Borodovsky 1998, Delcher et al. 1999)	+	–	Uses HMMs	O	http://exon.gatech.edu/GeneMark/
Gismo (Krause et al. 2007)	+	+	Combines Pfam search with SVM	D	http://www.cebitec.uni-bielefeld.de/brf/gismo/gismo.html
Glimmer-3 (Delcher et al. 1999)	+	–	Employs interpolated context model. Uses dynamic programming to reduce overlapping genes and to refine gene starts	D	http://www.cbcb.umd.edu/software/glimmer/
MetaGene (Noguchi et al. 2006)	+		Employs pre-trained, GC content dependent codon-usage model. Applicable also for metagenomic fragments and draft genomes	D+O	http://metagene.cb.k.u-tokyo.ac.jp
Rast (Aziz et al. 2008)	+	+	Combines Glimmer-2 with homology searches	O	http://rast.nmpdr.org/
Reganor (McHardy et al. 2004b)	+	+	Combines Glimmer-2 and Critica	O	https://www.cebitec.uni-bielefeld.de/groups/brf/software/reganor/
RescueNet (Suzek et al. 2001)	+	+	Uses SOM	D	http://bioinf.nuigalway.ie/RescueNet/
SearchForRNAs (Niels Larsen et al., unpublished)			Identifies rRNA genes		Available from author upon request.
tRNAscan-SE (Lowe and Eddy 1997)			Predicts tRNA genes using covariance models	D+O	http://lowelab.ucsc.edu/tRNAscan-SE/
ZCurve (Guo et al. 2003)	+	–	Employs LDA. Performs Z-transformation of DNA	D	http://tubic.tju.edu.cn/Zcurve_B/

SVM: Support Vector Machine; SOM: Self-Organizing Map

At present, the most reliable gene structure predictions are obtained by mapping the sequence of gene transcripts onto the genome of interest. Many different programs have been devised for this task, including EST_GENOME (Mott 1997), AAT (Huang et al. 1997), SIM4 (Florea et al. 1998), GENESEQUER (Usuka et al. 2000), BLAT (Kent 2002), and GMAP (Wu and Watanabe 2005). The main disadvantage of this approach is that genes that are expressed at a low level or those expressed only in specific tissues or under specific conditions are likely to be missed.

Dual or multi-genome gene finding algorithms, such as SGP2 (Parra et al. 2003), SLAM (Alexandersson et al. 2003), TWAIN (Majoros et al. 2005), N-SCAN (Gross and Brent 2006), and TWINCAN (Korf et al. 2001) analyse the pattern of sequence conservation between two or more genomes of evolutionary related organisms. In cases where gene transcripts have not been sequenced, but genomes of close relatives are available this class of programs can provide accurate exon predictions. However, dual and multi-genome gene finding methods may achieve only a modest accuracy for the prediction of complete gene structures, miss genes without sequence homologies, and frequently mistake pseudogenes as functional. The latter problem has recently been addressed by combining a comparative method with a pseudogene detection using PPFINDER (van Baren and Brent 2006). If the sequence of a closely related genome is not available, gene finding methods using alignments with known proteins can be used to reliably identify exon locations (Huang et al. 1997, Slater and Birney 2005). The popular GENEWISE (Birney et al. 2004) for example employs a Hidden Markov model that incorporates a model for the alignment of protein sequences to a genome and a model of eukaryotic gene structure.

To complement alignment based approaches, intrinsic methods allow the identification of genes based on the evaluation of compositional sequence properties. The majority of these programs, such as GENSCAN (Burge and Karlin 1997), GENEMARK.hmm (Lomsadze et al. 2005), GLIMMERHMM (Majoros et al. 2004), FGENESH (Salamov and Solovyev 2000) and AUGUSTUS (Stanke and Waack 2003), employ HMMs or generalized HMMs (GHMMs) to partition genomic sequences into introns, exons, and intergenic regions. By using HMMs or GHMMs with states for introns, exons, intergenic regions, start and stop codons, splice sites, and polyadenylation signals, diverse sequence features can be combined into a coherent, probabilistic model. Before intrinsic methods can be applied to novel genomes, they need to be trained so that they learn the genome specific compositional sequences properties, such as codon usage and splice site patterns. While most intrinsic methods require supervised training on known genes, SNAP (Korf 2004), and GENEMARK.hmm ES (Lomsadze et al. 2005) are able to discover sequence properties of novel genomes in an unsupervised manner. For many of the already sequenced eukaryotic genomes, pre-trained intrinsic models are publicly available. In general, intrinsic gene finding programs perform fairly well when the identification of exons is considered, but they are still far from perfect when it comes to reconstructing the exon-intron structures of complete genes. Furthermore, owing to the high number of false positives produced, predictions that are not supported by external evidence, such as sequence similarities, are unreliable.

Several eukaryotic gene finding algorithms are able to exploit different types of evidence. Programs like N-SCAN, N-SCAN_EST (Wei and Brent 2006), TWINSCAN, AUGUSTUS, and GENIE (Reese et al. 2000) employ HMMs to incorporate intrinsic and extrinsic information into a single probabilistic model. The eukaryotic gene finder EUGENE (Schiex et al. 2001), which has been used for the annotation of *Arabidopsis thaliana*, employs a comparable approach. In EUGENE, intrinsic and extrinsic information is combined in a weighted directed acyclic graph, the most likely gene structure of the analysed DNA sequence is determined by the shortest path in that graph. Other programs, such as JIGSAW (Allen and Salzberg 2005), EXONHUNTER (Brejova et al. 2005), GLEAN (Elsik et al. 2007), EXOGEAN (Djebali et al. 2006), EVIDENCEMODELER (Haas et al. 2008), and AGUSTUS-any (Stanke et al. 2006), mimic the human gene identification process by evaluating evidence from diverse sources.

The accuracy of eukaryotic gene finding methods in general increases as the number and size of introns decreases. Accuracy levels of up to 70% are achieved for the de novo prediction of complete gene structures in compact genomes using comparative methods. For genomes of mammals with high proportions of intergenic regions, complex gene structures and alternative splice sites, usually a considerably lower accuracy is obtained.

In 2005 the EGASP project was launched to evaluate the accuracy of existing gene finding methods on the ENCODE regions of the human genome (Guigo and Reese 2005, Guigo et al. 2006). The best programs correctly predicted at least one transcript for almost 70% of the annotated genes. However, when alternative splicing variants were taken into account, the accuracy dropped to values between approximately 40 and 50%. At the coding exon level, the best evaluated methods achieved a sensitivity and specificity of more than 80%, and close to 90% at the coding nucleotide level. Programs relying on diverse information, in particular sequences of gene transcripts or proteins, were the most accurate, followed by methods relying on sequence comparisons across two or more genomes. Algorithms evaluating only intrinsic sequence features achieved the lowest accuracy. The performance for predicting non-coding exons was low for almost all evaluated programs.

To conclude, the best contemporary eukaryotic gene finding methods achieve good accuracy levels in general for the identification of exons, but the prediction of complete gene structures in complex genomes is still highly challenging. To obtain predictions of the highest possible quality it is recommended to combine programs that detect repetitive regions and pseudogenes with methods that rely on a mapping of gene transcripts and dual, multi-genome, and intrinsic gene finders. Several popular gene finding algorithms can easily be run via a public web-interface, including: AUGUSTUS (http://augustus.gobics.de/), GENEMARK (http://exon.gatech.edu/GeneMark/), TWINSCAN/NSCAN (http://mblab.wustl.edu/software/twinscan/), GLIMMERM (http://www.tigr.org/tdb/glimmerm/glmr_form.html) and GLIMMERHMM (http://nbc9.biologie.uni-kl.de/framed/left/menu/auto/right/glimmerhmm/).

9.3.3 Genome Annotation and Beyond

With either gene prediction or EST assembly being finished, the next step in working with genomic sequences is the prediction and analysis of the functions and roles a gene or assembled EST might have. Almost all *in-silico* methods are based on the correlation between gene sequences, their corresponding amino acid sequence, its two- and three-dimensional structure and the function the protein may carry out. Since a large number of sequences from different sources are already known and their function has been identified and described in scientific publications, searching for similar sequences has proven to be a good method for analysing novel genes and genomes.

9.3.3.1 Introduction to Sequence Similarity

Comparing sequences requires an exact measure for the level of similarity and the degree of difference between them; without this measure a computer is unable to automate the processing. In Bioinformatics "alignments" are used to describe the similarity of sequences. They represent the necessary steps to convert one sequence into another sequence. Each single step may either be (i) match or mismatch of an amino acid or nucleotide base, (ii) insertion of an amino acid or nucleotide base, (iii) deletion of an amino acid or nucleotide base.

Combined with a metric that scores the steps in an alignment, e.g. by putting a penalty on insertions and deletions and rewarding preservation, the best or "optimal" alignment may be deduced. In the simplest implementation all possible alignments are generated and scored to find the optimal one; unfortunately the number of alignments grows exponentially with the length of the sequences, tripling the number of alignments for each nucleotide base or amino acid added to the process.

Thus a software implementing algorithms for sequence similarity has to filter out most alignments and reduce them to the best one, depending on how the quality of alignments is estimated. This may involve the use of simple scores for the different steps up to highly sophisticated models that include mutation events, recombination and other events that change sequences. In 1970 Needleman and Wunsch (1970) described a dynamic programming algorithm whose processing time grows in relation to the product of the length of the two sequences, thus making the alignment of two large sequences feasible. These "global" alignments work well if the sequences are well conserved. In 1981 Smith and Waterman presented a similar approach for "local" alignments that works well if parts of the sequences are not conserved (Smith and Waterman 1981).

With the amount of published sequences growing, efforts were made to collect them in central repositories and make them available to the community. While the first release of "GenBank" in 1982 contained about 600 entries, the repository size has grown in an exponential manner, and in 2008 it now contains 80,000,000 DNA sequences. The first widely used program to search the repositories for similar sequences was "FASTA", released by Pearson and Lipman (1988). This program uses special optimizations of the Needleman-Wunsch algorithm that made working

with large sequence repositories feasible. In 1990, Altschul et al. published the first release of "BLAST", the basic local alignment search tool. BLAST implements a fast algorithm to find similar sequences in repositories based on a local alignment (Altschul et al. 1990).

Although other methods for searching for similar sequences in large repositories exist, FASTA and BLAST are still the most well known examples for sequence similarity based tools, and almost every large genome sequencing and annotation effort is based on them.

9.3.3.2 From Gene Annotation to Genome Annotation

Using the various bioinformatics tools available, the analysis of a novel gene and the prediction of its function based on similarity are quite easy. Many institutes provide services for analysing single genes with BLAST or other tools, with easy to use web interfaces and nice graphical post processing of results.

Things change dramatically if a gene set of a complete genome or a collection of ESTs has to be analysed. Running all the necessary prediction tools by hand is not feasible if several thousand sequences need to be processed. Data management and automation are the keys to successfully annotating a complete genome or a large set of EST sequences. Several integrated systems exist that aid biologists to carry out the annotation process, manage data, handle the various tools and present the results in a well-arranged manner. Based on the stored information most of the systems also offer higher-level functionality like metabolic reconstruction, integration of experimental data and comparisons with other genomes.

An easy to use example for a genome annotation system is "Artemis", developed at the Sanger Centre and published in 2000 (Rutherford et al. 2000). It offers visualisation and annotation of local sequence data, provided in simple sequence files, tabular files or structured files that already contain information about genes and their annotation. Installation is simply a matter of downloading and unpacking a file. It also allows the integration of various function prediction tools and adaptation to the local computer systems. Due to the focus on local files Artemis does not allow multiple users to work on the same genome at the same time.

The multiple user problem is addressed by annotation systems like GenDB (Meyer et al. 2003) and SAMS (see also Section 9.3.1.b). Sequences and information about genes, tool results and annotations are stored in a central relational database. Access is provided by a web interface, making a locally installed software unnecessary. Support from large computer clusters and pipeline processing make GenDB and SAMS powerful tools, and a well defined and open programming interface allows developers to add functionality with ease. Both systems support complete automated annotation of sequence data and may also guide manual annotation by a group of users.

Most large sequencing and bioinformatics centres have developed their own processing pipelines and are offering other institutes access to their systems. Much of this development has been carried out in the context of large eukaryotic genome projects. The Ensembl system is a good example of this kind of system (Flicek

et al. 2008). It acts as a central repository for annotated eukaryotic genomes, ranging from *Saccharomyces cerevisiae* to human. The web interface allows the genome to be browsed on several different levels, starting at the chromosomes and ending at single bases. It also includes a sophisticated data mining mechanism.

The systems described here are only examples. Many more systems exist, with new and improved systems being published every year. Almost every large sequencing project also offers its own web site with background information and specialized databases.

9.3.3.3 Protein Annotation Tools

A protein function is often associated with the presence of a particular signature or motif. Understanding the function of a protein is also often correlated with the determination of its subcellular location. Consequently it is better to use multiple methods and information resources for protein sequence annotation than a single sequence similarity search tool such as those presented in the previous section.

In the first part, we will focus on InterPro, an integrated database for protein function and structure analysis and InterProScan, a search tool, which combines the protein signature recognition methods from the InterPro database and allows the user to query an unknown sequence against the database in one step in order to aide prediction of protein function.

In the second part, we will present two reliable prediction programs used for the determination of the subcellular location: TMHMM for the detection of membrane-spanning segments and topology, and SignalP for the detection of signal peptides.

(a) InterPro and InterProScan

InterPro (Mulder et al. 2007) stands for Integrated Resource of Protein Families, Domains and Functional Sites. Proteins or protein domains belonging to a particular family usually share conserved regions often in correlation with evolution. Some are directly important for function and some play a role in the preservation of the protein 3D structure. The identification of such areas of sequence similarity allows the determination of a unique signature for a particular protein family or domain. The different proteins belonging to a protein family can then be distinguished from other proteins from another family by their signature.

The InterPro consortium is currently composed of eleven member databases located in Europe and the USA: UniProt (see also Section 9.3.5.2), PROSITE, Pfam, PRINTS, ProDom, SMART, TIGRFAMs, PIRSF, SUPERFAMILY, Gene3D, and PANTHER. The member databases have been constructed with different internal representations such as profiles or position specific scoring schemes (Table 9.4). Apart from protein signatures from the ProDom database, which are automatically generated from the UniProt sequence database, the protein signatures from the other member databases are manually curated. Refer to the member database homepages for further details on the methodology and criteria they utilize.

By combining several protein signature databases with different strengths and weaknesses, InterPro provides an integrated tool for protein function and

Table 9.4 Two categories of methods used by the member databases for building the protein signatures

Sequence-motif methods			Sequence cluster method
Regular expression and profiles	Motifs	Hidden Markov models or HMMs	Sequence clustering derived from the UniProtKB database
PROSITE	PRINTS	Pfam, SMART, TIGRFAMs, PIRSF, SUPERFAMILY, PANTHER, Gene3D	ProDom

structure analysis as well as for sequence annotation. Each InterPro entry is manually curated and composed of one or more protein signatures from one or several member database. For example, two signatures predicting the same domain of a protein will be assigned to the same InterPro entry. There are different ways of integration within InterPro. More information about this can be found in the user manual on the InterPro website.

InterPro release 17.0 contains 16,583 entries representing domains, families, post-transcriptional modifications or PTMs, repeats, and active, binding or conserved sites.

An example of an InterPro entry is shown at this link: http://www.ebi.ac.uk/interpro/DisplayIproEntry?ac=IPR000719

An InterPro entry is divided into eight fields:

- Protein matches (UniProt matches, accession number, type, signatures)
- InterPro relationships (parent/child, contains/found in)
- InterPro annotation (abstract, structural and database links)
- Taxonomic coverage
- Overlapping InterPro entries
- Example proteins (graphical view)
- Publications
- Additional reading

The protein matches or hits can be represented as a table, a simple graphical overview, a detailed graphical view (with specific colours for each member database), or as an InterPro domain architecture view (a very useful display for visualising the organization of multi-domain proteins). The structural information is available at the bottom of the view representing the mapping of SCOP and CATH structural domains to UniProt protein sequences. This information is based on the InterPro, UniProt, and Macromolecular Structural Database or MSD (http://www.ebi.ac.uk/msd/) collaboration. The InterPro graphical interface

shows the location of structural domains on a sequence via the residue-by-residue mapping between the PDB (Protein Data Bank, see also Section 9.3.5.4) chain(s) and the UniProt sequences obtained using data from the MSD. Only the PDB chains representing non-overlapping regions are shown. Although protein structure is more difficult to determine than sequence, representative structures are currently known for about 2,000 protein families. Structures are more conserved than sequences, and often reveal evolutionary relationships, which are hidden at the sequence level. Structural data is also essential for providing detailed insights into a protein's function, catalytic mechanism and interactions with other proteins.

