Pierre Pontarotti *Editor*

Evolutionary Biology

Convergent Evolution, Evolution of Complex Traits, Concepts and Methods

Springer

Evolutionary Biology

Pierre Pontarotti
Editor

Evolutionary Biology

Convergent Evolution, Evolution of Complex Traits, Concepts and Methods

Springer

Editor
Pierre Pontarotti
CNRS, Laboratoire Evolution Biologique et
Modélisation
Université d'Aix-Marseille
Marseille
France

ISBN 978-3-319-41323-5 ISBN 978-3-319-41324-2 (eBook)
DOI 10.1007/978-3-319-41324-2

Library of Congress Control Number: 2016942867

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG Switzerland

Preface

For the ninth year we publish a book on evolutionary biology concept and application we try to catch the evolution and progress of this field for this goal we are really helped by the Evolutionary Biology Meeting at Marseilles. The goal of this annual meeting is to allow scientists of different disciplines, who share a deep interest in evolutionary biology concepts, knowledge and applications, to meet and exchange and enhance interdisciplinary collaborations. The Evolutionary Biology Meeting at Marseilles is now recognised internationally as an important exchange platform and a booster for the use of evolutionary-based approaches in biology and also in other scientific areas.

The book chapter have been selected from the meeting presentations and from a proposition born by the interaction of meeting participants.

The reader of the evolutionary biology books as well as the meeting participants would, maybe like me, witness years after years during the different meetings and book editions a shift on the evolutionary biology concept. The fact that the chapters of the book are selected from a meeting enables the quick diffusion of the novelties.

I would like to underline that the nine books are complementary one to another and should be considered as tomes.

The articles are organised in the following categories
Convergent evolution (Chaps. 1–8)
Evolution of complex traits (Chaps. 9–14)
Concepts (Chaps. 15–19)
Methods (Chaps. 20–22)

Marseille, France
April 2016

Pierre Pontarotti

Acknowledgement

I would like to thank all the authors, the meeting participants, the sponsors of the meeting: Aix Marseille Université, CNRS, ITMO, ECCOREV FEDERATION, Conseil Départemental 13, ITMO, Ville de Marseilles.

I wish to thank the team of the non profit organisation: Association pour l'Etude de l'Evolution Biologique (AEEB) for the organisation of the meeting.

I also wish to thank the Springer's edition staff and in particular Andrea Schlitzberger for her competence and help.

I am also thankful to the director of the AEEB Marie-Hélène Rome for the coordination of the meeting and the book.

Marseilles France
April 2016

Pierre Pontarotti

Contents

Part I
Convergent Evolution

Chapter 1
Road Map to Study Convergent Evolution: A Proposition for Evolutionary Systems Biology Approaches

Pierre Pontarotti and Isabelle Hue

Abstract Every evolutionary biologist will surely acknowledge that convergent evolution, the independent evolution of similar features in different evolutionary lineages, is an important phenomenon of the organic evolution. However, the concept is complex and can have several related but conceptually distinct meanings in the literature, including parallel evolution (independent mutations in orthologous genes) and homoplasy (any convergent traits, including reversion). Some authors, for example, use the term "parallel evolution" differently from "convergent evolution", and they reserve the latter term for more "unlikely" (or "more independent") examples of phenotypic similarity across lineages, those not predisposed by genomic similarity. Semantic arguments in science are often fruitful, but can also prevent efficient scientific exchanges in the field. Hence, the goal of this article was (1) to define convergent evolution in a better way by applying a multilevel biological-level approach and (2) to propose a road map to help researchers navigate their routes in studying this phenomenon.

State of the art including new concepts needed to better understand this phenomenon

In order to add depth and (we hope) some clarity to the conceptual understanding of evolutionary convergence, it is important to describe the following aspects: at the phenotype level, to distinguish isoconvergent/alloconvergent evolution (see text below), species global convergence versus character convergence (multivariate convergence vs single trait convergence) and identify what is the type of character that has evolved, whether it is morphology versus physiology, for instance, and whether it is a continuous or discrete character.

P. Pontarotti (✉)
CNRS Centrale Marseille, I2M UMR 7373 équipe EBM (Evolution Biologique Modélisation), Aix Marseille Université, 13453 Marseille cedex, France
e-mail: pierre.pontarotti@univ-amu.fr

I. Hue
UMR BDR INRA ENVA, Université Paris Saclay, 78350 Jouy en Josas, France

P. Pontarotti (ed.), *Evolutionary Biology*, DOI 10.1007/978-3-319-41324-2_1

We argue here that it is essential to first characterize phenotype-level convergence, before defining convergence at other levels. Indeed, once the phenomenon at the phenotypic level is clearly described, one can define the phenomenon at other biological levels in cases where the different biological levels are linked, and address the question about the origins of the genetic changes de novo versus the pre-existing ones and the consequences of the genetic variation at the coding, epigenetic, transcriptional and higher biological level.

1.1 Concept, Definition: Convergence at Phenotypic Level

1A) **Iso- versus Alloconvergent evolution**

We start this section with a short epistemologic analysis about the use of the term *parallel and convergent evolution* in the literature. The term "convergent evolution" is generally used in the following way: «Convergent evolution: the independent evolution of similar features in different evolutionary lineages» (Losos 2011). However, the convergent evolution of a given character can result from the evolution of a similar or a different character.

The distinction, using the ancestral state information, is found in some of the molecular phylogeny analyses, under the term of parallel and convergent evolution (Zhang and Kumar 1997). Here, the term "parallel evolution" is used if the ancestral amino acid is the same, and "convergent evolution" is used if the ancestral amino acid is different.

However, most of the time, when convergent evolution is found at the phenotype level, the distinction between parallel and convergent evolution is not based on the evolutionary history of the characters, but on the similarity of the genetic mechanisms that are involved in the repeated phenotype. If the molecular mechanisms are the same, the evolution is said to be parallel; if the genetic mechanisms are different, the evolution is said to be convergent (see, for example, Rosenblum et al. 2014). Some authors use also the phylogenetic proximity; when the species are phylogenetically close, the term "parallel evolution" is used; when they are phylogenetically distant, they use the term "convergent evolution". This makes sense, regarding the above use of both terms, since closest species will tend to evolve using the same mechanisms, either because the variation that underlines the phenotype variation corresponds to the standing variation (Elmer and Meyer 2011), or because of the interaction between the different genes of the genome (or epistasis). With a similar background, more similar fixation of substitutions are excepted but if the interactions are different between the genes, it is likely that the evolutionary trajectory will be different (for a great example, see Ujvari et al. 2015).

In order to help in the conceptual understanding of evolutionary convergence, we propose two neologisms that can be applied to all the biological levels: iso-convergent and alloconvergent evolution (iso from the same ancestral state and allo from a different ancestral state). This distinction is not done at the phenotype level

in the literature (Conway Morris 2003; McGee 2011; Losos 2011; Gordon et al. 2015). However, it has to be noted that scientists who work at the genetic level, when talking about convergent evolution at the phenotype level, intuitively mean isoconvergence at the phenotype level (see Fig. 1.1 in Martin and Orgozozo 2013; Rosenblum 2014). As an illustration, this is what Rosenblum et al. wrote (2014) in a chapter called Linking phenotype and phylogeny: "*Before investigating the molecular mechanisms of convergence, researchers must first ensure that the phenotype of interest is convergent. Researchers should define convergent evolution in a phylogenetic context. This requires an explicit integration of phenotypic data with a molecular phylogeny, and should incorporate uncertainty in the phylogeny and the model. For example, the independent evolution of similar phenotypes can be identified from ancestral state reconstruction, comparisons of phylogenetic and genetic distance, or inferred shared selective regime* (e.g. Muschick et al. 2012; Ingram and Mahler 2013)". Here, Rosenblum et al. used the word "convergent evolution" of a phenotype, but they mean, in our terms, isoconvergent evolution of a phenotype.

Thus, either the distinction is not noted, or the authors (mainly the authors interested by the genetic mechanisms) when discussing about convergent evolution think about "isoconvergent evolution" implicitly. In our view, it is, however, important to explicitly make the distinction between iso- and alloconvergence at the phenotype level as, above all, this will allow testing further the relationship between isoconvergent evolution at the phenotypic level and isoconvergence at other biological levels (Stern and Orgozozo 2008). If this is the case, isoconvergent evolution of a phenotypic level can be used to decipher biological mechanisms at the genetic, epigenetic, transcriptional and other biological levels (Kopp 2009).

A key example: biochemical/enzymatical function

Alloconvergent evolution of enzyme function can be seen in two distinct, but sometimes joint, effects. The first is when non-homologous enzymes deliver the same transformation, as expressed by the same four-digit enzyme commission (EC) number. These enzymes are named transformational analogues. The second situation is when the same (same four-digit EC) or related (same three-digit EC) enzyme transformation is effected by a similar disposition of residues in the active site, as exemplified in the Ser-His-Asp catalytic triad shared by the trypsin family and subtilisin. Such enzymes are mechanistic analogues. These two situations are not exclusive because two enzymes are assigned to both classes if they perform exactly the same overall reaction with the same mechanism (Doolittle 1994; Gherardini et al. 2007).

The transformational analogues evolved at the protein level from different ancestors via divergent evolution as the transformational function evolved likely from different ancestral transformational functions, and these functions evolved therefore in an alloconvergent manner (see, for an example, the case of b lactamase B, Alderson et al. 2014).

In the case of enzymatic isoconvergent evolution, the same function is present in the ancestor (see, for example, Dick et al. 2012).