How can these annotations of known proteins (within the InterPro database) be used to investigate the functions of novel protein sequences? That is where InterProScan comes into use. InterProScan (Quevillon et al. 2005) is a sequence search tool, which searches against the signature databases (as opposed to BLAST or FASTA which look for similarity to individual sequences in the databases; see Section 9.3.3.1). The input protein sequence format should be either free text/raw, in FASTA format or in the UniProt format. Input nucleotide sequence format used is either free text/raw, FASTA, or DDBJ/EMBL/GenBank format. You can use InterProScan via the European Bioinformatics Institute (EBI) web interface (http://www.ebi.ac.uk/InterProScan/) or the standalone version of InterProScan (ftp://ftp.ebi.ac.uk/pub/databases/interpro/iprscan/) by downloading and installing it locally on your computer.

Two tutorials from the 2can website (the bioinformatics educational resource from the EBI) provide help with using InterProScan as a tool to characterize and annotate sequences: http://www.ebi.ac.uk/2can/tutorials/

(b) TMHMM and SignalP

TMHMM is dedicated to the identification of transmembrane proteins, whereas SignalP detects signal peptides. Transmembrane^{TM} proteins are polypeptide chain(s) that pass through the lipid bilayer. They are involved in a wide variety of cellular functions such as transport and inter and intra-cellular communication. The majority of proteomes are predicted containing between 20 and 25% of transmembrane proteins. The two major classes are: alpha-helix bundle proteins, which are found in any type of membrane, and beta-barrel proteins, which are present mostly in the outer membranes (of gram-negative bacteria, but also mitochondria or chloroplasts). The identification of TM domains is not easy as few 3D structures are available. There are currently less than 1,000 TM protein 3D structures among the almost 52,000 3D structures available in the Protein data bank (PDB) as of July 2008. Hydrophobic regions of precursor proteins are also often mistaken for TM regions and aliphatic helices are not always recognized as a TM region. The most reliable prediction program used for the detection of membrane-spanning segments and topology is TMHMM 2.0 (http://www.cbs.dtu.dk/services/TMHMM/), which is based on a hidden Markov model. It predicts the location and orientation of alpha helical regions (Sonnhammer et al. 1998, Krogh et al. 2001). Unfortunately, transmembrane topology prediction programs sometimes predict signal peptides as transmembrane segments near the N-terminus. A signal peptide

contains a hydrophobic core and can therefore be misclassified as the putative first TM segment.

A signal peptide is a short peptide chain that directs the transport of a protein out of the cytosol of the cell. Different types of signal peptides exist, depending on the destination in the cell. SignalP 3.0 (http://www.cbs.dtu.dk/services/SignalP/) is a tool that can predict the presence of a signal peptide and thus provide information about subcellular location. SignalP can predict the cleavage site within the signal peptide in both in eukaryotic and prokaryotic sequences (Bendtsen et al. 2004, Emanuelsson et al. 2007). Unfortunately, signal peptide prediction programs sometimes identify N-terminal TM segments as signal peptides. The Phobius web server (http://phobius.binf.ku.dk), which combines transmembrane topology and signal peptide prediction, was developed to address this problem (Kall et al. 2007).

The Table 9.5 summarizes the prediction programs presented in this section. Other prediction programs have been developed by G. Von Heijne's group such as ChloroP (Emanuelsson et al. 1999), which identifies the presence and location of chloroplast transit peptides, and TargetP (Emanuelsson et al. 2007). TargetP predicts the subcellular location of chloroplast transit peptides, mitochondrial targeting peptides, or secretory pathway signal peptides.

Table 9.5 Programs that provide a reliable prediction of subcellular location

Prediction program	TMHMM 2.0 (Phobius server)	SignalP 3.0 (Phobius server)	ChloroP	TargetP
Subcellular location	Transmembrane helices	Signal peptide cleavage sites	Chloroplast transit peptide	Any N-terminal presequence
Organism	Prokaryotes and eukaryotes	Prokaryotes and eukaryotes	Plants	Eukaryotes

9.3.4 Comparative Genomics and Functional Classification

With the exponential growth in the number of completely sequenced prokaryotic and eukaryotic genomes available, there is clearly a need for accurate, consistent and automated annotations of gene functions.

9.3.4.1 Homology and Similarity

Chothia et al. proposed that most proteins have been formed by gene duplication, recombination, and divergence (Chothia et al. 2003). In order to describe the evolutionary relation of proteins, the terms homology, orthology, and paralogy are used. In this context, homology means that two proteins or sequences share a common ancestor. Homology among protein or DNA sequences can only be concluded on the basis of sequence similarity, because the true evidence for homology would require the analysis of the common ancestor and all intermediate forms (Reeck et al. 1987). If two genes have an almost identical DNA sequence, it is likely that they

are homologous. However, homology is not in all cases the reason for sequence similarity. Short sequences may be similar by chance or sequences may be similar because both were selected to bind to a particular protein, such as a transcription factor. Two principal types of homology can be distinguished. Orthologous sequences are homologous sequences that were separated by a speciation event. A gene that existed in an ancestral species, which then diverged into two species, will then exist as two copies. These copies are called orthologues and will typically have the same or a similar function.

Homologous sequences that have been separated by a gene duplication event in an ancestral genome are called paralogous. Paralogous genes may mutate and acquire new functions because the selective pressure is reduced. The three terms and their relation are depicted in Fig. 9.4.

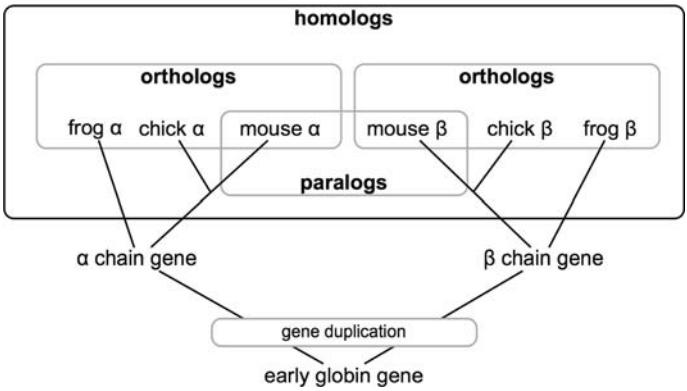

Fig. 9.4 Relation of the terms homology, orthology and paralogy. Genes that have a common ancestor are homologous, orthologous genes are separated by a speciation event. A gene duplication event generates two paralogous copies of a single gene. This figure is based on the figure at the following website: http://www.ncbi.nlm.nih.gov/Education/BLASTinfo/Orthology.html

Functional annotation of genomes can be described as the process of identifying the functions of particular regions of sequence data that would otherwise be almost devoid of information (Overbeek et al. 2005). By assigning functions to coding regions, the understanding of a complete genome is extended. The most direct and reliable method to obtain the function of a coding region is to carry out biological experiments. This approach is time consuming and expensive and therefore not practical for the vast amount of genomic data that has been generated.

Proteins that are encoded by genes with similar sequences often have a similar three-dimensional structure. Since the three dimensional structure determines the function of a protein, the assumption can be made that genes with a similar sequence encode proteins with a similar function. The connection between the sequence of a gene and the function of the corresponding gene product can be exploited to obtain a functional annotation for an unknown gene. For this purpose, its nucleotide or amino acid sequence is compared to the sequences of genes with known and verified function. If a relevant sequence similarity can be found, the function of the known gene

can be assigned to the unknown gene. A variety of computational tools for detecting sequence similarity are available with the Basic Local Alignment Search Tool (BLAST) (Altschul et al. 1990) being one of the best known and most widespread.

9.3.4.2 Protein Domains

Protein sequences can also be analysed in detail for the presence of conserved domains. Domains can be seen as the functional building blocks of proteins. Specialized databases such as that constructed by the InterPro Consortium have been established for this task (see Section 9.3.3.3).

Sequence analyses and similarity searches against databases produce evidence that support the functional annotation of a gene. If the detected sequence similarity is low or the evidence is contradictory, an assignment of the function of a gene is difficult and requires human interaction. Otherwise, a propagation of false annotations may be introduced in the genomic databases. Functional annotation is in general thought to be of best quality when performed by a human expert.

9.3.4.3 Use of Gene Clusters in Functional Annotation

As described before, sequence similarity can be used to determine the function of genes. The results derived from sequence analyses can be enriched by incorporating the evidence of higher-level analyses of the genomic data. The order of genes on the chromosome of an organism often yields additional information about functionally related genes, particularly for prokaryotes. In bacteria the sequence of genes on the chromosome (gene order) is well preserved at close phylogenetic distances (Tamames et al. 1997), but the order and composition of genes in two diverging organisms changes over time. Events like duplication or loss of genes as well as horizontal gene transfer change the composition of genes in the genome. Translocation, transposition, inversion, and chromosome fission and fusion affect the gene order.

Interestingly, sets of genes with strong conservation of composition and order can also be detected in distantly related species. The term "gene cluster" was used for the first time by Bauerle and Margolin (1966) who described the tryptophan gene cluster in *Salmonella typhimurium*. Through the detection of further occurrences of the tryptophan cluster in other organisms, Tatsuov et al. (1996) showed in the mid-90s that, although gene order is generally under no selective pressure in prokaryotic genomes, certain genes tend to conserve their chromosomal neighbourhood. This circumstance can be explained by evolutionary advantages that derive from the chromosomal neighbourhood of the respective genes for the organisms. Functionally related genes occur in close proximity for example, if they are collectively regulated in an operon in prokaryotic genomes. If their gene products interact, it is advantageous for the cell to produce them at a close distance (Dandekar et al. 1998).

Another reason for conserved gene order across phylogenetic distances is the occurrence of horizontal gene transfer (Lawrence and Roth 1996). It is reasonable to use the information present in the chromosomal context of the gene as presented

in Overbeek et al. (1999) to infer functional coupling of genes and thereby generate additional evidence for functional annotation. Especially, if a gene with unknown function is found in a conserved cluster among several genomes, this contextual information can support a potential gene function (Tamames et al. 1997).

9.3.4.4 Existing Resources for Comparative Analyses

A variety of annotation systems have been developed to support the process of genome annotation (see Section 9.3.3.2). With the availability of more than 1,000 completely sequenced genomes the comparative genome annotation strategy has become of major interest. It incorporates existing genomic data from related or all available genomes for the annotation of a novel genome. Here we present some annotation systems that support this strategy as well as relevant ontologies and databases.

MAGPIE: MAGPIE is an annotation system developed by Gaasterland et al. (2000). It provides a graphic interface that annotators can use to navigate a genome and it assists in automated data collection, analysis, and annotation. The system uses the internal HERON tool to generate automated annotations based on evidence from homology searches. The decision process of a human annotator is modelled in a protocol that tries to select the best annotation from the description lines of high scoring matching sequences.

ERGO: The ERGO genome analysis and discovery suite integrates data from genomics, biochemistry, high-throughput expression profiling, genetics, and peer-reviewed journals (Overbeek et al. 2003). 500 genomes at various levels of completion have been integrated into the system. The functional assignment of genes is supported by evidence derived from comparative analysis including co-regulation, fusion events, chromosomal neighbourhood of functionally related genes. Reconstructions of cellular pathways are contained in the database for genomes from all three domains of life. The system is no longer publicly available and is only accessible through a fee-based subscription.

The SEED: The SEED (Overbeek et al. 2004) system provides the functionality to annotate the exponentially growing number of completely sequenced genomes. It is an open source successor to the commercial WIT (Overbeek et al. 2000) and ERGO annotation systems. The SEED system replaces the one-genome-at-a-time annotation strategy with an approach involving simultaneous annotation of all available genomes. The annotation process is supported by comparative analyses.

The sequences of organisms that are available to the public are present in the system's database and an all-against-all similarity matrix for the genomic features is pre-computed and allows the detection and visualisation of chromosomal clusters and functional coupling of genes. Furthermore, the pre-computed similarity information present in the database enables the annotator to find sets of orthologous genes and annotate them consistently across the complete set of organisms. The system also provides functionality to organize related genes on a higher level. The SEED system was the first to introduce the concept of subsystems (Overbeek et al. 2005).

KEGG and KOBAS: The Kyoto Encyclopedia of Genes and Genomes (Kanehisa and Goto 2000) combines databases that represent molecular interaction networks in the context of biochemical pathways. These feature enzymes, compounds, and connection points to related pathways. The recently published KOBAS (Mao et al. 2005) system uses the KEGG Ontology (KO) to produce automated annotations of large sets of genes. KO is a controlled vocabulary similar to the Gene Ontology (Ashburner et al. 2000) that offers the connection of genes to metabolic pathways present in the KEGG database. KOBAS is an automated annotation system written in Python, which is able to identify and annotate complete pathways in sets of sequences. KO terms are assigned to query genes if they show a high sequence similarity to genes that are already annotated and connected to KEGG maps.

STRING: The STRING (von Mering et al. 2005) database stores information about genomic associations between genes. The most relevant associations between two genes are conserved chromosomal proximity of genes in phylogenetically distant organisms and gene fusion events. In addition, information about similar expression patterns in microarray experiments and the co-occurrence of gene names in the literature are included in the database. The combined information can be used to predict unknown protein–protein interactions. The database is a pre-computed global resource for the exploration of functional interactions between genes and is available online.

COG: The COG database (Clusters of Orthologous Groups of proteins) (Tatusov et al. 2003) was created by Tatusov and co-workers in 1997. It was designed to classify genes from completely sequenced organisms based on their common origin. Initially a set of 21 complete genomes was used and an all-against-all sequence comparison was performed. Clusters of Orthologous Groups were created by applying the criterion of genome specific best hits to a comparison of all coding sequences. The COG categories were derived from the clusters. They act as a controlled hierarchical vocabulary that can be used to describe the function of proteins. For the initial creation of the database, 2,091 COGs could be created that included up to 83% of the gene products of a single organism. The clusters have continuously been extended with the genes of novel sequenced organisms.

GO: The Gene Ontology (GO) project was started as a collaborative effort to address the need for consistent descriptions of gene products in different databases (Ashburner et al. 2000). Driven by the fact that functional conservation of genes can be found across all three domains of life, a common language for the annotation of gene products has been established. Thereby, the interoperability of genomic databases is simplified. Three structured controlled vocabularies (ontologies) have been defined that describe gene products in terms of their associated *biological processes*, *cellular components,* and *molecular functions* in a species-independent manner. The Gene Ontology can be found online at http://www.geneontology.org, existing terms and their relations can be browsed via the AmiGO ontology browser.

9.3.5 Major Public Sequences Databases and Other Resources

This section presents the major public sequence databases and other resources that aim to provide a comprehensive coverage of sequences and annotations available to the scientific community.

In the first part, we will initially introduce the three leading nucleotide sequence centres and explain how to access the collected data. Following this, the different procedures of submission required by these centres, depending on the type of data being submitted, are described. The second part focuses on UniProt, which is a database for protein sequences. Finally, an introduction is provided to some other resources that may be useful to marine biologists.

9.3.5.1 Major Public Nucleotide Sequences Databases

The three major public database centres are located in Europe, Japan, and the USA (see Table 9.6). These databases form the International Nucleotide Sequence Database Collaboration (INSDC, www.insdc.org) and have a long-established collaboration of over 18 years involving the daily exchange of new and updated nucleotide sequence and annotation data (Brunak et al. 2002). Collectively, the databases aim to provide a comprehensive coverage of sequences and annotations available within the public domain.

Table 9.6 The major public nucleotide sequences databases and their database centres

Major public database centres		
Europe	Japan	USA
EBI (European Bioinformatics Institute)	CIB (Center for Information Biology) at the NIG (National Institute of Genetics)	NCBI (National Center for Biotechnology Information)
Established in 1993	Established in 1995	Established in 1988
www.ebi.ac.uk	www.ddbj.nig.ac.jp	www.ncbi.nlm.nih.gov
EMBL-Bank created in 1981 (Cochrane et al. 2008)	DDBJ created in 1986 (Sugawara et al. 2008)	GenBank created in 1982 (Benson et al. 2008)

Three activities are central to the ongoing INSDC effort. Firstly, the databases provide submission services for data generators, so that information can be submitted as easily as possible, while retaining important contextual biological information (such as the biological source of the sequence) and functional interpretations of the sequence data in the form of an annotation. Secondly, the databases develop structures and formats through which the sequence and annotation data can be accurately and concisely represented. The focus is on the usability for the users; instruments include the INSDC Feature Table Definition document and associated vocabularies,

such as those in the /country and /db_xref qualifiers. Finally, the databases strive to promote public availability of sequence and annotation data through a collaboration with publishers.

Each centre strives to impose a quality control upon submitted sequences and annotation, but the data generator retains editorial responsibility for the biological content of his/her entries. Increasingly, in the face of ever growing data volumes, these quality control measures rely on automated validation procedures. Unique and permanent database entry accession numbers (e.g., AF123456) are provided by the receiving database upon submission to allow future identification of the entry. Database accession numbers are required prior to publication by the publishers of the major molecular biology journals.