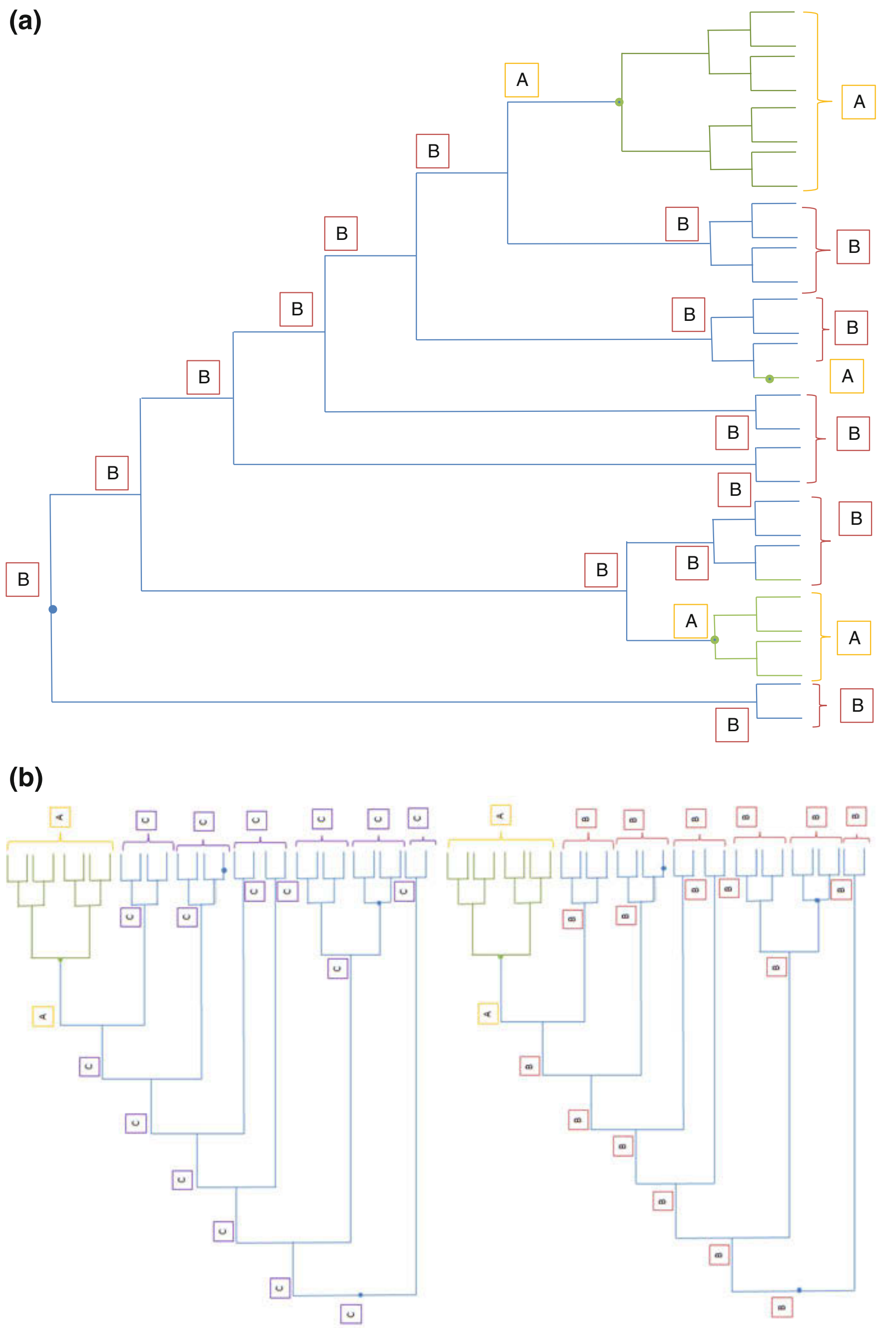


Fig. 1.1 Isoconvergence versus alloconvergence. Legend: A: Derived state; B and C: Ancestral state

1B) **Assumption of "convergent evolution" via adaptation to a new environment**

Another aspect of the definition of convergent evolution is that some authors have no assumption concerning the evolutionary driving force, while others see the action of the adaptive evolution. Convergent evolution, defined as being the result of adaptive evolution, is commonly used in the literature as synthesized by Stayton (2015). A scientific approach should be first to evidence convergent evolution and then test for adaptation or any other phenomenon such as constraint or evolutionary relaxation (Arbuckle et al. 2014). Indeed, in many articles, the convergence at the phenotype level is suspected by the fact that the species that live in the same environment will be adapted to it in the same manner and will have the same phenotypic response, and the case of the marine mammals is one of the examples (Foote et al. 2015; McGowen et al. 2014; Mirceta et al. 2013). Furthermore, many examples are found in the case of "parallel speciation" for species that are close phylogenetically (this is why the term *parallel* is used for different populations of the same species; Jones et al. 2012; Elmer and Meyer 2011; Soria-Carrasco et al. 2014). It should be noted that in these species the genetic origin of the convergence could be due to the standing variation or hybridization.

1C) **Convergence at the character or the organism levels**

The definition of convergence can be based on specific traits (see, for example, Pankey et al. 2014) or based on the total organism: global similarities based on several characters (see, for example, Losos et al. 2011; Malher et al. 2013).

1.2 Road Map

1.2.1 Re-analyses of Known Cases and Analyses of New Cases of Convergent Evolution at the Phenotypic and Environmental Level

Concerning these three points, our proposition is to re-analyse the described and new cases on a phylogenetic basis and to extract on one side the cases corresponding to isoconvergent evolution and on the other those corresponding to alloconvergence.

1.2.1.1 Bibliographic Analysis and Research of Undescribed Cases

Research of known cases

This is the strategy we propose for the research of known cases. Research can be done on the ISI database (http://www.webofknowledge.com/) using convergent* AND evolution* keywords. In order to complete this search, several books can be

used: the McGhee (2011), the Conway Morris (2003) and the Sanderson and Hufford (1996), but also the website dedicated to the convergent evolution and coordinated by Conway Morris, i.e. map of life at http://www.mapoflife.org/, the Wikipedia site at http://en.wikipedia.org/wiki/List_of_examples_of_convergent_evolution, as well as the compilation realized by Connie Barlow at www.thegreatstory.org/convergence.pdf. General reviews like that of Gordon and Notar (2015) will also be used.

Research of new cases
Following the strategy previously proposed by Hiller et al. (2012), one can perform a systematic search for all the characters of a phylogenetic group and find one or several characters that have a paraphylogenetic distribution; in other words, to map all the characters on a phylogenetic tree and decipher the character that evolves in an independent manner. A database on phenotypes in the context of the phylogeny already exists: the morphobank, and is available online at http://www.morphobank.org/ (O'Leary and Kaufman 2011).

Habitat shift convergence
The search can be done by screening the ISI database using the key words convergent* AND environment* or, in an indirect manner, convergent* evolution and adaptation*. We can perform analysis showing the paraphyletic environmental distribution for a given clade, for example the aquatic way of life of mammals. At this stage, it is important to define the environment; an international effort is on the way (Buttigieg et al. 2013).

1.2.1.2 Definition of the Phenotype

Functional level, morphological level, distinction between discrete and continuous characters. Character convergence/Global convergence.

We now consider the different levels of biological organization to which convergence concepts can be applied. At the physiological level, one can distinguish different sublevels: biochemical (metabolism), electrophysiological, mechanical and physical. It is important to add here another level: the one of the environmental demand and the possibility that different solutions can be found to face the same problem. One of the classical examples is predators that confront prey containing toxic substances and may either evolve resistance or avoid eating the part of the body that contains the toxin. It should also be noted that very different morphologies may produce similar functional capabilities. For example, labrid fish with many different jaw structures can produce the same suction force (Alfaro et al. 2005).

Further, a physiological character can be linked to several morphological traits. For example, talking about ultrasonic detection, the deformation of the cochlea is isoconvergent in the ultrasonic hearing lineage but bats have external ears that allow amplifying the signal which is not the case for the dolphins. A given function or physiology can also be the addition of several functions. The example of

echolocation is a good one as it corresponds in fact to the addition of several different functions, and it is important to dissect the functions and then look if each function (subfunction) is supported by one or several morphological characters.

Once the character is defined, we then have to classify the cases as isoconvergence or alloconvergence. In the case of alloconvergence, because the character under investigation occurred in an independent manner, it can be used for statistical analysis as an independent observation and link, for example, to other parameters such as environment. However, the link with the other biological levels cannot be done (see above). In the case of isoconvergent evolution, many tests can be done (see after); furthermore, the link with the other biological levels can be investigated, and this will be discussed in the following paragraph.

1.2.1.3 Isoconvergence Detection and Test

Methods have been developed for both discrete and continuous traits in the case of isoconvergent evolution with well-defined characters.

Discrete Characters, Character per Character at the Phenotype Level

Identifying isoconvergence starts by an ancestral state reconstruction of the isoconvergent trait. For example, this method has provided support for isoconvergent evolution of plumage coloration in *Icterus* orioles (Omland and Lanyon 2000) and the origin of photosphores in squids (Pankey et al. 2014). In such an analysis, the phenotype is reconstructed over the phylogeny, and independent origins are taken as an evidence of isoconvergence.