Accession number prefixes depend on the issuing database and the type of data submitted (e.g. direct submission from DDBJ, genome project data from EMBL, EST from GenBank, patent from JPO, EPO, or USPTO). A complete list of prefix codes used so far can be found here: http://www.ddbj.nig.ac.jp/sub/prefix.html.

Note that the GI prefix identifiers (e.g. GI:26117688) are internal to Genbank and are not primary INSDC accession numbers. To resolve a GenInfo identifier or GI, the Entrez website from NCBI can be used to retrieve the primary INSDC accession number (in this case it is U00089).

(a) DDBJ at CIB

The DNA Data Bank of Japan (DDBJ) (Sugawara et al. 2008) was created in 1986 at the National Institute of Genetics (NIG) which was then reorganized as the Center for Information Biology and DNA Data Bank of Japan (CIB-DDBJ) in 2001. The centre mainly collects submissions from Japanese laboratories.

A sample entry in DDBJ flat file format can be found in the regularly updated DDBJ/EMBL/GenBank Feature Table file on the DDBJ website.

Several useful DDBJ annotation examples (e.g. ribosomal RNA, EST, microsatellite, transposon) can be found on their website.

The DDBJ nucleotide sequence database can be accessed in different ways (see Table 9.7).

Table 9.7 The different ways to retrieve data at DDBJ

Sequence retrieval at DDBJ	
Getentry	Data retrieval of nucleotide sequences mainly by accession numbers
ARSA	All-round Retrieval of Sequence and Annotation. Search of sequence libraries such as DDBJ and UniProt, sequence-related libraries such as PROSITE and Pfam, protein 3D structures, and metabolic pathways
SRS	Sequence Retrieval System offering term search
GIB	Genome information broker or data retrieval and comparative analysis system for completed genomes

(b) EMBL at the EBI

EMBL-Bank (Cochrane et al. 2008) was created in 1981 at the European Molecular Biology Laboratory (EMBL) in Heidelberg, Germany. It moved in 1993 to the European Bioinformatics Institute (EBI) outstation in Hinxton near Cambridge, UK, to be closer to genome research institutes such as the Wellcome Trust Sanger Institute. EMBL-Bank now forms part of the Protein and Nucleotide Database Group (PANDA) at the EBI.

A sample entry in EMBL flat file format can be found in the regularly updated DDBJ/EMBL/GenBank Feature Table file on the EMBL website.

Several other useful EMBL annotation examples can be found on the EMBL website.

The EMBL nucleotide sequence database can be accessed in a number of ways (see Table 9.8).

Table 9.8 The different ways to retrieve data at EMBL

Sequence retrieval at EMBL	
Simple sequence retrieval (embl fetch)	Sequence retrieval by accession number
SRS	Query all databases by term search, including EMBL-Bank standard, EST, STS and GSS data
EMBL sequence version archive	Repository of all current and historical EMBL entries
Browse data by geography	Geographical origin of sequenced samples
FTP server	Complete latest EMBL release, completed genomes, contigs, WGS sequences, patent sequences, etc.
Genomes	Access to completed genomes
Genome reviews	Genome annotation of Archaea, Bacteria, bacteriophages and selected Eukaryota
Ensembl genome browser	Annotation of large eukaryotic genomes
Integr8	Proteome analysis information

(c) GenBank at NCBI

The GenBank nucleotide sequence database (Benson et al. 2008) was created in 1982 at the Los Alamos National Laboratory (LANL) in the USA. It moved in 1993 to the National Institute of Biotechnology Information (NCBI) in Bethesda, Maryland.

A sample entry in GenBank flat file format can be found in the regularly updated DDBJ/EMBL/GenBank Feature Table file on the NCBI website.

Several useful annotation examples in GenBank flat-file format can be found on the BankIt website.

The GenBank nucleotide sequence database can be accessed in a number of ways (see Table 9.9).

Table 9.9 The different ways to retrieve data at GenBank

Sequence retrieval at GenBank	
Entrez nucleotide browser	Searches for sequences in Genbank, RefSeq and PDB
dbEST searching	Expressed sequence tags from one of the 6 major organism groups (Archea, Bacteria, Eukaryota, Viruses, Viroids, and Plasmids)
dbSTS searching	Sequence tagged sites
dbGSS searching	Genome survey sequences
FTP	Full release and daily updates of GenBank
Genomes	Views provided for genomes from one of the 6 major organism groups (Archea, Bacteria, Eukaryota, Viruses, Viroids, and Plasmids)

Table 9.10 Submission tools available at the major public database centres

Direct submissions of DNA sequences			
	DDBJ (CIB)	EMBL (EBI)	GenBank (NCBI)
Submission tool	Sakura[1]	Webin[2]	BankIt[3] or Sequin[4]

[1] http://sakura.ddbj.nig.ac.jp/top-e.html
[2] http://www.ebi.ac.uk/embl/Submission/webin.html
[3] http://www.ncbi.nlm.nih.gov/BankIt/index.html
[4] http://www.ncbi.nlm.nih.gov/Sequin/index.html

(d) Nucleotide Sequence Submissions

How to submit data to one of the nucleotide sequence databases?

If you have one or more sequences to submit (as is often the case for small laboratories), the direct submission procedures from one of the three sites are recommended (Table 9.10). For long or complex submissions, the database in question should be contacted in advance for assistance.

If you have many sequences to submit (often the case in large-scale sequencing centres), each of the three sites offers a bulk submission procedure such as the mass submission system (MSS) at DDBJ.

For submissions to NCBI of Expressed Sequence Tags (ESTs), Genome Survey Sequence (GSS), and Sequence Tagged Sites (STS) data, please have a look at the NCBI website.

For all submissions, it is recommended that the submitter studies the database website for up to date information on submission at one of the three sites (DDBJ, EMBL, or GenBank). The databases can be contacted by e-mail at the following addresses:

DDBJ: ddbjsub@ddbj.nig.ac.jp
EMBL: datasubs@ebi.ac.uk
GenBank: gb-sub@ncbi.nlm.nih.gov

(e) Third Party Annotation

The INSDC Third Party Annotation (TPA) section (Cochrane et al. 2006) was created in 2002 in order to accept high-quality annotations of nucleotide sequences from submitters who have not themselves generated those nucleotide sequences. The quality of the annotation is supported by experimental or inferential analyses.

TPA data are divided into two tiers:

- with experimental evidence (e.g., BK000016)
- with inferential evidence – where the source molecule or its product(s) have not been the subject of direct experimentation (e.g., BK000554).

More information about TPA data can be found at the DDBJ, EMBL, or GenBank websites.

9.3.5.2 Major Public Protein Sequences Database: UniProt

Dr. Margaret Oakley Dayhoff (1925–1983) was a pioneer in the bioinformatics field of computers in chemistry and biology, and produced the Atlas of Protein Sequence and Structure, published from 1965 to 1978 by the National Biomedical Research Foundation (NBRF). The Protein Information Resource (PIR) was then established in 1984 by the NBRF as a first resource of protein sequence information. It produced the protein sequence database called PIR-PSD until the end of 2004.

Swiss-Prot, an annotated sequence database, was created in 1986 by Amos Bairoch in Geneva, Switzerland and the first entry was human cytochrome c (http://www.expasy.ch/uniprot/P99999). The Swiss Institute of Bioinformatics (SIB) has hosted the Swiss-Prot group since 1998 and also currently maintains the Expert Protein Analysis System (ExPASy) proteomics server (http://www.expasy.org). This server is dedicated for more than 15 years to the analysis of protein sequences and structures as well as two-dimensional gel electrophoresis (2D Page electrophoresis).

To cope with the growing amount of sequence data, TrEMBL was created at the EBI (European Bioinformatics Institute) in 1996. It is a computer-annotated protein sequence database containing translations of all coding sequences (CDS) present in the DDBJ/EMBL/GenBank Nucleotide Sequence Databases, which are not yet in Swiss-Prot. SIB and EBI combined efforts from the beginning to jointly produce Swiss-Prot and TrEMBL.

In 2002, the three institutes (EBI, PIR, and SIB) pooled their resources and expertise. The Universal Protein Resource (UniProt) Consortium (Consortium 2008) was born to provide a single database of high-quality and comprehensive protein sequence and functional information (http://www.uniprot.org).

(a) Organisation of the UniProt Databases (Table 9.11)

The UniProt Knowledgebase (UniProtKB) is composed of two databases: UniProtKB/Swiss-Prot, a high-quality manually annotated and non-redundant

Table 9.11 Organisation of the UniProt databases

The universal protein resource			
UniProtKB: protein knowledgebase	UniRef: sequence clusters	UniMES	UniParc
UniProtKB/Swiss-Prot (manually curated annotation) UniProtKB/TrEMBL (automatic annotation)	UniRef100 UniRef90 (at least 90% sequence identity) UniRef50 (at least 50% sequence identity)	Metagenomic and environmental sample sequences	UniProt archive

protein sequence database, which brings together experimental results and computed features and UniProtKB/TrEMBL, a high quality automatically annotated database. TrEMBL entries are manually annotated and integrated into Swiss-Prot, keeping their unique accession number.
UniProtKB contains all the protein sequences available except for the following ones:

- Most non-germline immunoglobulins and T-cell receptors
- Synthetic sequences
- Most patent application sequences
- Small fragments encoded from nucleotide sequence (<8 amino acids)
- Pseudogenes
- Fusion/truncated proteins
- Not a real protein (when enough evidence that the existence of a protein looks dubious)

The first five types of sequences are identified automatically during the creation of UniProtKB/TrEMBL. The last two types are manually identified by curators (e.g., sequences derived from gene predictions from genomic sequences, which were wrongly predicted to code for proteins) and then removed.

All these seven types of excluded sequences are available in UniParc with a reason for their exclusion from UniProtKB (see paragraph below).

More information about UniProtKB can be found on the UniProt website.

The UniProt Reference Clusters (UniRef) provide clustered sets of closely related sequences from the UniProt Knowledgebase to allow fast searches. UniRef90 and UniRef50 are composed of sequences that have at least 90% or 50% sequence identity, respectively. More information about UniRef can be found on the UniProt website.

The UniProt Metagenomic and Environmental Sequences database (UniMES) currently contains data from the Global Ocean Sampling Expedition (GOS), which were originally submitted to the International Nucleotide Sequence Databases

(INSDC). UniMES is available on the FTP site in FASTA format with a "UniMES matches to InterPro methods" file.

The UniProt Archive (UniParc) is a non-redundant database that aims to capture all publicly available protein sequences. UniParc stores each unique sequence only once and gives it a stable and unique identifier (UPI), making it possible to identify the same protein from different source databases. A UPI is never removed, changed, or reassigned. The basic information stored within each UniParc entry is the identifier, the sequence, cyclic redundancy check number, source database(s) with accession and version numbers, and a time stamp. If a UniParc entry does not have a cross-reference to a UniProtKB entry, the reason for the exclusion of that sequence from UniProtKB is provided (e.g. patent).

More information about UniParc can be found on the UniProt website.

(b) Release, Submission, and Access

UniProt is updated every 3 weeks and a major release is also produced three times per year. The first major release was in December 2003. The latest release (release 13, February 26, 2008) includes 5,751,608 entries as follows:

- 356,194 UniProtKB/Swiss-Prot entries (release 55) and 11,290 different species,
- 5,395,414 UniProtKB/TrEMBL entries (release 38) and 1,552,882 different species.

If you have a new protein sequence you can submit it directly to UniProtKB using the web-based tool SPIN at the EBI.

The UniProt website can be accessed at http://www.uniprot.org where examples with the protein accession number, the UniRef entries, and UniParc unique identifier can be found.

9.3.5.3 RefSeq

The Reference Sequence (RefSeq) database is a non-redundant collection of transcripts, proteins, and genomic regions (http://www.ncbi.nlm.nih.gov/RefSeq/) produced by NCBI (Pruitt et al. 2007). RefSeq is limited to major organisms for which sufficient data is available (almost 5,000 distinct organisms as of January 2008, release 27), while GenBank includes sequences for any organism submitted (more than 250,000 different named organisms). Each RefSeq entry represents a single sequence from one organism. The annotation status within the entries varies and includes not annotated (inferred, model, predicted, provisional or WGS) or annotated (validated and reviewed) entries.

The Reference Sequence (RefSeq) database can be accessed in different ways, either directly by querying or indirectly through links provided from several NCBI resources including Gene, Entrez, PubMed, and Map Viewer.

RefSeq uses the following prefixes with two characters followed by an underscore character ("_") such as NP_010000. More information on the RefSeq accession number format can be found on the RefSeq website.

Table 9.12 Comparison RefSeq vs UniProt

RefSeq	UniProt	
Coding sequences from the NCBI's set of genomic, transcript and protein reference sequences. Only the entries with validated or reviewed status are annotated entries	UniProtKB/Swiss-Prot: manually annotated protein sequences mostly derived from TrEMBL	UniProtKB/TrEMBL: automatically annotated protein sequences derived from coding sequences in nucleotide sequence database
Release 27, January 11, 2008: 4,426,609 protein entries and 4,926 organisms	Release 13, February 26, 2008: 356,194 entries (release 55) and 11,290 different species	Release 13, February 26, 2008: 5,395,414 entries (release 38) and 1,552,882 different species
Bi-monthly release	Major release 3 times per year, minor release every 3 weeks	
Limited to major organisms	All organisms	
Exclusive NCBI database	Produced by EBI, PIR, and SIB	
Protein and nucleotide data	Protein data only	

In Table 9.12 you can find different characteristics of RefSeq versus UniProt.

9.3.5.4 Other Resources

In the following, a few other resources that contain more specific data are presented.

(a) Organism Specific Databases

The databases briefly presented here contain information specific to an organism or genome. The list below is not exhaustive. These databases vary greatly in the classes of data captured and how these data are stored. Please have a look for updates at the Database issue from Nucleic Acids Research published in January each year:

- FlyBase or A Database of *Drosophila* Genes and Genomes (http://www.flybase.org)
- GDB or Human Genome Database (http://www.gdb.org/)
- MGI or Mouse Genome Informatics (http://www.informatics.jax.org)
- SGD or *Saccharomyces* Genome Database (http://www.yeastgenome.org)
- TAIR or the *Arabidopsis* info resource (http://www.arabidopsis.org)
- WormBase or Databases on the genetics of *Caenorhabditis elegans* and related nematodes (http://www.wormbase.org)
- ZFIN or the zebrafish model organism database (http://www.zfin.org)

(b) Marine Databases

Over the last years, modern high-throughput techniques in genome and post-genome research have also entered the marine sciences. Today, massively parallel DNA sequencing or hybridization approaches allow identification not only of the gene repertoire but also of the gene regulatory networks of an organism. Below is a list

of recent marine genomics projects and databases that are trying to catalogue these new data. This list is not exhaustive.

- Moore Marine Microbial Genome Sequencing Project (http://www.moore.org/microgenome/). The Moore Foundation's Microbial Genome Sequencing Project was launched in April 2004. More information about the sequencing project can be found at https://research.venterinstitute.org/moore/.
- The Megx.net (http://www.megx.net) database resource for marine ecological genomics provides specialized databases and tools for genome-wide analyses of marine bacteria and metagenomics.
- The Marine Genomics Project, Charleston South Carolina (http://www.marinegenomics.org/). The Marine Genomics pipeline automates the processing, maintenance, storage, and analysis of ESTs (Expressed Sequence Tags) or 16S RNA sequences and microarray experiments from 35 different marine species.
- Marine Genomics Europe or MGE (http://www.marine-genomics-europe.org/) is devoted to the development, utilization and spreading of high-throughput approaches for the investigation of the biology of marine organisms. Within the bioinformatics platform located at Bielefeld University, a bioinformatics portal (http://www.cebitec.uni-bielefeld.de/brf/cooperations/mge.html) has been created for the Marine Genomics Europe (MGE) community that provides a central access point for all data sets and various software tools (e.g. GenDB, SAMS, EMMA).

(c) wwPDB

The Protein Data Bank (PDB) was founded in 1971 at Brookhaven National Laboratory (Long Island, USA) to archive experimentally determined three-dimensional structures of biological macromolecules. In 1974, 12 structures with atomic coordinates were available. Nowadays the PDB contains the coordinates and related information of more than 50,000 structures determined using X-ray crystallography, Nuclear Magnetic Resonance (NMR) and electron microscopy techniques (Henrick et al. 2008).

The importance of protein structures Although the protein structure is more difficult to determine than protein sequence, representative structures are currently known for about 2,000 protein families. Structures are more conserved than sequences and often reveal evolutionary relationships hidden at the sequence level. Structural data is also essential for providing detailed insights into a protein's function, catalytic mechanism and interactions with other proteins. Because of the increase in the number of deposited structures during the last 10 years and because of structural genomics projects emerging that will generate a large number of 3D structures, a worldwide protein data bank was needed to maintain a single archive of publicly available macromolecular structural data.