Continuous Trait at Phenotype Level

Most of the time this approach is used for multicharacter traits to test isoconvergent evolution at the global (species) level, but it can also be used for a simple character trait, biochemical activity, for example.

Detection of the isoconvergent evolution

A first approach has been developed by Muschick et al. (2012) that used a phenotype approach to test for convergence in cichlid fishes by considering that convergence should result in a pattern of reduced phenotypic differentiation when compared with phylogenetic distance. They calculated Euclidean distances between species for the morphological traits of interest and plotted them against the phylogenetic distances. They then used simulations to identify instances where phenotypic divergence was significantly lower than expected, based on phylogenetic distance.

A second approach was described by Ingram and Mahler (2013), which explicitly modelled trait evolution onto a phylogeny to identify convergent

evolution. Their method (called SURFACE) takes a continuous trait and fits Ornstein–Uhlenbeck models, with varying numbers of selective regimes and with shifts at varying points on the tree. Akaike's information criterion is then used to select the best fitting model. Convergence is identified by the independent adoption of the same similar shift at multiple points on the phylogeny. This method represents a technique to identify when convergence has occurred.

Test to evidence whether we have more isoconvergence than expected
In order to quantify the strength of isoconvergence, Arbuckle et al. (2014) developed an index (Wheatsheaf) that quantifies the strength of convergence as the ratio of the average pairwise distance between all the taxa in the dataset to the average pairwise distance between all putatively convergent taxa. This kind of analysis is done usually for a group of characters in order to test a global convergence (see Malher et al. 2013; Moen et al. 2016 and for a review Stayton 2015).

1.2.2 Isoconvergence: Link Between the Different Biological Levels: From the Genotype to the Phenotype

Because it is possible that isoconvergence at the phenotype level could be linked to similar mechanisms at other biological levels (multilevel isoconvergence), it is important to show here how the different biological levels are linked; this is described in the following paragraph (Fig. 1.2).

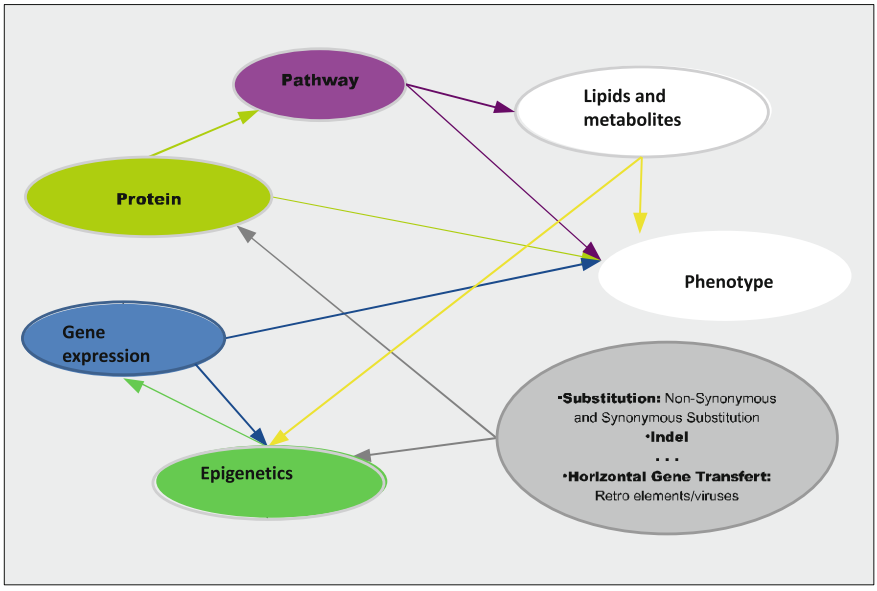


Fig. 1.2 The different biological levels to be analysed

1.2.2.1 Genetics/Epigenetics Mechanism Linking the Genotype to the Phenotype

The causative mutation for a phenotypic shift should be found at the DNA level: point mutation, indel deletion, horizontal gene transfer, at the coding level or non-coding level (repetitive element, non-coding RNA, micro RNA…). At the non-coding level, the consequence of a mutation can be at epigenetic and transcriptional level (Lynch et al. 2015). To better analyse the consequence of the mutation in the broader context of convergent evolution, we argue that it is important to have a network view instead of a gene-centred view (as genes do not work in isolation).

The most integrative way to understand the role of the different genes and corresponding networks of genes in development and in the relation "phenotype to genotype" is given by the work of Davidson and collaborators in seminal papers on gene regulatory network (GRN) and differentiation gene battery (DGB). The GRN shows a hierarchical organization. The GRNs establishing this initial postgastrular regulatory state, including the signalling interactions that help to establish domain boundaries, could be termed GRN level 1 (GRN1). The progenitor fields for the future adult body parts are later demarcated by signals plus local regulatory spatial information, and given regulatory states are established in each such field by the earliest body part-specific GRNs. Many such progenitor fields are thus set up during postgastrular embryogenesis. This corresponds to the second hierarchical level 2 (GRN2). Each progenitor field is then divided up into the subparts that will together constitute the body part, where the subdivisions are initially defined by installation of unique GRNs producing unique regulatory state GRN 3. The next-level termination of the developmental process in each region of the late embryo, the GRNs specifying the several individual cell types and deployed in each subpart of each body part, is the level 4. These GRNs control the expression of differentiation gene batteries, the final outputs of each cell type.

The evolution of forms or the apparition of a new structure is a developmental process that can be due to a neo-expression of the GRN levels 1, 2, 3. The apparition of a new cell type can be due to the neo-expression of GRN but at different levels 3 and 4 as well as from the neo-expression of the differentiation gene battery (due to the neo-expression of GRN or due to new promoter elements in cis position that could come from a retro-element, for example) (Lynch et al. 2015). As a consequence, the repeated variation should also be analysed at the transcriptional and epigenetic level (Pankey et al. 2014; Gallant et al. 2014; Pfenning et al. 2014).

Concerning physiological character (function), as said above, a new function could correspond to the set-up of complex organs. At the physiological level, we can distinguish several sublevels: biochemical (metabolism), electrophysiological, mechanical and physical. For the biochemical and the electrophysiological levels, the function will be supported by the DGB even though the transcriptional state of the involved genes may be regulated at an upper level in the GRN (see, for example, the electrogenesis case: Gallant et al. 2014; Zakon et al. 2006). In the case of a simple biochemical shift, this can be explained by a mutation on a DGB (Hiller

et al. 2012). This will be also the case for metabolic enzyme (Hiller et al. 2012) or for a gene coding a protein that gives a resistance to a poison or an antibiotic-resistant enzyme. Note also that a shift in the activity could be due to a difference at transcriptional or post-transcriptional level in the case of biochemical shift. It will thus be important to scan at the protein level but also at the transcriptional and epigenetic level.

1.2.2.2 Link Between the Different Biological Levels

It has to be noted that isoconvergence at one level could be correlated with either iso- or alloconvergence, or difference at the other biological levels; one of the paradigmatic examples is the flying capacity of bats and birds which came from an organ that allowed their common ancestor to walk. This is a clear case of isoconvergent evolution at the functional level, but in that case the organ involved in this function evolved via divergent evolution from the same ancestral character. Another example, the fly for butterfly, birds and mammals, is a case of alloconvergent evolution, since at the functional level the ancestral structure was not involved in the walk in the insect ancestor. The organ evolved via different arms in the case of birds/mammals, lateral lobes in the case of insects via divergent evolution. Another example can be found in the case of enzyme, the function of which evolved via alloconvergent evolution, with different ancestral functions and the same derived function. However, they evolved from a different structure in a divergent manner.

In the case of alloconvergent evolution, it is therefore difficult to make the link between structure and function. For example, in the case of enzyme, we cannot predict the involved amino acids at the sequence or biophysical levels. When the ancestor is different, many ways are possible to obtain the derived function, except if alloconvergent evolution is also present at the structural level. Therefore, the link between the evolution of the structure and the evolution of the function is not possible in that case.

In the case of enzymatic isoconvergent evolution, where the process starts likely from the same structure and the same function, it is plausible that the same change occurred at that protein level (amino acid sequence, biophysical properties). Several reports show that for receptors or enzymes: Mirceta et al. (2013), Ujvari et al. (2015). This phenomenon could be due to epistasis and pleiotropy (Martin and Orgogozo 2013).

In the case of isoconvergence at functional level, the link with the other level will be less evident, as seen above it is at the superior biological level than at the morphological one. So if isoconvergent evolution is present at the physiological level, the morphological level has to be checked (whenever possible) before proceeding to the other levels (genetic, epigenetic). If one has access only at the physiology, without knowledge of the supporting structure (e.g. adaptation to aquatic life), the relation with the other biological levels will be less evident. However, as described by many authors, the more similar the species, the higher the chance to have the same evolving mechanism (Conte et al. 2012).

1.2.2.3 Origin of the Genetic Variation: De Novo Versus Standing Variation

In the case of iso- or alloconvergent evolution at the nucleic acid and amino acid levels, the convergence could be due to independent mutations in the different lineages (de novo mutations) or due to standing variation or any other processes corresponding to horizontal gene transfer. This has been explained in the literature by Stern (2013), or Martin and Orgogozo (2013). The question is how to distinguish standing variation from de novo mutations. This point is important since, depending on the origin of the mutations, we cannot use the same model; in the case of de novo mutation, we need to use a model based on the phylogeny, whereas in the case of transfer or substitutions coming from the standing variation, different models are needed.

In general, the longer the time of separation between species where isoconvergence is found ancient, the weaker the chances to find pre-existing mutations between such species. Several mechanisms could explain why a pre-existing mutation will be shared across "distant" species: (i) incomplete lineage sorting due to rapid speciation, (ii) balancing selection, (iii) hybridization, (iv) paleo-hybridization in a group that separated but entered in contact again, followed by a new separation or any case of horizontal gene transfer (Bird et al. 2015). The different trait included at the genome level will not follow a bifurcative story, even if the paleo-hybridization event is old. In the case of paleo-hybridization, the conflicting topology will occur at the time the events occurred and a little bit after that, due to the allele sorting. For example, there is a probability that hybridization occurred at the base of the mammalian radiation, and this is why we may have conflicting histories between the mammalian lineages (Hallström and Janke 2010).

1.2.2.4 De Novo Isoconvergent Evolution from the Genotype to the Phenotype

Once the isoconvergent evolution has been identified at the phenotype level, the hypothesis of co-convergent evolution with the other levels can be tested (see Figs. 1.2 and 1.3). To do so, one needs to perform ancestral reconstruction and show that the apomorphy occurred in a convergent manner. This should be done for each biological level. In cases where the events at the different levels are concomitant (co-convergence), a possible cause-to-effect relationship across the levels will be evidenced (Fig. 1.2 and 1.3).

Direct link from genome to phenotype

Amino acid substitution that could cause a character shift (Fig. 1.3).

Zhang and Kumar (1997) as well as Foote et al. (2015) linked the convergent evolution at the amino acid level with the convergent evolution at the phenotypic level. The probable evolution of the amino acid is evaluated using a likelihood-based model. The next step is then to detect the shift from the plesiomorphic to the