The worldwide Protein Data Bank (wwPDB) was thus established in 2003 (Berman et al. 2003). The founding members are RCSB PDB (USA), the macromolecular structure database (MSD) at the EBI (Europe), and the Protein Data Bank

Japan (PDBj) at Osaka University. The BioMagResBank or BMRB group (USA) joined the wwPDB in 2006. These wwPDB sites (Table 9.13) share the responsibilities in data deposition, processing, and distribution of the PDB archive, and agree to support a single, standardized archive of structural data. The archive is currently updated weekly.

Table 9.13 Public database centres for biological macromolecular 3D structures

wwwPDB data access sites			
BMRB: http://www.bmrb.wisc.edu	MSD-EBI: http://www.ebi.ac.uk/msd	PDBj: http://www.pdbj.org	RCSB-PDB: http://www.pdb.org
USA	Europe	Japan	USA

9.4 Transcriptome Analysis Using High-Throughput Technology

Modern high-throughput sequencing technology has produced large amounts of biological sequence data. For marine organisms these data sets may consist of whole microbial genome sequences or of large EST libraries of marine eukaryotes as described in the previous sections. The available structural genomics data raise the need for further analyses to determine the functions of genes and other sequences. The analysis of intrinsic sequence features and sequence comparisons applied to existing genomic data provide initial information on the function of novel sequences. Still, there is a need for quantitative experiments to infer new hypotheses about genes of unknown function and to test hypotheses of sequence function.

The central dogma of molecular biology describes protein expression as a directed flow of information: from DNA through the intermediate of messenger RNA (mRNA) towards the end product, a protein. Regulation occurs at several stages of gene expression, from the DNA, at the level of regulation of transcription, to the level of translation into amino acids, and post-translational modification. Gene-expression depends on the internal state of the organisms and environmental conditions. The control of regulatory networks is mediated by complex signalling networks within the cell. It is often assumed that quantitative measurements of gene-expression under certain experimental conditions can be used to infer the function of genes. An approach, which is followed by a large number of researchers, is to determine gene function by patterns of common regulation between genes. Often genes, which function in a similar metabolic pathway or share another common function, show similar patterns of gene-expression in transcription profiling experiments. This approach has been termed “guilt-by-association” (Quackenbush 2003).

Transcriptomics is a relatively new field in functional genomics that aims at measuring abundances of mRNA molecules. Several methods can be used to quantify mRNA-abundance, including quantitative Real-Time Reverse Transcription PCR (qRT-PCR) (Iizuka et al. 1994), Serial Analysis of Gene Expression (SAGE)

(Velculescu et al. 1995), and microarrays. Each of these methods will be discussed in the following sections.

Quantitative real-time reverse-transcription PCR Quantitative real-time RT-PCR is a highly sensitive method for the detection and quantification of low abundance mRNA. In this method an mRNA sample is transcribed into cDNA using the enzyme *reverse transcriptase*. cDNAs corresponding to a few genes of interest are amplified using gene specific primers in the standard PCR technique. The gene specific primers guarantee that only the desired gene is amplified. The amplification of sample cDNA is monitored using fluorescent dyes as reporters for the amount of DNA synthesised. There are several types of dyes available, the most simple and most used ones are double strand binding dyes like SYBR Green I. These dyes emit fluorescence when bound to dsDNA. The increase in fluorescence intensity correlates with the amount of dsDNA produced, and thereby allows an assessment of the amount of PCR product that has been synthesised. The initial amount of cDNA can be inferred from the time it takes until the fluorescence reaches an initial background level, identified as the so called crossing point (CP). The earlier the crossing point is reached, the higher the inital cDNA amount. The drawback of this technique is that it is relatively low throughput, as only the transcription level of one gene is measured per assay. However, larger-scale commercial qPCR systems are available with capacities ranging from 32 to 384 assays per qPCR run.

A detailed description of all aspects of this technique such as platforms, reagents, and data normalization would go beyond the scope of this chapter, but a comprehensive overview of qPCR related issues, including an updated list of publications regarding qPCR, is provided at http://www.gene-quantification.info. As far as the bioinformatics analysis of the data is concerned, very few open source software solutions suited for a full-fledged analysis of qPCR data are available. One example is the CAmpER system (available at the Center for Biotechnology, Bielefeld University, http://www.cebitec.uni-bielefeld.de), which is freely accessible.

Serial Analysis of Gene Expression (SAGE) and other sequencing based approaches: Serial Analysis of Gene Expression is a high-throughput method for the quantitative analysis of mRNA. It is based on sequencing short fragments of cDNA, so-called tags. In contrast to EST sequencing, the tags are only short fragments (11–25 bp) of the full-length cDNA.

In the original protocol, cDNA is generated using biotinylated oligo(dT) primers which bind the poly-A tail at the 5′ end of the mRNA. The resulting double-stranded cDNA is bound to streptavidin beads at the primer site and afterwards cleaved with an anchoring restriction enzyme (NlaIII) to obtain shorter fragments.

The beads carrying bound cDNA are then divided in two aliquots. The cDNA in the two aliquots is linked to two different linker oligos (A and B) and cleaved from the attached beads using a tagging enzyme (BsmFI). This enzyme has the property of binding at a certain nucleotide sequence (CATG) and cleaving the DNA 11 bp upstream of its binding site. The tags obtained from the two aliquots are further ligated into di-tags, which can be amplified using PCR-primers specific to A and B.

After a sufficient level of amplification is achieved, the linkers A and B are cleaved using NlaIII again. The resulting di-tags are concatenated randomly into

long chains of tags. In this classical protocol, the tags are then cloned into a plasmid vector and their sequence is obtained using Sanger sequencing. The specific cleavage site remains in the DNA-vector and can be used as a separator.

If genomic sequences of the organism are known, either in the form of ESTs or as whole genome, the tags obtained from sequencing can be mapped onto the longer sequences, thereby enabling a quantitative analysis of differential gene-expression. The number of similar tags found can be assumed to correlate (although not linearly due to the involvement of PCR-amplification) with the original amount of cDNA fragments from which these tags are uniquely derived.

If no sequence information is available for the source organism, the sequence tags can still be analysed quantitatively, as the tag sequence is an (almost unique) identifier for the corresponding mRNA. However, mapping of short tags of 11 bp to available databases for sequence comparison will yield a large proportion of false positive hits.

An initial problem of the SAGE protocol was the high amount of source poly-A RNA required (approximately 2.5–5 μg). Therefore, Datson et al. (1999) developed MicroSAGE which uses strepatividin coated tubes instead of beads. The method requires only 1 ng of RNA, which is approximately the equivalent of 100,000 mammalian cells

Another drawback of the initial SAGE approach was the short tag-length. Several enhancements of the technique have been published which provide an improved tag-length by using different restriction enzymes. Tag lengths of 20 bp (LongSAGE) (Saha et al. 2002), and 26 bp (SuperSAGE) (Matsumura et al. 2003), have been reported. These provide increasingly significant database hits.

Another improvement has been the development of gene identification methods based on paired-end di-tags (GIS-PETs) (Ng et al. 2005). In contrast to SAGE, PETs are derived from the 3′ and 5′ signatures of mRNA. For the 3′ end, primers specific to the capping structure of the processed mRNA are added. As a result, paired di-tags contain sequences from both termini of the same mRNA. Thus, the exact transcription boundaries can be identified.

In combination with second generation sequencing methods, a vast increase in throughput can be obtained (MS-PET) (Ng et al. 2006). The PETs have an average length of 40 bp per di-tag and therefore two di-tags correspond to one sequencing read of a first generation 454 sequencer.

It should be mentioned however, that SAGE and all derived methods involve highly complex laboratory protocols. The increased throughput of next-generation sequencing methods has shifted the bottleneck of sequencing to the adaptation of these protocols. Also, it must be noted that the protocols are not all applicable to prokaryotes, as the they are based on mRNA structures (5′ capping, poly-A tail) which are found in eukaryotes but not prokaryotes.

Shotgun-transcriptomics approaches are also suitable for the analysis of trancription. In principle, the approach is similar to the EST approach described before. Messenger RNA is reverse transcribed into cDNA which is directly fragmented and sequenced. Instead of using Sanger sequencing, high-throughput methods can

be applied. The resulting sequence reads can be mapped onto the genome and be analysed quantitatively. This method also applies to prokaryotes, and allows for a high coverage of complete transcripts. Despite its high potential, it has currently only been used for a limited number of studies and clear guidelines for the statistical analysis of shotgun-transcriptomics data still need to be developed.

9.4.1 Fundamentals of Microarray Technology

Microarrays allow the parallel measurement of the abundance of mRNA corresponding to thousands of genes (Lipshutz et al. 1995, Schena et al. 1995). In consequence, microarrays are considered to be a high-throughput technique (Lipshutz et al. 1999, Miron and Nadon 2006, Küster et al. 2007). Microarray technology evolved rapidly during the late 1990s and this technology has marked a turning point in functional genomics due to their wide range of applications and relative cost-efficiency. A multitude of diverse array technologies, protocols for their application, and statistical methods for data evaluation have been developed.

Despite the many technological differences at the detailed level, the common principle of all microarray platforms is rather straightforward. DNA molecules with a defined nucleotide sequences are attached to the surface of a solid support, usually coated glass. Molecules of the same type share a small region on the surface. These regions are called features or spots and are arranged in a grid pattern. Current technology enables a density of more than 10,000 features per cm^2. To measure messenger RNA, it has to be extracted and converted into cDNA by reverse transcription using reverse transcriptase. The cDNA is labelled with a fluorescent marker allowing quantification of the number of DNA copies. Some protocols also allow the use of mRNA directly, without reverse transcription. The solution of labelled molecules (also called targets) is then brought into contact with the surface of the microarray. In a process of parallel hybridization, the labelled single-stranded RNA or cDNA molecules hybridize with their single-stranded counterparts, representing the nucleotide sequence of the complementary strand on the surface of the microarray. The approximate number of target molecules bound to a given feature is measured by a detection device. Often a laser scanner producing visible images is used. The laser scanners excite fluorescent dyes with laser radiation of a defined bandwidth and this results in the emission of light of a defined emission bandwidth. The resulting images are processed and transformed into intensity measurements by image analysis software.

Many companies commercially produce microarrays. The technologies vary in many details, for example the array substrate, and in the process of generating the probe sequences. Spotted microarrays are produced by a deposition of small amounts of DNA solution on the substrate by robotic spotters. Other companies (for example Agilent) use ink-jet technology to deposit DNA. Affymetrix® and NimbleGen arrays are produced using a photolithographic process in which

nucleotides are used to synthesise longer oligonucleotides directly on the substrate. One of the most important parameters is the length of the probe sequences that can be deposited or synthesized. Sequence lengths can vary from a few hundred base-pairs, when cDNA is spotted, to short oligonucleotides. Commercial arrays are normally produced using short oligonucleotides. Affymetrix® arrays, for example have multiple different probe sequences per gene of 25 bp in length. Combining measurements made with several different probes is necessary to compensate for non specific binding of target DNA. Other oligonucleotide arrays contain oligos of between 50 and 80 bp, which tend to be more specific. Another important difference between microarray platforms is the number of channels they support. Two-colour microarrays support direct comparison of two experimental conditions per array. Single channel arrays can be used for single conditions or for a comparative study using multiple arrays. Furthermore, commercial array providers have different set-up costs and times for individual array designs and the number of features they support.

9.4.1.1 Variation and Replication

As with any other biological experiment, microarray experiments will show a certain amount of measurement variability and measurement error. Sources of variation in a microarray experiment can be divided into technical and biological variation.

Deviations in the technical process can have a large influence on the results of microarray experiments. The extent of these influences depends on the technology and microarray platform used and can be observed in replicated experiments in the same laboratory and also between laboratories. Primary causes are variations in the application of protocols. Further technical variation stems from the microarray production process such as variations in feature sizes and concentrations (Bammler et al. 2005). Some studies have also shown the large impact of differential probe sequences as a source of cross-platform variation. Other technical problems include scanner settings, as well as image segmentation and quantification (Yauk et al. 2004, Repsilber and Ziegler 2005, Yauk et al. 2005).

Even assuming perfect measurement technology, there would still be biological variation within the organism, either between different cells or between individuals. This variation is due to the individual genetic characteristics of the organisms and also due to variations in environmental conditions. The level of biological variation is assessed by performing the experiment multiple times under the same conditions and by harvesting several samples. This is termed replication. In general, biological variation seems to be a more important factor than technical variation for the assessment of the significance of the results obtained. Biological replications are preferable to technical replications (Allison et al. 2006).

Biological replicates will be influenced by both technical and biological variation, thereby serving to assess the overall variability of the experiment. If the number of replications exceeds the number of available microarrays, a pooling procedure is often used to generate a mixture of samples.

9.4.1.2 How Many Replicates?

It is important to note that any gene expression measurement pipeline will exhibit biological variation, not only those using microarrays. Any analysis should therefore contain biological replicates. The harvesting of cells to extract sufficient amounts of RNA from the organism can be particularly complicated for microarray experiments, especially when working with eukaryotes. The necessary number of replicates required in an experiment is thus an important question.

Unfortunately, there is no simple answer. Three replicates, however, is an often cited quantity (Lee et al. 2000, Yang and Speed 2002), and could serve as the rule of thumb for the lowest possible number. There also seems to be a growing trend for journals to require a minimum of three biological replicates for publication as well as experimental validation of results by a different technique. However, the suitable number of replicates depends on several factors. The biological variability is the most important, together with the size of the effect which the observer expects to observe. If variability is high, the number of replicates should be increased. Similarly, the smaller the measured effect, the more replicates are required to detect it. The required precision of the study is another important factor. If the experiment is aimed to discover a few candidate genes for further studies, it might be justified to reduce the number of replicates.

So called power analysis methods can serve to calculate the approximate number of replications necessary. They assist in experimental design by providing an estimate of the number of replicates, given the desired power (the ability to detect a large proportion of the differentially expressed genes), the confidence level, and the variability of the data – see for example Pan et al. (2002), Black and Doerge (2002), Li et al. (2005), and in particular Page et al. (2006) who have implemented the PowerAtlas software for power analysis based on publicly available data.

9.4.2 Gene Expression Analysis

The pipeline used for analysing the data obtained from a microarray experiment depends on the purpose of the experiment and the experimental question. Despite this, common analysis steps can be identified which are based on the features of the data. Analysis steps, which are often found in the literature, will be detailed in the following paragraphs.

9.4.2.1 Image Analysis

Data analysis of typical microarray experiments starts with the analysis of image data generated by the scanner software. Some novel microarray platforms provide direct signal readout via electro-chemical reactions. For these platforms, the image analysis step is not necessary. For approaches that do require image analysis, images for each channel are segmented. In other words, the locations of the features on the surface need to be identified. Most image analysis software can be calibrated

using a master grid for semi-automatic segmentation of spot images. There are also algorithms that can do a fully automated segmentation, but special care should be taken. Errors during segmentation can lead to erroneous results, because missing the correct location of a number of features can lead to measurements being assigned to the wrong feature.

In the next step, the image analysis software computes the intensity for each feature, and statistics such as measures of variation and background estimates are usually also calculated at the same time. Most software adds measures of signal quality. These are often called flags and can be used to filter out spots with low signal intensity or irregular shape. Automatic flags should be treated with care, as it is often not clear how these measures are derived, as each software does the analysis in a different way. If in doubt, it might be preferable to ignore flags and leave the removal for later stages in the analysis.

9.4.2.2 Normalization

To answer an experimental question, the various measurements coming from the image analysis need to be condensed to a single value or at least a lower number of values describing the intensity or the differential intensity of a feature.

The aim of applying normalization to microarray data is to make the data from different microarrays within an experiment comparable. Therefore, it is necessary to remove systematic bias from the datasets (Quackenbush 2002). A systematic bias in the data might originate from differences in RNA concentrations between samples, differences in scanner settings, and differences in the labelling, bleaching, and detection behaviour of the dyes. From an inspection of technical replicate arrays hybridized with the same labelled extract, it can be concluded that scanner settings also contribute to a large degree to the between-array bias. Cyanin-dyes 3 and 5 (Cy3 and Cy5) are currently the most frequently used fluorescent markers used for two-channel microarrays. These dyes emit different light intensities with respect to the number of hybridized target and this relation is not linear. This can lead to non-linear distortions, which can be visualised in scatter plots of the measured intensities of two channels (see Fig. 9.5).

Yang and Speed (2002) have developed a normalization method that uses a so called scatterplot smoothing function. They also propose a special logarithmic transformation of the raw microarray intensities that is suitable for plotting differential expression on a log-scale. This transformation maps the optionally background subtracted intensity values to a log-ratio (M) and a log-intensity measure (A):

$$M_i = \log R_i - \log G_i$$

$$A_i = \log R_i - \log G_i$$

where R_i and G_i denote the intensity of each channel for the ith feature. This representation has the advantage over normal ratios of providing a symmetric measure

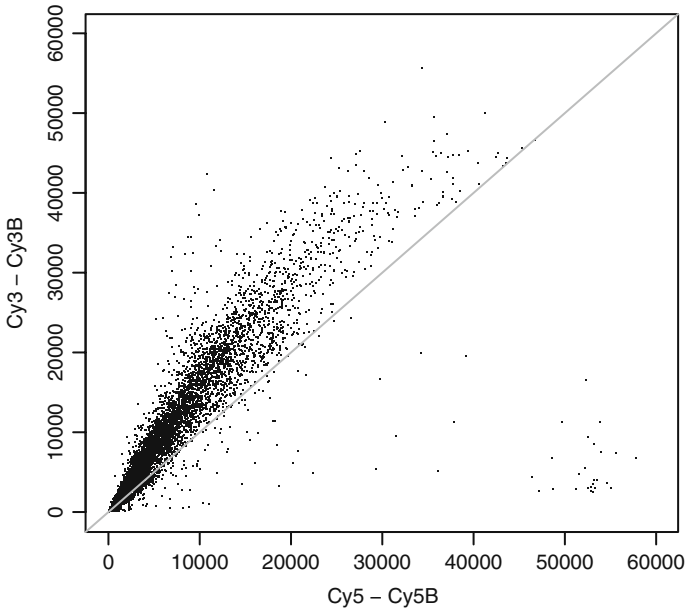

Fig. 9.5 Scatterplot of the raw intensities of the first microarray in the Swirl demo data from the microarray package for R. The raw channel intensities are background adjusted for each channel and plotted for each spot. The main diagonal is plotted as a *grey line*. The data distribution shows a visible deviation from the main diagonal

of differential expression (M) such that the absolute values of up-regulated genes are the same as the absolute values of down-regulated genes. Yang and Speed have proposed so called MA-plots in which both measures are combined as a means to inspect systematic variation and dye bias. These plots have since then become a standard tool in the analysis of microarray data (see Fig. 9.6).

9.4.2.3 Detecting Significant Changes

The most basic question to ask after performing a microarray experiment is which genes are significantly (differentially) up- or down-regulated in a sample or in a comparison of two samples. The inference step is of primary importance as for many experiments it is the only relevant analysis step (previous data acquisition and processing steps can be seen as preparations for the inference step). Also for machine learning steps, inference statistics play an important role for data reduction. In the earliest microarray studies fixed cut-offs for ratios or log-ratios were used. The choice of an ad hoc cut-off value is, however, arbitrary and was soon regarded as bad practice (Quackenbush 2001). Such a so-called fold-change approach fails to provide an estimate of measurement error. Without an estimate of variability, it is impossible to assess the probability of observing an event (in this case a specific M-value) within a sample just by chance.

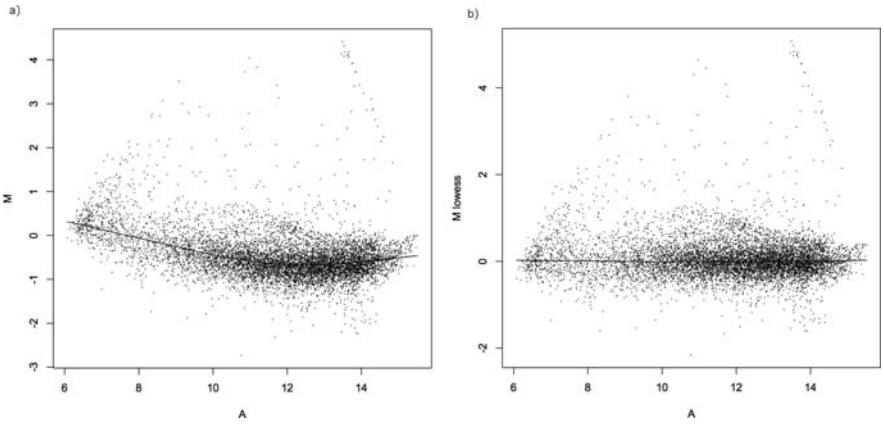

Fig. 9.6 The same data as in Fig. 9.5 after transformation. The y-axis corresponds to the logarithmic differential expression (M-value), the x-axis represents the logarithmic absolute expression measure (A-value). (**a**) depicts the data before and (**b**) after lowess-normalization. The lowess function used for normalizing the data is plotted as a *curve* following the centre of the data distribution

To analyse differential expression with respect to variability between replicates, methods of statistical inference or statistical tests are applied. Several statistical tests have been developed recently specifically for microarray data. These are based on classical test-theory, established by W.S. Gosset, Fischer, and many others in the early twentieth century. A hypothesis test is always based on the same principle:

There are two contradicting hypotheses: The null hypothesis (H0), here this means the transcript levels are not different, and the alternative hypothesis (H1) that the transcript levels differ significantly.

The null hypothesis (H0) is assumed to be true, unless there is enough evidence to reject it and to assume H1. A value, called a test statistic, is computed from sampled data to describe the empirical distribution of the data. With the test statistic as a summary of the data, the decision whether to reject the null hypothesis or not can be made. A threshold of the test statistic is computed for the rejection of the null hypothesis. The threshold is set in such a way that there is a sufficiently low probability of observing the value just by chance, when H0 is in fact true. Given that the distribution of the test-statistic is known, we can compute the probability of observing a statistic that is at least as extreme, in case that H0 is true. This probability is named the p-value. Unfortunately, there is often some confusion about the interpretation of p-values. With microarrays for example, the p-value can be interpreted as the probability of obtaining measurements such that we would call the gene differentially expressed, even though no such measurement may exist in a particular experiment.

There are several methods that can be used to statistically test microarray data. The classical t-test is among the most frequently used methods. It is applicable when the input data are normally distributed. Other methods can be applied should this

not be the case, for example, the Wilcoxon's Rank Sum Statistic, or the Significance Analysis of Microarrays (SAM) (Tusher et al. 2001) method. Other methods, such as the CyberT method (Baldi and Long 2001), attempt to address the problem of small numbers of replicates. The principal of CyberT is to borrow variance information from other genes on the microarray having a similar expression measurement.

9.4.2.4 Cluster Analysis

Cluster analysis (or simply "clustering") has become a very popular method to detect hidden structures in multivariate microarray data, and to detect co-regulated genes. The popularity of cluster-analysis is understandable, as it requires no or few prior hypotheses about the data. The application of cluster analysis is motivated by the "guilt-by-association" assumption. If genes share a common mechanism of regulation (for example the same transcription factors) they could also be functionally related; hence, it is useful to find groups of genes with similar expression profiles.

Clustering algorithms group measurements of gene expression into clusters. The cluster assignment may be a hard assignment to a single class or a weighted gradual assignment to multiple classes. The measurements are compared by their pair-wise dissimilarity. Therefore, the notion of dissimilarity or distance measure plays a central role for all clustering algorithms.

Hierarchical clustering algorithms construct a hierarchy of similar objects that can be represented as a rooted binary tree dendrogram. Two types of hierarchical clustering exist: agglomerative and divisive clustering. Agglomerative clustering is the most popular approach.

Agglomerative clustering uses a bottom up approach. All objects start as singleton clusters and the most similar clusters are joined to form bigger clusters in each step. The opposite approach is taken in divisive clustering algorithms where all the genes initially constitute a single large cluster and are divided into smaller and smaller clusters. The use of hierarchical clustering for the analysis of microarray data was popularized by Eisen et al. (1998) who also developed a software tool to perform cluster analyses on microarrays. An appealing method for visualisation of the results of the hierarchical clustering as a heat map is also presented in this publication. The expression values are represented by colour codes; a red-green representation is used to denote the measured values. Negative log-ratios are projected as green values and positive as red values, yielding black for values close to zero.

9.4.2.5 Classification

Sometimes, there is prior information about the origin of a certain sample. In such cases, microarray data can be used to predict class information from the expression profiles. The process of classification can be defined as assigning measurements to discrete class labels.

All classification methods share the concept of a training phase and a classification phase. The data with known origin is used as the training set. Then a sample with unknown origin can be classified into a class respective to its origin. During the

classification phase, the trained classifier is applied to data and predicts class labels for these.

One of the first applications of this method was the study by Golub et al. (1999) who used microarray data for the classification of different types of leukaemia. One of the most simple but highly useful methods is the k nearest-neighbour (kNN) classifier which makes class predictions for an unknown object based on the majority of the k closest labelled examples (Cover and Hart 1967). Wu et al. (2005) have used a kNN classifier to compare the merit of different normalization methods.

Vapnik (1999) developed the Support Vector Machine (SVM) approach for classification. SVMs are suited for finding an optimal separating function and allow for linear and non-linear decision boundaries. Examples of applications of SVMs to microarray data are frequently found in the literature (see for example, Pavlidis et al. 2002, Liu et al. 2005) and they are becoming more and more popular.

9.4.2.6 Microarray Software

An almost unmanageable amount of software related to microarrays has been developed over the recent years. The domains of application of these software tools also often overlap. Here we will, therefore, concentrate on several examples. Unfortunately, there is currently no perfect solution and all software packages have their advantages and drawbacks. In particular there is no all-in-one solution that offers both highly flexible analysis methods and an easy to use graphical user interface. For most analysis tasks, this means that the analysis will involve the use of multiple software programs. Commercial software is available for most tasks, with a large range of license fees as well as free open source solutions. Another important aspect when choosing software is its hardware and software requirements. Some software packages, especially database-based systems, require a significant effort for installation and professional administration. Several categories of software are described in the following sections.

9.4.2.7 Image Analysis Software

Image analysis software serves the purpose of analysing the resulting images obtained from the scanner software. This process is usually twofold. A segmentation step is performed to identify the locations of the features and their boundaries within the image, corresponding to spots of the microarray. Depending on the software, this step can be carried out manually, semi-automatically if predefined grid information is used, or fully automatically.

The TIGR Spotfinder software for example, is an academic open-source project, which is available for several operating systems. A package called Spot also exists. This is implemented as a package for the statistical environment R. Users of Affymetrix® microarrays have to use the MAS5 algorithm from Affymetrix® for image analysis.

9.4.2.8 Pure Analysis Systems

Pure analysis systems provide variable collections of analysis algorithms, but they normally do not take care of data-management. Data is normally stored in files and user interaction is possible via a graphical user interface (GUI). One of the first examples was the Cluster and TreeView software, which was developed by Eisen et al. (1998), and focused on clustering algorithms. Cluster and TreView work only on Windows. The Genesis software is another application that provides various clustering methods (Sturn et al. 2002). It is implemented in Java and thus operating system independent. ArrayNorm (Pieler et al. 2004) and MIDAS (Saeed et al. 2003) are programs that provide normalization for two-colour arrays and some statistical tests.

Users of Affymetrix® microarrays can use the dChip software that provides normalization, clustering, and classification methods for oligonucleotide arrays (Li and Wong 2001, Lin et al. 2004).

In summary, pure analysis tools are recommended for small to medium size laboratories. In general, they have minimal resource requirements, but data-management and collaborative functions are not provided. For larger projects with many participants these systems can be useful to complement a database-based system.

9.4.2.9 General Purpose Database Systems

There is also a class of systems that combine general data analysis functions with data-management capabilities. These systems use a database system to systematically store and retrieve data. Furthermore, all allow the annotation of experiments by storing protocols. These systems include the open source systems BASE (Saal et al. 2002), MARS (Maurer et al. 2005), MADAM (Saeed et al. 2003), and EMMA (Dondrup et al. 2003) (Dondrup et al. 2009). All systems provide user and account management and collaborative functions such as making experimental data publicly available. Often, the software is accessed via a web-browser. Unfortunately, these systems have relatively high requirements with respect to installation and administration of the server, which make them impractical for small laboratories. For larger institutions it might be a good option to set up and maintain a central installation of such a system, in particular if researchers work as part of an international collaboration.

9.4.2.10 R and BioConductor

The statistical environment R (Team 2008) has a special and prominent position among all analysis systems. It is a general purpose statistical environment and also a powerful programming language. The BioConductor project bundles and provides add-on packages for many bioinformatics applications including microarrays and sequence analysis (Gentleman et al. 2004). In comparison to other analysis systems R is highly flexible and provides the largest amount of analysis algorithms. General statistical functions also applicable to microarrays comprise various statistical test

methods and graphical displays. Within the BioConductor packages there are normalization methods for two-colour arrays as well as for Affymetrix® arrays. As a downside of this flexibility, the system requires a high level of expertise. User interaction is mediated by a command line interface, and complex actions might even involve programming. Anyway, we would recommend gaining an insight into and to trying to work with R and BioConductor, as this combination provides maximal flexibility in the choice of methods and their combination and full control over the analysis process. Also, novel methods are often first implemented as BioConductor packages. We refer the reader to Gentleman (Gentleman et al. 2005) for further reference.

9.4.3 Data Sharing and Public Repositories

Gene expression experiments involve highly complex protocols and produce high-volume data. Compared with a genome sequencing approach the experimental annotations need to be much more precise, as a good annotation of a functional genomics experiment also involves environmental conditions. During growth, harvesting protocols, and further laboratory procedures, samples are transformed many times. Also complex array designs are involved which can carry up to many hundreds of thousands of features.

The experimental results of a microarray experiment need to be accompanied with all this meta-information. Otherwise, it is almost impossible to interpret the resulting data or even reproduce the experiment independently, which is still a difficult and underestimated task. To ensure scientific standards, almost all journals in functional genomics now require the submission of microarray data to a public repository prior to publication of results (Ball et al. 2004).

The most frequently used repositories are:

- ArrayExpress hosted by the EBI (http://www.ebi.ac.uk) (Parkinson et al. 2007).
- Gene Expression Omnibus (GEO) at the NCBI (http://www.ncbi.nlm.nih.gov/) (Barrett et al. 2007).
- Stanford Microarray Database (SMD) at Stanford University (http://www.stanford.edu) (Demeter et al. 2007).

All repositories have web-forms to query for data, and they offer some data-analysis features such as normalization, filtering, and clustering. They all offer a web-based submission process, which is recommended for small to medium-scale experiments. The choice of repository is a matter of personal preference, but we recommend that each laboratory uses one repository consistently.

The Microarray Gene Expression Data (MGED) Society (http://www.mged.org) has developed and promoted standards, recommendations, and tools for sharing microarray data. The most important in the context of publishing microarray results is the Minimal Information About a Microarray Experiment (MIAME) standard

(Brazma et al. 2001). It is aimed at standardizing the necessary content of a submission to a public database.

The MicroArray Gene Expression Markup Language (MAGE-ML) format was developed by the MGED Society based on XML to provide a standardized document format to exchange microarray data (Spellman et al. 2002). However, this document format is highly complex and should therefore be only used for data exchange between software applications. It has been complemented by a tab-delimited spreadsheet-based format called MAGE-TAB (Rayner et al. 2006), which is much simpler in structure and should now be preferred for large-scale submissions.

9.4.4 Summary of the Gene Expression Analysis Section

Gene expression analysis is a highly flexible and promising approach within marine genomics and expression data can greatly aid the inference of functions from sequence data. Quantitative analysis of transcriptomes can be performed using microarrays in combination with quantitative RT-PCR for validation. If only a few known genes are to be investigated, the use of microarrays is not necessary but quantitative RT-PCR could be used instead.

While qRT-PCR and microarrays require the sequence of the transcript to be known but the SAGE method generates transcript sequence data during the experimentation and can therefore be applied to genomes that have not been sequenced. Genome wide expression profiling is only possible with the SAGE and microarrays approaches. The most popular method of these two is the microarray technology, as it allows mRNA levels of thousands of genes to be studied in parallel in a cost efficient manner. Also, SAGE can be applied only to Eukaryotes.

Good experimental design and an appropriate level of replication are crucial for microarray experiments. The optimal setup depends on the biological question. In principle biological replicates should be preferred over technical replicates and at least three biological replicates should be made.

There are a growing number of commercial providers of microarrays and related services. For large scale comparative studies there is no clearly preferable platform, thus the choice of array provider should be based on criteria such as budget, availability of designs and services, and in particular practical experience with the methods. However, setting up an in-house array production pipeline is not recommended for small to medium sized labs because of high setup costs and the investment in time. For large labs or consortia, this might still be an option, while for occasional applications a full-service provider should be considered.

Each experiment has a certain level of unavoidable experimental variation, however technical variation can be reduced by following experimental protocols rigidly.

Data analysis is a complex task and requires bioinformatics and statistical expertise. Depending on the experimental setup, simple analyses such as statistical tests

are often sufficient. Normalization of microarray data however is crucial. Initially, it is preferable to use well-established and well-understood statistical methods (e.g. t-tests, ANOVA) to identify genes with significantly different patterns of expression, while critically assessing the merits of novel tools for the given application. In almost every case, several competing methods should be tested.

For small laboratories and occasional use, stand-alone software with low administrative requirements should be adopted initially. For large labs and international collaborations, database based systems may be used, providing collaborative functions such as data sharing. To reduce administrative effort, a central installation of such a database and analysis software can be provided. Due to its high flexibility, the use of the R and BioConductor environment should be considered, at least as a complementary tool.