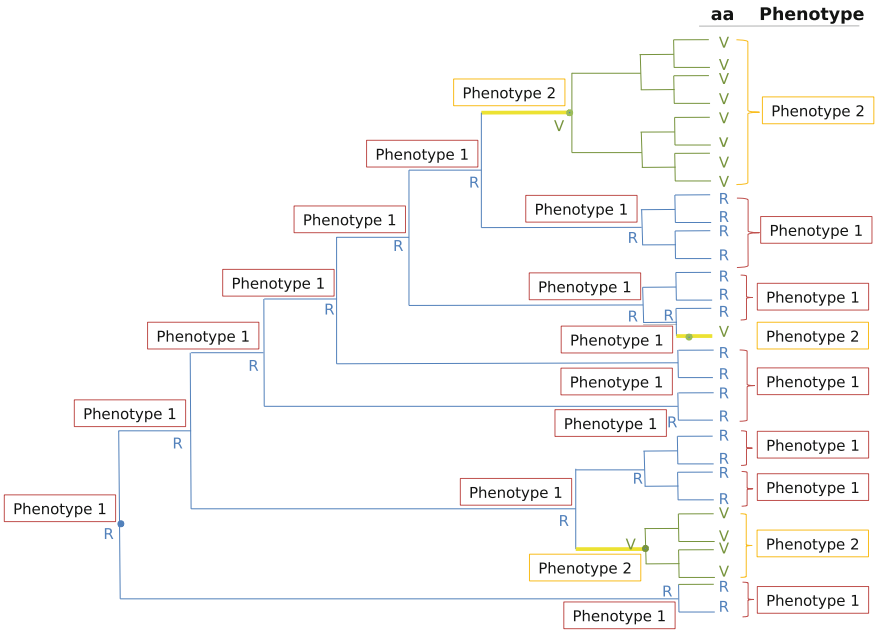


Fig. 1.3 Schematic example of how to link the phenotype to the other biological levels on a phylogenetic tree (here the link between a mutation in a protein and a given function). Step 1: Phenotype history reconstruction. Step 2: Map of the phenotypic shifts in our phylogenetic tree. Step 3: Proteome history reconstruction. Step 4: Map of all the convergent (allo or iso) substitutions in our phylogenetic tree, selecting in *yellow* the ones that co-occur with the phenotypic shift. Step 5: Ortholog genes having convergent substitutions are candidates for the phenotype shift. This ranking can then be redefined with other predictive analysis such as function prediction (see Parker et al. 2013; Foote et al. 2015). Step 6: Experimental task

isoconvergent apomorphic sites. This is done by reading automatically the phylogenetic tree. However, in these studies and in all the studies published so far that linked the phenotype to the genotype, the reconstruction of the phenotypic character evolution (most of time discrete characters) is performed with an algorithm based on the Mirkin et al.'s (2003) approach, even if the authors do not write it explicitly. This algorithm evidences where the characters are present on a group of leaves and where they are absent, to conclude that the characters appeared after the separation of the groups having or not the characters, which is an approximation, except if one has access to fossils or other information as in the case of the marine mammals where one can guess that the ancestral reconstruction of the way of life is correct since lots of information show that the ancestors of the marine groups were terrestrial.

It should be noted that in general, character reconstruction could be done using stochastic approaches (maximum likelihood; for example, see for review Royer Carenzi et al. 2013).

Of course, convergence at amino acid levels due to noise is possible and gives a false signal. The significance increases if many positions are needed to get the new phenotype (and if several genes are needed for a given phenotype). However, it has to be noted that only few amino acid changes can lead to a shift of the function for a given protein and that, even if the statistical test is inconclusive, the changes can be involved in the phenotypic shift.

Zhang and Kumar (1997) developed a test where they looked whether the number of alloconvergent or isoconvergent amino acids that occurred in two or more orthologues is above what is expected. A similar but different test has been developed by Castoe et al. (2009). In their analysis, these authors were interested to test whether two species evolved more convergently (at least) for a part of their genomes. The authors tested all the branches and tested whether the chosen species evolved more convergently than the other couple of species or against simulated data.

Detection of evolutionary isoconvergent "characteristics" at genomic level
Coding sequence

If the isoconvergent evolution at the phenotypic level results from adaptation, it is possible that the causal mutation at the genome level has been selected, and a test to detect positive selection can be used so that the protein that evolved under positive selection in a convergent manner can therefore be detected. The nice thing with this method is that even if different amino acids are involved in a similar functional shift in the coding sequence, the genotype to the phenotype link is possible (this approach has been used by Parker et al. (2013), Foote et al. (2015), for review: Levasseur et al. 2007). Concerning the shift on the constraints seen on common or different sites of the same protein (Gu et al. 2001; Lichtage et al. 1996; Gribaldo et al. 2003), a general relaxation of a protein (different sites on the same protein) could mean that some of its functions are less important for the fitness of the species. Here again, we can test whether we have more relaxation than expected. An extreme case is the loss of function (pseudogenization). This has been nicely shown in the case of the convergent loss of the GULO gene involved in the vitamin C synthesis (Hiller et al. 2012).

We could also have the case of different genes involved in the same convergent phenotype but belonging to the same network, for example: all the orthologues of a network where at least one gene evolves under positive selection or has the same constraint in the different species presenting the convergent phenotypic character. The protein corresponding to these genes could be involved in the same cascade, the same network, the same pathway (Foote et al. 2015).

Non-coding sequence

In analyses at the promoter level, one first needs to identify the promoters. This is possible for species that are close phylogenetically as promoter conservation is possible. In any cases, it is easier to detect convergent evolution in the case of a deregulation (loss of the expression in a given tissue, for example). In that case, the promoters should evolve quicker than expected, as reported for Shaven Baby

(Frankel et al. 2012). Therefore, the cause of an expression loss should be possible to detect.

However, it would be really difficult to evidence the causal mutation of a new expression territory (especially if the mutation is not the same). Nonetheless, many examples in the literature show that a new expression territory could be due to the integration of a retro-element in the gene promoter (Lynch et al. 2015) or a new function could be due to an LTR coding sequence (Naville et al. 2016; Pavlicev et al. 2015), and one can look for such sequences surrounding the orthologous genes in the species that show convergent evolution at phenotypic level and evidence their absence in the species where the phenotype is absent.

Molecular level: convergent evolution of physical–chemical protein characteristics

A new protein function could not be due to specific amino acid changes but due to the overall physical properties. In that case, it is possible to perform an ancestral reconstruction of the physical properties of the protein families and looked for isoconvergent shift and then looked for the co-isoconvergence for the studied function (see Mirceta et al. 2013; Ujvari et al. 2015).

Detection of genetic events other than substitutions

Other genetic events can be evidenced (Gouret et al. 2011; Dainat et al. 2012; Dainat and Pontarotti 2014; Le et al. 2012; Paganini et al. 2012) and linked to the phenotype shift (Cayrou et al. 2012; Levasseur et al. 2012).

1.2.2.5 Detection of Isoconvergent Evolution at Intermediate Biological Levels

Expression data

The difference is observed directly in one gene of the GRN. As said above, it will be really difficult to identify the causative mutation. Therefore, in that case, one needs to investigate the expression data of the concerned tissues. Once the gene co-opted is evidenced, one can look at the cis regulatory region or search for an epigenetic signature such as the methylation status of the gene that has been shown to be over- or underexpressed as reported by Lynch et al. (2015).