Before publication, experimental data and annotations should be submitted to one of the public repositories. For high-volume submissions, an automated submission based on MAGE-TAB should be considered.

Despite the recent success of microarrays, we expect sequencing based methods will become increasingly popular and efficient in the mid-term. For classical gene-expression studies, sequencing methods and especially shotgun-transcriptomics are likely to surpass microarrays in terms of cost efficiency and precision of measurement in the coming years.

Acknowledgments We are grateful to the CeBiTec at Bielefeld University, the BMBF Competence Network GenoMik-Plus (grant 0313805A), the International NRW Graduate School in Bioinformatics and Genome Research, the EU FP6 Network of Excellence Marine Genomics Europe (contract No. COGE-CT-2004-505403) and Nestlé Research Center for financial support of our work. Special thanks to our native speaker Sita Lange, the chapter would not have been the same without her efforts. The authors would also like to thank Guy Cochrane, Naryttza Diaz, Michele Magrane, Nicky Mulder, Kai Runte and Rafael Szczepanowski, who read sections of the chapter and provided valuable comments. Many thanks to our present and former colleagues from the Junior Group Computational Genomics for their patience during the writing.

References

Adams CP, Kron SJ, Mosaic Technologies USA (1997) Method for performing amplification of nucleic acid with two primers bound to a single solid support. US Patent 5,641,658.

Alexandersson M, Cawley S, Pachter L (2003) SLAM: cross-species gene finding and alignment with a generalized pair hidden Markov model. Genome Res 13(3):496–502

Allen JE, Salzberg SL (2005) JIGSAW: integration of multiple sources of evidence for gene prediction. Bioinformatics 21(18):3596–3603

Allison DB, Cui X, Page GP et al (2006) Microarray data analysis: from disarray to consolidation and consensus. Nat Rev Genet 7(1):55–65

Altschul SF, Gish W, Miller W et al (1990) Basic local alignment search tool. J Mol Biol 215(3):403–410

Ashburner M, Ball CA, Blake JA et al (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25(1):25–29

Aziz RK, Bartels D, Best AA et al (2008) The RAST Server: rapid annotations using subsystems technology. BMC Genomics 9:75

Badger JH, Olsen GJ (1999) CRITICA: coding region identification tool invoking comparative analysis. Mol Biol Evol 16(4):512–524

Baldi P, Long AD (2001) A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes. Bioinformatics 17(6): 509–519

Ball CA, Brazma A, Causton H et al (2004) Submission of microarray data to public repositories. PLoS Biol 2(9):E317

Bammler T, Beyer RP, Bhattacharya S et al (2005) Standardizing global gene expression analysis between laboratories and across platforms. Nat Methods 2(5):351–356

Barrett T, Troup DB, Wilhite SE et al (2007) NCBI GEO: mining tens of millions of expression profiles-database and tools update. Nucleic Acids Res 35(Database issue):D760–D765

Bartels D, Kespohl S, Albaum S et al (2005) BACCardI-a tool for the validation of genomic assemblies, assisting genome finishing and intergenome comparison. Bioinformatics 21(7):853–859

Bauerle RH, Margolin P (1966) The functional organization of the tryptophan gene cluster in *Salmonella typhimurium*. Proc Natl Acad Sci U S A 56(1):111–118

Bekel T, Henckel K, Küster H et al (2009) The sequence analysis and management system – SAMS-2.0: data management and sequence analysis adapted to changing requirements from traditional sanger sequencing to ultrafast sequencing technologies. J Biotechnol 140(1–2):3–12

Bendtsen JD, Nielsen H, von Heijne G et al (2004) Improved prediction of signal peptides: SignalP 3.0. J Mol Biol 340(4):783–795

Benson DA, Karsch-Mizrachi I, Lipman DJ et al (2008) GenBank. Nucleic Acids Res 36:D25–D30

Berman H, Henrick K, Nakamura H (2003) Announcing the worldwide Protein Data Bank. Nat Struct Biol 10(12):980

Besemer J, Borodovsky M (2005) GeneMark: web software for gene finding in prokaryotes, eukaryotes and viruses. Nucleic Acids Res 33:W451–W454

Besemer J, Lomsadze A, Borodovsky M (2001) GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. Nucleic Acids Res 29(12):2607–2618

Birney E, Clamp M, Durbin R (2004) GeneWise and Genomewise. Genome Res 14(5):988–995

Black MA, Doerge RW (2002) Calculation of the minimum number of replicate spots required for detection of significant gene expression fold change in microarray experiments. Bioinformatics 18(12):1609–1616

Brazma A, Hingamp P, Quackenbush J et al (2001) Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. Nat Genet 29(4):365–371

Brejova B, Brown DG, Li M et al (2005) ExonHunter: a comprehensive approach to gene finding. Bioinformatics 21(Suppl 1):i57–i65

Brent MR (2007) How does eukaryotic gene prediction work? Nat Biotechnol 25(8):883–885

Brunak S, Danchin A, Hattori M et al (2002) Nucleotide sequence database policies. Science 298(5597):1333

Burge C, Karlin S (1997) Prediction of complete gene structures in human genomic DNA. J Mol Biol 268(1):78–94

Chen YA, Lin CC, Wang CD et al (2007) An optimized procedure greatly improves EST vector contamination removal. BMC Genomics 8:416

Chothia C, Gough J, Vogel C et al (2003) Evolution of the protein repertoire. Science 300(5626):1701–1703

Cochrane G, Bates K, Apweiler R et al (2006) Evidence standards in experimental and inferential INSDC Third Party Annotation data. Omics 10(2):105–113

Cochrane G, Akhtar R, Aldebert P et al (2008) Priorities for nucleotide trace, sequence and annotation data capture at the ensembl trace archive and the EMBL nucleotide sequence database. Nucleic Acids Res 36:D5–D12

Codd EF (1990) The relational model for database management: version 2. Addison-Wesley Longman Publishing Co., Inc, New York.

Conesa A, Gotz S, Garcia-Gomez JM et al (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 21(18): 3674–3676

Consortium U (2008) The universal protein resource (UniProt). Nucleic Acids Res 36:D190–D195

Cover T, Hart P (1967) Nearest neighbor pattern classification. IEEE Trans Inf Theory 13(1):21–27
Dandekar T, Snel B, Huynen MA et al (1998) Conservation of gene order: a fingerprint of proteins that physically interact. Trends Biochem Sci 23(9):324–328
Datson NA, van der Perk-de Jong J, van den Berg MP et al (1999) MicroSAGE: a modified procedure for serial analysis of gene expression in limited amounts of tissue. Nucleic Acids Res 27(5):1300–1307
Delcher AL, Bratke KA, Powers EC et al (2007) Identifying bacterial genes and endosymbiont DNA with Glimmer. Bioinformatics 23(6):673–679
Delcher AL, Harmon D, Kasif S et al (1999) Improved microbial gene identification with GLIMMER. Nucleic Acids Res 27(23):4636–4641
Demeter J, Beauheim C, Gollub J et al (2007) The Stanford microarray database: implementation of new analysis tools and open source release of software. Nucleic Acids Res 35:D766–D770
Djebali S, Delaplace F, Crollius HR (2006) Exogean: a framework for annotating protein-coding genes in eukaryotic genomic DNA. Genome Biol 7(Suppl 1):S7–S10
Dondrup M, Goesmann A, Bartels D et al (2003) EMMA: a platform for consistent storage and efficient analysis of microarray data. J Biotechnol 106(2-3):135–146
Dondrup M, Albaum S, Griebel T et al (2009) EMMA 2 – A MAGE-compliant system for the collaborative analysis and integration of microarray data. BMC Bioinformatics 10(1):50
Dressman D, Yan H, Traverso G et al (2003) Transforming single DNA molecules into fluorescent magnetic particles for detection and enumeration of genetic variations. Proc Natl Acad Sci U S A 100(15):8817–8822
Durbin R, Eddy S, Krogh A et al (1998) Biological sequence analysis. Cambridge University Press, Cambridge.
Edwards RA, Rodriguez-Brito B, Wegley L et al (2006) Using pyrosequencing to shed light on deep mine microbial ecology. BMC Genomics 7:57
Eisen MB, Spellman PT, Brown PO et al (1998) Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci U S A 95(25):14863–14868
Elsik CG, Mackey AJ, Reese JT et al (2007) Creating a honey bee consensus gene set. Genome Biol 8(1):R13
Emanuelsson O, Nielsen H, von Heijne G (1999) ChloroP, a neural network-based method for predicting chloroplast transit peptides and their cleavage sites. Protein Sci 8(5):978–984
Emanuelsson O, Brunak S, von Heijne G et al (2007) Locating proteins in the cell using TargetP, SignalP and related tools. Nat Protoc 2(4):953–971
Ewing B, Hillier L, Wendl MC et al (1998) Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res 8(3):175–185
Fedurco M, Romieu A, Williams S et al (2006) BTA, a novel reagent for DNA attachment on glass and efficient generation of solid-phase amplified DNA colonies. Nucleic Acids Res 34(3): e22
Fleischmann RD, Adams MD, White O et al (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science 269(5223):496–512
Flicek P, Aken BL, Beal K et al (2008) Ensembl 2008. Nucleic Acids Res 36:D707–D714
Florea L, Hartzell G, Zhang Z et al (1998) A computer program for aligning a cDNA sequence with a genomic DNA sequence. Genome Res 8(9):967–974
Gaasterland T, Sczyrba A, Thomas E et al (2000) MAGPIE/EGRET annotation of the 2.9-Mb *Drosophila melanogaster* Adh region. Genome Res 10:502–510
Gartemann KH, Abt B, Bekel T et al (2008) The genome sequence of the tomato-pathogenic actinomycete *Clavibacter michiganensis* subsp. michiganensis NCPPB382 reveals a large island involved in pathogenicity. J Bacteriol 190(6):2138–2149
Gentleman R, Huber W, Carev VJ (eds) (2005) Bioinformatics and computational biology solutions using R and bioconductor. Springer, New York.
Gentleman RC, Carey VJ, Bates DM et al (2004) Bioconductor: open software development for computational biology and bioinformatics. Genome Biol 5(10):R80

Goesmann A, Linke B, Bartels D et al (2005) BRIGEP-the BRIDGE-based genome-transcriptome-proteome browser. Nucleic Acids Res 33:W710–W716
Goldberg SMD, Johnson J, Busam D et al (2006) A Sanger/pyrosequencing hybrid approach for the generation of high-quality draft assemblies of marine microbial genomes. Proc Natl Acad Sci U S A 103(30):11240–11245
Golub TR, Slonim DK, Tamayo P et al (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 286(5439):531–537
Gordon D, Abajian C, Green P (1998) Consed: a graphical tool for sequence finishing. Genome Res 8(3):195–202
Gordon D, Desmarais C, Green P (2001) Automated finishing with autofinish. Genome Res 11(4):614–625
Gouy M, Gautier C (1982) Codon usage in bacteria: correlation with gene expressivity. Nucleic Acids Res 10(22):7055–7074
Green P (2002) Whole-genome disassembly. Proc Natl Acad Sci U S A 99(7):4143–4144
Gresham D, Ruderfer DM, Pratt SC et al (2006) Genome-wide detection of polymorphisms at nucleotide resolution with a single DNA microarray. Science 311(5769):1932–1936
Gross SS, Brent MR (2006) Using multiple alignments to improve gene prediction. J Comput Biol 13(2):379–393
Guigo R, Reese MG (2005) EGASP: collaboration through competition to find human genes. Nat Methods 2(8):575–577
Guigo R, Flicek P, Abril JF et al (2006) EGASP: the human ENCODE Genome Annotation Assessment Project. Genome Biol 7(Suppl 1):S2–S31
Guo FB, Ou HY, Zhang CT (2003) ZCURVE: a new system for recognizing protein-coding genes in bacterial and archaeal genomes. Nucleic Acids Res 31(6):1780–1789
Haas BJ, Salzberg SL, Zhu W et al (2008) Automated eukaryotic gene structure annotation using EVidenceModeler and the program to assemble spliced alignments. Genome Biol 9(1): R7
Henrick K, Feng Z, Bluhm WF et al (2008) Remediation of the protein data bank archive. Nucleic Acids Res 36:D426–D433
Herring CD, Raghunathan A, Honisch C et al (2006) Comparative genome sequencing of *Escherichia coli* allows observation of bacterial evolution on a laboratory timescale. Nat Genet 38(12):1406–1412
Huang X, Madan A (1999) CAP3: a DNA sequence assembly program. Genome Res 9(9):868–877
Huang X, Adams MD, Zhou H et al (1997) A tool for analyzing and annotating genomic sequences. Genomics 46(1):37–45
Iizuka M, Yamauchi M, Ando K et al (1994) Quantitative RT-PCR assay detecting the transcriptional induction of vascular endothelial growth factor under hypoxia. Biochem Biophys Res Commun 205(2):1474–1480
Iseli C, Jongeneel CV, Bucher P (1999) ESTScan: a program for detecting, evaluating, and reconstructing potential coding regions in EST sequences. Proc Int Conf Intell Syst Mol Biol 7:138–148
Ju J, Kim DH, Bi L et al (2006) Four-color DNA sequencing by synthesis using cleavable fluorescent nucleotide reversible terminators. Proc Natl Acad Sci U S A 103(52): 19635–19640
Kaiser O, Bartels D, Bekel T et al (2003) Whole genome shotgun sequencing guided by bioinformatics pipelines-an optimized approach for an established technique. J Biotechnol 106(2–3):121–133
Kall L, Krogh A, Sonnhammer EL (2007) Advantages of combined transmembrane topology and signal peptide prediction-the Phobius web server. Nucleic Acids Res 35:W429–W432
Kanehisa M, Goto S (2000) KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res 28:27–30
Kent WJ (2002) BLAT-the BLAST-like alignment tool. Genome Res 12(4):656–664
Korf I (2004) Gene finding in novel genomes. BMC Bioinformatics 5:59

Korf I, Flicek P, Duan D et al (2001) Integrating genomic homology into gene structure prediction. Bioinformatics 17(Suppl 1):S140–S148

Krause A, Ramakumar A, Bartels D et al (2006) Complete genome of the mutualistic, N2-fixing grass endophyte *Azoarcus* sp. strain BH72. Nat Biotechnol 24(11):1385–1391

Krause L, McHardy AC, Nattkemper TW et al (2007) GISMO-gene identification using a support vector machine for ORF classification. Nucleic Acids Res 35(2):540–549

Krogh A, Larsson B, von Heijne G et al (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J Mol Biol 305(3):567–580

Küster H, Becker A, Firnhaber C et al (2007) Development of bioinformatic tools to support EST-sequencing, in silico- and microarray-based transcriptome profiling in mycorrhizal symbioses. Phytochemistry 68(1):19–32

Lafay B, Lloyd AT, McLean MJ et al (1999) Proteome composition and codon usage in spirochaetes: species-specific and DNA strand-specific mutational biases. Nucleic Acids Res 27(7):1642–1649

Lagesen K, Hallin P, Rodland EA et al (2007) RNAmmer: consistent and rapid annotation of ribosomal RNA genes. Nucleic Acids Res 35(9):3100–3108

Lander ES, Waterman MS (1988) Genomic mapping by fingerprinting random clones: a mathematical analysis. Genomics 2(3):231–239

Larsen TS, Krogh A (2003) EasyGene-a prokaryotic gene finder that ranks ORFs by statistical significance. BMC Bioinformatics 4:21

Lawrence JG, Roth JR (1996) Selfish Operons: horizontal transfer may drive the evolution of gene clusters. Genetics 143(4):1843–1860

Lee ML, Kuo FC, Whitmore GA et al (2000) Importance of replication in microarray gene expression studies: statistical methods and evidence from repetitive cDNA hybridizations. Proc Natl Acad Sci U S A 97(18):9834–9839

Li C, Wong WH (2001) Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. Proc Natl Acad Sci U S A 98(1):31–36

Li SS, Bigler J, Lampe JW et al (2005) FDR-controlling testing procedures and sample size determination for microarrays. Stat Med 24(15):2267–2280

Lin M, Wei LJ, Sellers WR et al (2004) dChipSNP: significance curve and clustering of SNP-array-based loss-of-heterozygosity data. Bioinformatics 20(8):1233–1240

Linke B, McHardy AC, Neuweger H et al (2006) REGANOR: a gene prediction server for prokaryotic genomes and a database of high quality gene predictions for prokaryotes. Appl Bioinformatics 5(3):193–198

Liolios K, Mavromatis K, Tavernarakis N et al (2008) The genomes on line database (GOLD) in 2007: status of genomic and metagenomic projects and their associated metadata. Nucleic Acids Res 36:D475–D479

Lipshutz RJ, Fodor SP, Gingeras TR et al (1999) High density synthetic oligonucleotide arrays. Nat Genet 21(1 Suppl):20–24

Lipshutz RJ, Morris D, Chee M et al (1995) Using oligonucleotide probe arrays to access genetic diversity. Biotechniques 19(3):442–447

Liu JJ, Cutler G, Li W et al (2005) Multiclass cancer classification and biomarker discovery using GA-based algorithms. Bioinformatics 21(11):2691–2697

Lomsadze A, Ter Hovhannisyan V, Chernoff YO et al (2005) Gene identification in novel eukaryotic genomes by self-training algorithm. Nucleic Acids Res 33(20):6494–6506

Lowe TM, Eddy SR (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res 25(5):955–964

Lukashin AV, Borodovsky M (1998) GeneMark.hmm: new solutions for gene finding. Nucleic Acids Res 26(4):1107–1115

Majoros WH, Pertea M, Salzberg SL (2004) TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. Bioinformatics 20(16):2878–2879

Majoros WH, Pertea M, Salzberg SL (2005) Efficient implementation of a generalized pair hidden Markov model for comparative gene finding. Bioinformatics 21(9):1782–1788

Mangalam H (2002) The Bio∗ toolkits-a brief overview. Brief Bioinform 3(3):296–302
Mao X, Cai T, Olyarchuk JG et al (2005) Automated genome annotation and pathway identification using the KEGG Orthology (KO) as a controlled vocabulary. Bioinformatics 21(19):3787–3793
Mardis ER (2008) Next-generation DNA sequencing methods. Annu Rev Genomics Hum Genet 9:387–402
Margulies M, Egholm M, Altman WE et al (2005) Genome sequencing in microfabricated high-density picolitre reactors. Nature 437(7057):376–380
Mathe C, Sagot MF, Schiex T et al (2002) Current methods of gene prediction, their strengths and weaknesses. Nucleic Acids Res 30(19):4103–4117
Matsumura H, Reich S, Ito A et al (2003) Gene expression analysis of plant host-pathogen interactions by SuperSAGE. Proc Natl Acad Sci U S A 100(26):15718–15723
Maurer M, Molidor R, Sturn A et al (2005) MARS: microarray analysis, retrieval, and storage system. BMC Bioinformatics 6:101
McHardy AC, Pühler A, Kalinowski J et al (2004a) Comparing expression level-dependent features in codon usage with protein abundance: an analysis of ′predictive proteomics′. Proteomics 4(1):46–58
McHardy AC, Goesmann A, Pühler A et al (2004b) Development of joint application strategies for two microbial gene finders. Bioinformatics 20(10):1622–1631
Meyer F, Goesmann A, McHardy AC et al (2003) GenDB-an open source genome annotation system for prokaryote genomes. Nucleic Acids Res 31(8):2187–2195
Millar CD, Huynen L, Subramanian S et al (2008) New developments in ancient genomics. Trends Ecol Evol 23(7):386–393
Miron M, Nadon R (2006) Inferential literacy for experimental high-throughput biology. Trends Genet 22(2):84–89
Moore JE, Lake JA (2003) Gene structure prediction in syntenic DNA segments. Nucleic Acids Res 31(24):7271–7279
Mott R (1997) EST_GENOME: a program to align spliced DNA sequences to unspliced genomic DNA. Comput Appl Biosci 13(4):477–478
Mulder NJ, Apweiler R, Attwood TK et al (2007) New developments in the InterPro database. Nucleic Acids Res 35:D224–D228
Nagaraj SH, Deshpande N, Gasser RB et al (2007) ESTExplorer: an expressed sequence tag (EST) assembly and annotation platform. Nucleic Acids Res 35:W143–W147
Nakano M, Komatsu J, Matsuura S-i et al (2003) Single-molecule PCR using water-in-oil emulsion. J Biotechnol 102(2): 117–124
Needleman SB, Wunsch CD (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol 48(3):443–453
Nekrutenko A, Chung WY, Li WH (2003) ETOPE: evolutionary test of predicted exons. Nucleic Acids Res 31(13):3564–3567
Ng P, Wei C-L, Sung W-K et al (2005) Gene identification signature (GIS) analysis for transcriptome characterization and genome annotation. Nat Methods 2(2):105–111
Ng P, Tan JJS, Ooi HS et al (2006) Multiplex sequencing of paired-end ditags (MS-PET): a strategy for the ultra-high-throughput analysis of transcriptomes and genomes. Nucleic Acids Res 34(12):e84
Noguchi H, Park J, Takagi T (2006) MetaGene: prokaryotic gene finding from environmental genome shotgun sequences. Nucleic Acids Res 34(19):5623–5630
Ou HY, Guo FB, Zhang CT (2004) GS-Finder: a program to find bacterial gene start sites with a self-training method. Int J Biochem Cell Biol 36(3):535–544
Overbeek R, Disz T, Stevens R (2004) The SEED: a peer-to-peer environment for genome annotation. Commun ACM 47(11):47–51
Overbeek R, Fonstein M, D'Souza M et al (1999) The use of gene clusters to infer functional coupling. Proc Natl Acad Sci U S A 96:2896–2901
Overbeek R, Larsen N, Pusch GD et al (2000) WIT: integrated system for high-throughput genome sequence analysis and metabolic reconstruction. Nucleic Acids Res 28(1):123–125

Overbeek R, Larsen N, Walunas T et al (2003) The ERGO genome analysis and discovery system. Nucleic Acids Res 31:164–171

Overbeek R, Begley T, Butler RM et al (2005) The subsystems approach to genome annotation and its use in the project to annotate 1,000 genomes. Nucleic Acids Res 33(17):5691–5702

Page GP, Edwards JW, Gadbury GL et al (2006) The PowerAtlas: a power and sample size atlas for microarray experimental design and research. BMC Bioinformatics 7:84

Pan W, Lin J, Le CT (2002) How many replicates of arrays are required to detect gene expression changes in microarray experiments? A mixture model approach. Genome Biol. research0022.

Parkinson H, Kapushesky M, Shojatalab M et al (2007) ArrayExpress-a public database of microarray experiments and gene expression profiles. Nucleic Acids Res 35:D747–D750

Parra G, Agarwal P, Abril JF et al (2003) Comparative gene prediction in human and mouse. Genome Res 13(1):108–117

Pavlidis P, Weston J, Cai J et al (2002) Learning gene functional classifications from multiple data types. J Comput Biol 9(2):401–411

Pearson WR, Lipman DJ (1988) Improved tools for biological sequence comparison. Proc Natl Acad Sci U S A 85(8):2444–2448

Pertea G, Huang X, Liang F et al (2003) TIGR Gene Indices clustering tools (TGICL): a software system for fast clustering of large EST datasets. Bioinformatics 19(5):651–652

Pieler R, Sanchez-Cabo F, Hackl H et al (2004) ArrayNorm: comprehensive normalization and analysis of microarray data. Bioinformatics 20(12):1971–1973

Prober JM, Trainor GL, Dam RJ et al (1987) A system for rapid DNA sequencing with fluorescent chain-terminating dideoxynucleotides. Science 238(4825):336–341

Pruitt KD, Tatusova T, Maglott DR (2007) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res 35: D61–D65

Quackenbush J (2001) Computational analysis of microarray data. Nat Rev Genet 2(6): 418–427

Quackenbush J (2002) Microarray data normalization and transformation. Nat Genet 32(Suppl):496-501

Quackenbush J (2003) Genomics. Microarrays-guilt by association. Science 302(5643):240–241

Quevillon E, Silventoinen V, Pillai S et al (2005) InterProScan: protein domains identifier. Nucleic Acids Res 33:W116–W1120

Rayner TF, Rocca-Serra P, Spellman PT et al (2006) A simple spreadsheet-based, MIAME-supportive format for microarray data: MAGE-TAB. BMC Bioinformatics 7:489

Reeck GR, de Haen C, Teller DC et al (1987) Homology in proteins and nucleic acids: a terminology muddle and a way out of it. Cell 50(5):667

Reese MG, Kulp D, Tammana H et al (2000) Genie-gene finding in *Drosophila melanogaster*. Genome Res 10(4):529–538

Repsilber D, Ziegler A (2005) Two-color microarray experiments. Technology and sources of variance. Methods Inf Med 44(3):400–404

Ronaghi M, Uhlén M, Nyrén P (1998) A sequencing method based on real-time pyrophosphate. Science 281(5375):363–365

Rutherford K, Parkhill J, Crook J et al (2000) Artemis: sequence visualization and annotation. Bioinformatics 16(10):944–945

Saal LH, Troein C, Vallon-Christersson J et al (2002) BioArray Software Environment (BASE): a platform for comprehensive management and analysis of microarray data. Genome Biol 3(8): SOFTWARE0003.

Saeed AI, Sharov V, White J et al (2003) TM4: a free, open-source system for microarray data management and analysis. Biotechniques 34(2):374–378

Saha S, Sparks AB, Rago C et al (2002) Using the transcriptome to annotate the genome. Nat Biotechnol 20(5):508–512

Salamov AA, Solovyev VV (2000) Ab initio gene finding in *Drosophila* genomic DNA. Genome Res 10(4):516–522

Sanger F, Nicklen S, Coulson A (1977) DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci USA 74:5463–5467

Schena M, Shalon D, Davis RW et al (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science 270(5235):467–470

Schiex T, Moisan A, Rouzé P (2001) Eugène: an eukaryotic gene finder that combines several sources of evidence. In: Computational Biology, selected papers from JOBIM′2000 number 2066 in LNCS, Springer Verlag, New York, pp. 111–125.

Schneiker S, Martins dos Santos VA, Bartels D et al (2006) Genome sequence of the ubiquitous hydrocarbon-degrading marine bacterium *Alcanivorax borkumensis*. Nat Biotechnol 24(8):997–1004

Schneiker S, Perlova O, Kaiser O et al (2007) Complete genome sequence of the myxobacterium Sorangium cellulosum. Nat Biotechnol 25(11):1281–1289

Shendure J, Mitra RD, Varma C et al (2004) Advanced sequencing technologies: methods and goals. Nat Rev Genet 5(5):335–344

Shendure J, Porreca GJ, Reppas NB et al (2005) Accurate multiplex polony sequencing of an evolved bacterial genome. Science 309(5741):1728–1732

Shendure JA, Porreca GJ, Church GM (2008) Overview of DNA sequencing strategies. Curr Protoc Mol Biol Chapter 7: Unit 7:1

Skovgaard M, Jensen LJ, Brunak S et al (2001) On the total number of genes and their length distribution in complete microbial genomes. Trends Genet 17(8):425–428

Slater GS, Birney E (2005) Automated generation of heuristics for biological sequence comparison. BMC Bioinformatics 6:31

Smith MW, Feng DF, Doolittle RF (1992) Evolution by acquisition: the case for horizontal gene transfers. Trends Biochem Sci 17(12):489–493

Smith TF, Waterman MS (1981) Identification of common molecular subsequences. J Mol Biol 147(1):195–197

Sonnhammer EL, von Heijne G, Krogh A (1998) A hidden Markov model for predicting transmembrane helices in protein sequences. Proc Int Conf Intell Syst Mol Biol 6: 175–182

Spellman PT, Miller M, Stewart J et al (2002) Design and implementation of microarray gene expression markup language (MAGE-ML). Genome Biol 3(9): RESEARCH0046.

Stanke M, Waack S (2003) Gene prediction with a hidden Markov model and a new intron submodel. Bioinformatics 19(Suppl 2):ii215–ii225

Stanke M, Tzvetkova A, Morgenstern B (2006) AUGUSTUS at EGASP: using EST, protein and genomic alignments for improved gene prediction in the human genome. Genome Biol 7(Suppl 1):S11–S18

Sturn A, Quackenbush J, Trajanoski Z (2002) Genesis: cluster analysis of microarray data. Bioinformatics 18(1):207–208

Sugawara H, Ogasawara O, Okubo K et al (2008) DDBJ with new system and face. Nucleic Acids Res 36:D22–D24

Suzek BE, Ermolaeva MD, Schreiber M et al (2001) A probabilistic method for identifying start codons in bacterial genomes. Bioinformatics 17(12):1123–1130

Tamames J, Casari G, Ouzounis C et al (1997) Conserved clusters of functionally related genes in two bacterial genomes. Mol Evol 44:66–73

Tatsuov RL, Mushegian AR, Bork P et al (1996) Metabolism and evolution of *Haemophilus influenza* deduced from a whole-genome comparison with *Escherichia coli*. Curr Biol 6(3):279–291

Tatusov RL, Fedorova ND, Jackson JD et al (2003) The COG database: an updated version includes eukaryotes. BMC Bioinformatics 4:41

Team RDC (2008) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Tech M, Meinicke P (2006) An unsupervised classification scheme for improving predictions of prokaryotic TIS. BMC Bioinformatics 7:121

Thieme F, Koebnik R, Bekel T et al (2005) Insights into genome plasticity and pathogenicity of the plant pathogenic bacterium *Xanthomonas campestris* pv. vesicatoria revealed by the complete genome sequence. J Bacteriol 187(21):7254–7266

Tusher VG, Tibshirani R, Chu G (2001) Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci U S A 98(9):5116–5121

Usuka J, Zhu W, Brendel V (2000) Optimal spliced alignment of homologous cDNA to a genomic DNA template. Bioinformatics 16(3):203–211

van Baren MJ, Brent MR (2006) Iterative gene prediction and pseudogene removal improves genome annotation. Genome Res 16(5):678–685

Vapnik VN (1999) The nature of statistical learning theory. Springer, New York.

Velculescu VE, Zhang L, Vogelstein B et al (1995) Serial analysis of gene expression. Science 270(5235):484–487

von Mering C, Jensen LJ, Snel B et al (2005) STRING: known and predicted protein-protein associations, integrated and transferred across organisms. Nucleic Acids Res 33:433–437

Vorhölter FJ, Schneiker S, Goesmann A et al (2008) The genome of Xanthomonas campestris pv. campestris B100 and its use for the reconstruction of metabolic pathways involved in xanthan biosynthesis. J Biotechnol 134(1–2):33–45

Wei C, Brent MR (2006) Using ESTs to improve the accuracy of de novo gene prediction. BMC Bioinformatics 7:327

Wilkinson MD, Links M (2002) BioMOBY: an open source biological web services proposal. Brief Bioinform 3(4):331–341

Wu J, Mao X, Cai T et al (2006) KOBAS server: a web-based platform for automated annotation and pathway identification. Nucleic Acids Res 34:W720–W724

Wu TD, Watanabe CK (2005) GMAP: a genomic mapping and alignment program for mRNA and EST sequences. Bioinformatics 21(9):1859–1875

Wu W, Xing EP, Myers C et al (2005) Evaluation of normalization methods for cDNA microarray data by k-NN classification. BMC Bioinformatics 6:191

Yang YH, Speed T (2002) Design issues for cDNA microarray experiments. Nat Rev Genet 3(8):579–588

Yauk C, Berndt L, Williams A et al (2005) Automation of cDNA microarray hybridization and washing yields improved data quality. J Biochem Biophys Methods 64(1):69–75

Yauk CL, Berndt ML, Williams A et al (2004) Comprehensive comparison of six microarray technologies. Nucleic Acids Res 32(15):e124

Zhang MQ (2002) Computational prediction of eukaryotic protein-coding genes. Nat Rev Genet 3(9):698–709

Zhang Z, Schwartz S, Wagner L et al (2000) A greedy algorithm for aligning DNA sequences. J Comput Biol 7(1–2):203–214

Glossary

2R hypothesis hypothesis suggesting that the genomes of the early vertebrate lineage underwent two rounds of whole-genome duplications, reflected e.g. by the existence of 4 remnants of Hox clusters.

Accession number An identifier supplied by the curators of the major biological databases upon submission of a novel entry (usually sequence data) that uniquely identifies that database entry.

AFLP This technique produces DNA fragments that are separated by size (length of sequence) using polyacrylamide gel electrophoresis or capillary sequencers. Bands at a particular site on the gel (equivalent to alleles) are counted as being present or absent for the analysis.

Algae all photosynthetic eukaryotes other than land plants.

Algal bloom rapid multiplication of one or a small number of algal species. In some cases the algae are toxic.

Alignment overview of both differences and similarities between two or more DNA, RNA or protein sequences created by aligning residues that are putatively homologous. Gaps may be introduced to optimally align the sequences.

AMD Acid Mine Drainage.

API Application Programming Interface.

BAC bacterial artificial chromosome. a cloning vector carrying up to 300 kbp of insert sequence.

Barcoding A specific piece of DNA (usually part of the mitochondrial cytochrome c oxidase sub unit I gene) is sequenced many times from different species. There are sufficient interspecies differences in this gene, for the sequence of DNA to act as a unique code for a species and therefore be used as a taxonomic aid.

Bioinformatics The application of information technology to the field of molecular biology.

J.M. Cock et al. (eds.), *Introduction to Marine Genomics*,
Advances in Marine Genomics 1, DOI 10.1007/978-90-481-8639-6,

BioJava, BioPerl Programming frameworks to handle the most common file formats and tools that are used by bioinformaticians.