We therefore need to have access to the concerned tissues and perform expression studies using a methodology similar to the one described by Pankey et al. to evidence the similarity between the two isoconvergent derived tissues and the difference with the others. These analyses can then be integrated in an interaction network, signalling pathway, metabolic pathway or gene regulation network (Gallant et al. 2014; Pankey et al. 2014; Pfenning et al. 2014).

It has to be noted that in that case, the authors did not use ancestral reconstruction followed by the search of the shift. They performed a comparative transcriptomic analysis between the new cells or organs, as compared to the other cells or organs of

the species where the convergent cell tissues have been evidenced, and looked for what is common for the isoconvergent tissues and different for the others.

Integration of the different biological levels
The further step is to integrate the different levels seen in Fig. 1.1, in order to explain the cause/consequence effect from the genotype to the phenotype. To the best of our knowledge, only one example of co-convergence at sequence, expression and phenotype has only been done in one case where mutation in the promoter was linked to the expression loss, leading to the loss of a phenotype (see for a complete story Stern and Frankel 2013). In that case, the promoters should evolve quicker than expected. This is what is happening for Shaven Baby (Frankel et al. 2012), a GRN gene that controls several differential battery genes. In that case, the loss of a phenotypic trait should be possible to detect. However, to evidence the causal mutation of a new expression territory would be really difficult. However, as many examples in the literature show that a new expression territory could be due to a new integration of a retro-element in the gene promoter (Lynch et al. 2015), the identification of similar retro-element in front of differential expressed gene should be possible.

1.2.3 Database Organization

Based on the points discussed in this article, it will be useful to develop a database and mandatory that this database uses an ontology that takes into account the different levels and links one to the other (Fig. 1.2). The database could be filled in by the different collaborators. The database we propose will offer a new way to analyse the different types of convergent evolution: separate apparition of the same derived state (apomorphic) in at least two distinct lineages from the same ancestral state (isoconvergence) or from different ancestral states (alloconvergence), considering the different biological levels. In other words, an amino acid substitution could induce a shift in an enzymatic function. The latter could be linked to a shift in the metabolite that can infer a shift in the phenotype at morphological and/or at physiological levels. Thus, because of the complexity of the reality and the wide variety of the data, this database will be flexible enough to store biological data as well as to incorporate different biological levels.

1.3 Conclusion/Road Map

In conclusion, we propose (1) to re-analyse all the cases of convergent evolution at the phenotype level described in the literature and sort out cases of isoconvergence, (2) to identify undescribed cases of isoconvergent evolution, (3) to use the strategies

developed in this article to study these cases and on selected school case studies as, for example, the evolution of the active electro-localization in fishes and (4) to create an international effort by the access of the database that is able to integrate the different cases of isoconvergence at different biological levels.

Acknowledgments We would like to acknowledge our colleagues from the EBM team, as well as Benoit Heulin, Olivier Sandra and Laurent Journot for helpful discussions and Christophe Klopp and Mike Speed for critical reading of the manuscript.

References

Alderson RG, Barker D, Mitchell JB (2014) One origin for metallo-β-lactamase activity, or two? An investigation assessing a diverse set of reconstructed ancestral sequences based on a sample of phylogenetic trees. J Mol Evol 79(3–4):117–129

Alfaro ME, Bolnick DI, Wainwright PC (2005) Evolutionary consequences of many-to-one mapping of jaw morphology to mechanics in labrid fishes. Am Nat 165:E140–E154

Arbuckle K, Bennett CM, Speed MP (2014) A simple measure of the strength of convergent evolution. Methods Ecol Evol 5(7):685–693

Bird DM, Jones JT, Opperman CH, Kikuchi T, Danchin EG (2015) Signatures of adaptation to plant parasitism in nematode genomes. Parasitology 142(Suppl 1):S71–S84

Buttigieg PL, Morrison N, Smith B, Mungall CJ, Lewis SE, ENVO Consortium (2013) The environment ontology: contextualising biological and biomedical entities. J Biomed Semantics 11;4(1):43

Conway Morris S eds (2003) Life's solution: inevitable humans in a lonely universe. Cambridge University Press, Cambridge

Castoe TA, de Koning AP, Kim HM, Gu W, Noonan BP, Naylor G, Jiang ZJ, Parkinson CL, Pollock DD (2009) Evidence for an ancient adaptive episode of convergent molecular evolution. Proc Natl Acad Sci USA 106(22):8986–8991

Cayrou C, Henrissat B, Gouret P, Pontarotti P, Drancourt M (2012) Peptidoglycan: a post-genomic analysis. BMC Microbiol 12:294

Conte GL, Arnegard ME, Peichel CL, Schluter D (2012) The probability of genetic parallelism and convergence in natural populations. Proc Biol Sci 279(1749):5039–5047

Dainat J, Pontarotti P (2014) Methods to study the occurrence and the evolution of pseudogenes through a phylogenetic approach. Methods Mol Biol 116:87–99

Dainat J, Paganini J, Pontarotti P, Gouret P (2012) GLADX: an automated approach to analyze the lineage-specific loss and pseudogenization of genes. PLoS One 7(6)

Dick R, Rattei T, Haslbeck M, Schwab W, Gierl A, Frey M (2012) Comparative analysis of benzoxazinoid biosynthesis in monocots and dicots: independent recruitment of stabilization and activation functions. Plant Cell 24(3):915–928

Doolittle RF (1994) Convergent evolution: the need to be explicit. Trends Biochem Sci 19:15–18

Elmer KR, Meyer A (2011) Adaptation in the age of ecological genomics: insights from parallelism and convergence. Trends Ecol Evol 26(6):298–306

Foote AD, Liu Y, Thomas GW, Vinař T, Alföldi J, Deng J, Dugan S, van Elk CE, Hunter ME, Joshi V, Khan Z, Kovar C, Lee SL, Lindblad-Toh K, Mancia A, Nielsen R, Qin X, Qu J, Raney BJ, Vijay N, Wolf JB, Hahn MW, Muzny DM, Worley KC, Gilbert MT, Gibbs RA (2015) Convergent evolution of the genomes of marine mammals. Nat Genet 47(3):272–275

Frankel N, Wang S, Stern DL (2012) Conserved regulatory architecture underlies parallel genetic changes and convergent phenotypic evolution. Proc Natl Acad Sci USA 51:20975–20979