BLAST "Basic local alignment search tool", a well-known and established tool for finding homologue sequences in large sequence databases.

CAMERA Community Cyberinfrastructure for Advanced Marine Microbial Ecology Research and Analysis.

Clone the term "clone" can refer to a bacterium carrying a piece of cloned DNA, or to the cloned DNA itself.

Clustering EST sequences are clustered according to their sequence, so that they can be assembled afterwards. The set of all Clusters form a Clusterset.

COG/KOG The COG (Cluster of Orthologous Groups of proteins) database consists of 138,458 proteins, which form 4873 COGs and comprise 75% of the 185,505 (predicted) proteins encoded in the 66 genomes of unicellular organisms that were used for the database. The eukaryotic orthologous groups (KOGs) include proteins from 7 eukaryotic genomes.

Cohort Fish in a stock born in the same year.

Contig The term "contig" comes from a shortening of the word "contiguous". It can be used to refer to the final product of a shotgun sequencing project. When individual lanes of sequence information are assembled to infer the sequence of the larger DNA piece, the product consensus sequence is called a "contig".

Cryptic species A cryptic species is a species that is reproductively isolated (or at least genetically distinct) from a second species but which resembles the second species so closely that both have traditionally been considered a single species.

ddNTP dideoxy-nucleotide.

DGGE Denaturing gradient gel electrophoresis. Molecular fingerprinting technique that separates PCR products based on sequence differences that result in distinctly different denaturing characteristics of the DNA.

dNTP deoxy-nucleotide.

Dye A chemical compound used to label biological material such as cDNA allowing it to be detected. Often fluorescent compounds are used.

EGT Environmental Gene Tags.

ELISA Enzyme Linked Immuno Sorbent Assay; Biochemical test to detect and quantify an antigen in a sample using a specific antibody.

Emulsion PCR a technique for generating a clonally amplified piece of DNA in vitro along with a microbead, within a mineral oil emulsion.

Endosymbiosis An endosymbiont is an organism that lives within the body or cells of another organism. Endosymbiosis played an important in the evolution

of the eukaryotes as mitochondria and plastids were derived from endosymbiotic organisms captured by ancestral eukaryotic cells.

Epitope Region of a protein or peptide recognised by an antibody.

eQTL expression Quantitative Trait Loci; the transcriptome is associated with thousands of expression traits (see also QTL).

EST Expressed Sequence Tag. For a gene expression analysis, it is possible to extract the mRNA from a cell, translate it into cDNA and sequence the cDNA from either one or both ends. The reads are called ESTs. EST data may contain sequencing errors.

FISH Fluorescence in situ hybridisation.

Fosmid A cloning system based on the E. coli F factor. These clones have an average insert size of 40 Kbp, with a very small standard deviation.

Gametophyte In plants and algae that undergo alternation of generations, a gametophyte is the multicellular generation that produces gametes.

Gene chip Thousands of sequences are individually attached (spotted) onto a small glass slide (typically the size of a microscope slide). Each spot on the slide represents a gene sequence. These can be screened to generate expression profiles of a cell, tissue or organism(s) under different conditions. Also called a microarray.

Gene knockdown Refers to techniques such as RNAinterference by which the expression of a gene is reduced.

Gene library Collection of pieces of DNA or cDNA cloned into artificial vectors, which can be replicated, sequenced and screened.

Gene repertoire the set of genes (or gene families) encoded in the genome of an organism.

Genetic hitch-hiking A process by which an allele or mutation may spread through the gene pool because it is closely linked to a gene that is being selected for.

Genome All the genetic material of an organism.

Genomics the study of the structure and function of genomes (see genome).

GOS Global Ocean Sampling.

GSC Genomic Standards Consortium.

GUI Graphical user interface. A common means of interaction between humans and computers, which uses graphical representations of data and offers user interaction via pointing devices (e.g. a mouse).

HMMs Hidden Markov Models (HMMs) allow the integration of diverse sequence features into a coherent, probabilistic framework. For the task of gene identification,

HMMs may include states modeling introns, exons, intergenic regions, start and stop codons, splice signals, polyadenyation signals and ribosomal binding sites.

Homeodomain proteins a class of proteins characterized by the existence of a specific motif, the homeodomain, which is primarily involved in DNA binding.

Homologue In order to describe the evolutionary relation of proteins, the terms homology, orthology, and paralogy are used. In this context, homology means that two proteins or sequences share a common ancestor. Two principal types of homology can be distinguished (Orthology and Paralogy).

Horizontal (or lateral) gene transfer A process in which genes from one organism are integrated into the genome of a second organism (which may be very distantly related to the first) and subsequently inherited with the rest of the genetic material of the cell.

Hox cluster genomic array of Hox genes, which constitute a subgroup of homeodomain proteins. Spatial Hox gene expression during development tends to correlate with the spatial arrangements of genes in the genomes of a broad panel of animals. Hence, Hox clusters also serve as a paradigm to understand the commonalities and changes during the evolution of animal body plans.

HSP high-scoring segment pair.

IMG The Integrated Microbial Genomes system.

INSDC International Nucleotide Sequence Database Collaboration. It consists of the DDBJ (Japan), EMBL (Europe) and GenBank (USA) Nucleotide Sequence Database. The three databases exchange new and updated data on a daily basis to achieve optimal synchronization.

IUPAC International Union of Pure and Applied Chemistry. This non-governmental organisation has developed a system of naming chemical elements and their compounds such as amino acids or nucleotides.

JCoast Comparative Analysis and Search Tool.

JGI Joint Genome Institute, a sequencing centre in the USA.

KEGG The KEGG (Kyoto Encyclopedia of Genes and Genomes) is a bioinformatics resource for linking genomes to life and the environment.

Kb or Kbp kilobase or kilobase pair, region of DNA 1000 nucleotides long.

Linkage map Genetic map produced using recombination values of genes or other markers to identify the linear order and relative distance of these markers on a chromosome.

LPS Lipopolysaccharides; large molecules containing a lipid and a polysaccharide linked by a covalent bond; major components of the outer membrane of Gram-negative bacteria; induce strong immune responses in animals.

Local alignment Sequence alignment which focuses on matching part of one sequence with part of another.

MAGE-TAB A data format used to transfer microarray data to public repositories (e.g. ArrayExpress). The MAGE-TAB format consists mainly of spread sheets, which can be represented as tabulator- or comma-separated files.

MAS Marker-Assisted Selection; Selection of a genetic determinant of a trait of interest (e.g. productivity, disease resistance and quality) through the use of a marker (morphological, biochemical or one based on DNA/RNA variation).

MA-plot A certain kind of scatter-plot commonly used for microarray analysis where the intensity measurements from microarrays are transformed in a special way. The x-axis corresponds to the absolute intensity of the spot (A-value), the y-axis corresponds to a measure of differential expression (M-value).

Match position with identical bases or amino acids in a sequence alignment.

Mate-pair paired reads of the two ends of a cloned DNA molecule. Mate-pairs are often used by genome assembly programs to orient and order contigs, taking into account the distance between the ends of the DNA molecule and the orientation relationship of the sequences.

MHC Major Histocompatibility Complex; A large genomic region playing an important role in the immune system, autoimmunity and reproductive success.

MDA Multiple displacement amplification. Isothermal DNA amplification method using random hexamers and the DNA polymerase of bacteriophage $\phi 29$.

Metabolomics High throughput analysis of the metabolites present in a cell type, a tissue or an organism and the modifications to metabolite pools under different conditions.

Metagenomics (or environmental sequencing) Application of genomic analysis, particularly high-throughput sequencing to environmental samples such as uncultured microbial communities.

Metatranscriptomics Global analysis, usually by sequencing, of the expressed genetic information (gene transcripts) produced by the collection of organisms in an ecosystem.

MIAME Minimal Information About a Microarray Experiment. MIAME is a recommendation that was established by a joint group of microarray experts (the Microarray and Gene Expression Data society or MGED) and describes the minimal requirements for data and information thata submission to a public microarray database should contain. The aim is to make microarray data understandable and reproducible.

Microsatellite Small DNA stretches of a repeated core sequence of few base pairs (e.g. GT repeat units) that are highly polymorphic, particularly in terms of the number of repeated units. These are mainly used to create genetic maps.

MIMS Minimum Information about a Metagenome Sequence.

Mismatch position with different bases or amino acids in a sequence alignment.

Morpholino Synthetic molecules of usually 25 bases in length that bind to complementary sequences of RNA, blocking the access of cell components to those sequences. Morpholinos can block translation, splicing, miRNAs or their targets and ribozyme activity.

MPSS Massive Parallel Signature Sequencing; Tool to analyse the level of expression of virtually all genes in a sample by counting the number of individual mRNA molecules produced from each gene.

Northern blot A method for analysing RNA molecules involving separation on an agarose gel and detection of specific RNAs using radioactive or fluorescently labelled probes (also called an RNA gel blot).

Ontology In the context of computer and information sciences, an ontology defines a set of representational primitives with which to model a domain of knowledge or discourse. The Gene Ontology for example provides a controlled vocabulary to describe gene and gene product attributes in any organism.

ORF Open Reading Frame, a series of codons beginning with a start codon and ending with a stop codon, without any internal stop codons.

Organelle A differentiated structure within a cell, such as a chloroplast, mitochondrion or vacuole, which performs a specific function.

Orthologue Orthologous sequences are homologous sequences that have been separated by a speciation event. A gene that exists in a particular species that then diverges into two separate species, will be present in two copies afterwards. These copies are called orthologues and will typically have the same or a similar function.

PAMP Pathogen-Associated Molecular Patterns; Small molecular motifs conserved within a class of microbes; LPS (see above) is considered to be the prototypical PAMP.

Pangenome Unique set of proteins for a given species. In bacteria, the pangenome can be twice the size of the genomes of individual members of a species.

Paralogue Homologous sequences that have been separated by a gene duplication event in an ancestral genome are called paralogous. Paralogous genes may mutate and acquire new functions because the original selective pressure is reduced.

PCR Polymerase chain reaction. A technique for replicating a specific piece of DNA in-vitro. Oligonucleotide primers are added (which initiate the copying of each strand) along with nucleotides and Taq polymerase. By cycling the temperature, the

target DNA is repeatedly denatured and copied allowing it to be amplified in an exponential manner.

PDB Protein Data Bank. A single worldwide repository for the processing and distribution of 3-D biological macromolecular structure data.

Phytoplankton The photosynthetic organisms present in the plankton.

Picoeukaryotes microscopic eukaryote species with a cell size of less than 3 μm.

Primary endosymbiosis Used to describe the initial capture of a cyanobacterium by a eukaryotic cell that gave rise to the algae (see also endosymbiosis).

Primer Small oligonucleotide (anywhere from 6 to 50 nucleotides long) used to prime DNA synthesis.

Probe/Reporter An oligonucleotide attached to the surface of a microarray.

Proteomics High-throughput analysis of the proteins present in a cell type, a tissue or an organism.

Pseudogene A gene that is no longer functional either because its protein-coding sequence is disrupted or because it is no longer expressed.

PTM Post-transcriptional modification. Alterations made to pre-mRNA before it leaves the nucleus and becomes mature mRNA.

p-value The result of a statistical test can be summarised using a p-value as a measure of significance. It can be described as the confidence in the rejection of the null-hypothesis. In the case of microarrays, p-values are often used to assess whether a gene is differentially expressed.

Pyrosequencing Massively parallel DNA sequencing technique without the requirement of a prior cloning of the DNA.

Q-PCR, qPCR or qRT-PCR Quantitative real-time reverse-transcription PCR. A PCR-based assay, which involves the direct measurment of the incorportation of a fluorescent dye into the PCR product. The level of fluorescence is a direct reflection on the amount of prduct present. By comparing the point in the PCR reaction where the reaction enters the log phase of replication for a control and a treated sample and determining the difference between the two, a measure of the change in relative gene expression caused by the treatment is determined.

QTL Quantitative Trait Loci; A region of a chromosome that is associated with a particular measurable trait (see also eQTL).

RAPD Random Amplification of Polymorphic DNA; DNA fragments randomly amplified by PCR from genomic DNA with short primers (8–12 nucleotides) of arbitrary nucleotide sequence.

RAST Rapid Annotation using Subsystem Technology.

Replicate A repetition of the same experimental measurement. The replicate can either be technical, where the same biological extract is analyzed twice, e.g. by using the same mRNA for multiple microarrays; or it can be a biological replicate, when for example sample material is harvested from different individuals grown under the same conditions.

Reporter/Probe An oligonucleotide attached to the surface of a microarray.

RNA interference (RNAi) a genetic mechanism in which double-stranded RNA is cleaved into small fragments and acts a signal to either initiate the degradation of a complementary messenger RNA or to interfere with its translation.

ROS Reactive Oxygen species; Small molecules or ions formed by the incomplete one-electron reduction of oxygen; ROS contribute to oxidative stress/damage of DNA, lipids (oxidations of polydesaturated fatty acids), proteins (oxidations of amino acids) and inactivate specific enzymes (oxidation of co-factors).

SAGE Serial Analysis of Gene Expression. A high-throughput method to measure gene-expression by sequencing. Short fragments (tags) are generated from cDNA, which are then ligated to form long concatenated sequences, and sequenced.

Seaweed Brown, red or green macroalgae.

Scatter-plot A two-dimensional plot where each measurement value is depicted by a single dot. Scatter-plots are often used to inspect data distributions.

SDS-PAGE Sodium dodecylsulphate – polyacrylamide gel electrophoresis; method used to separate proteins involving electrophoresis on an acrylamide gel.

Secondary endosymbiosis An event in which a eukaryotic cell enslaves another eukaryotic cell that possesses a plastid derived from a primary endosymbiosis (see also endosymbiosis and primary endosymbiosis).

Sequence trace file Raw sequence data in the form of a chromatogram from a Sanger sequencing reaction. The NCBI Trace Archive (http://www.ncbi.nlm.nih.gov.gate1.inist.fr/Traces/home/) and the Ensembl Trace Server (http://trace.ensembl.org/) are a public repositories for this sort of data.

Short Read Archive (SRA) An NCBI database that stores raw data from sequencing platforms such as the Roche 454 System.

Singleton A single EST sequence that does not cluster with other ESTs.

SNP Single-nucleotide polymorphism; particular sites (base pairs) in a sequence that are polymorphic and can be used as markers.

SOM Self Organizing Maps.

Sporophyte A spore-producing plant.

Spot/feature A small area on the surface of a microarray containing a defined probe. Features can be produced using a spotter or by direct synthesis of oligonucleotides on the substrate.

SSR Simple Sequence Repeat; also known as microsatellite (see definition above).

SteN Statistical electronic Northern blot.

SVM Support Vector Machine, high performance machine learning technique that has been used to improve classification accuracy in biological applications such as gene prediction, detection of protein family members, RNA and DNA binding proteins and the functional classification of gene expression data. SVMs can solve non-linear classification problems by learning an optimally separating hyperplane in a higher-dimensional feature space. By use of non-linear kernel functions such as a Gaussian kernel, complex and non-linear decision functions can be realised.

Synteny evolutionary preservation of the order of genes in a genomic region that indicates that this order also reflects the arrangement of genes in the last common ancestor of two compared species. Subdivided into macro-synteny (conserved aspects on the chromosomal level) and micro-synteny (local, gene-by-gene synteny).

Target In a microarray experiment, the target is labelled population of molecules corresponding to the RNA or DNA extracted from mixtures of cells, tissues, cell cultures or other samples. This labelled population of molecules is often referred to erroneously as the probe.

TC Tentative Consensus Sequence. EST sequences can be clustered together using assembly programs to form TCs.

Tertiary endosymbiosis An event in which a eukaryotic cell enslaves another eukaryotic cell that possesses a plastid derived from a secondary endosymbiosis (see secondary endosymbiosis). This has occurred several times in the dinoflagellates, and the captured plastid is thought to have replaced a pre-existing secondary plastid (see also endosymbiosis).

TETRA A tool that can be used to calculate, how well tetranucleotide usage patterns in DNA sequences correlate. Such correlations can provide valuable hints about the relatedness of DNA sequences.

TGICL TIGR Gene Index Clustering tools.

Third Party annotation (TPA) sequence database A database designed to capture experimental or inferential results that support submitter-provided annotation for sequence data that the submitter did not directly determine but can be derived from DDBJ/EMBL/GenBank primary data.

Transcriptomics High-throughput analysis of the mRNA transcripts present in a cell type, a tissue or an organism.

Transcriptome the full complement of transcripts in a given species.

UTR untranslated region of a transcript. The complementary regions to the ORF (see ORF).

Web services Provide a standardized method to access information or perform computations over a network via the exchange of XML-based messages.Western blot?

Western blot A method for analysing proteins involving separation on by SDS-PAGE (see above) and detection of specific proteins using antibodies.

XML Extended Markup Language, a specification to create a common representation of different types of data.

Index