Gallant JR, Traeger LL, Volkening JD, Moffett H, Chen PH, Novina CD, Phillips GN Jr, Anand R, Wells GB, Pinch M, Güth R, Unguez GA, Albert JS, Zakon HH, Samanta MP, Sussman MR (2014) Genomic basis for the convergent evolution of electric organs. Science 344(6191): 1522–1525

Gherardini PF, Wass MN, Helmer-Citterich M, Sternberg MJE (2007) Convergent evolution of enzyme active sites is not a rare phenomenon. J Mol Biol 372(3):817–845

Gordon MS, Notar JC (2015) Can systems biology help to separate evolutionary analogies (convergent homoplasies) from homologies? Prog Biophys Mol Biol 117(1):19–29

Gouret P, Paganini J, Dainat J, Louati D, Darbo E, Pontarotti P, Levasseur A (2011) Integration of evolutionary biology concepts for functional annotation and automation of complex research in evolution: the multi-agent software system. In: P Pontarotti (ed) Evolutionary biology-concepts biodiversity, macroevolution and genome evolution, pp 71–87

Gribaldo S, Casane D, Lopez P, Philippe H (2003) Functional divergence prediction from evolutionary analysis: a case study of vertebrate haemoglobin. Mol Biol Evol 11:1754–1759

Gu X (2001) Maximum likelihood approach for gene family evolution under functional divergence. Mol Biol Evol 18:453–464

Hallström BM, Janke A (2010) Mammalian evolution may not be strictly bifurcating. Mol Biol Evol 27(12):2804–2816

Hiller M, Schaar BT, Indjeian VB, Kingsley DM, Hagey LR, Bejerano GA (2012) A "forward genomics" approach links genotype to phenotype using independent phenotypic losses among related species. Cell Rep 2(4):81723

Ingram T, Mahler DL (2013) SURFACE «A simple measure of the strength of convergent evolution»: detecting convergent evolution from comparative data by fitting Ornstein-Uhlenbeck models with stepwise Akaike Information Criterion. Methods Ecol Evol 4:416–425

Jones FC, Grabherr MG, Chan YF, Russell P, Mauceli E, Johnson J, Swofford R, Pirun M, Zody MC, White S, Birney E, Searle S, Schmutz J, Grimwood J, Dickson MC, Myers RM, Miller CT, Summers BR, Knecht AK, Brady SD, Zhang H, Pollen AA, Howes T, Amemiya C, Broad Institute Genome SequencingPlatform, Whole Genome Assembly Team, Baldwin J, Bloom T, Jaffe DB, Nicol R, Wilkinson J, Lander ES, Di Palma F, Lindblad-Toh K, Kingsley DM (2012) The genomic basis of adaptive evolution in threespine sticklebacks. Nature 484(7392):55–61

Kopp A (2009) Metamodels and phylogenetic replication: a systematic approach to the evolution of developmental pathways. Evolution 63(11):2771–2789

Le PT, Ramulu HG, Guijarro L, Paganini J, Gouret P, Chabrol O, Raoult D, Pontarotti P (2012) An automated approach for the identification of horizontal gene transfers from complete genomes reveals the rhizome of Rickettsiales. BMC Evol Biol 12:243

Levasseur A, Orlando L, Bailly X, Milinkovitch MC, Danchin EG, Pontarotti P (2007) Conceptual bases for quantifying the role of the environment on gene evolution: the participation of positive selection and neutral evolution. Biol Rev Camb Philos Soc 82(4):551–572

Levasseur A, Paganini J, Dainat J, Thompson JD, Poch O, Pontarotti P, Gouret P (2012) The chordate proteome data base. Evol Bioinform Online 8:437–447

Lichtarge O, Bourne HR, Cohen FE (1996) An evolutionary trace method defines binding surfaces common to protein families. J Mol Biol 257(2):342–358

Losos JB (2011) Convergence, adaptation, and constraint. Evolution 65(7):1827–1840

Lynch VJ, Nnamani MC, Kapusta A, Brayer K, Plaza SL, Mazur EC, Emera D, Sheikh SZ, Grützner F, Bauersachs S, Graf A, Young SL, Lieb JD, DeMayo FJ, Feschotte C, Wagner GP (2015) Ancient transposable elements transformed the uterine regulatory landscape and transcriptome during the evolution of mammalian pregnancy. Cell Rep 10(4):551–561

Mahler DL, Ingram T, Revell LJ, Losos JB (2013) Exceptional convergence on the macroevolutionary landscape in island lizard radiations. Science 341(6143):292–295

Martin A, Orgogozo V (2013) The Loci of repeated evolution: a catalog of genetic hotspots of phenotypic variation. Evolution 67(5):1235–1250

McGhee GR (2011) Convergent evolution: limited forms most beautiful. The MIT Press, Cambridge

McGowen MR, Gatesy J, Wildman DE (2014) Molecular evolution tracks macroevolutionary transitions in Cetacea. Trends Ecol Evol 29(6):336–346

Mirceta S, Signore AV, Burns JM, Cossins AR, Campbell KL, Berenbrink M (2013) Evolution of mammalian diving capacity traced by myoglobin net surface charge. Science 340:1234192

Mirkin BG, Fenner TI, Galperin MY, Koonin EV (2003) Algorithms for computing parsimonious evolutionary scenarios for genome evolution, the last universal common ancestor and dominance of horizontal gene transfer in the evolution of prokaryotes. BMC Evol Biol 3:2

Moen DS, Morlon H, Wiens JJ (2016) Testing convergence versus history: convergence dominates phenotypic evolution for over 150 million years in frogs. Syst Biol 65(1):146–160

Muschick M, Indermaur A, Salzburger W (2012) Convergent evolution within an adaptive radiation of cichlid fishes. Curr Biol 22(24):2362–2368

Naville M, Warren IA, Haftek-Terreau Z, Chalopin D, Brunet F, Levin P, Galiana D, Volff JN (2016) Not so bad after all: retroviruses and LTR retrotransposons as a source of new genes in vertebrates. Clin Microbiol Infect. doi:10.1016/j.cmi.2016.02.001. (Epub ahead of print)

O'Leary MA, Kaufman S (2011) MorphoBank: phylophenomics in the "cloud". Cladistics 27:529–537

Omland KE, Lanyon SM (2000) Reconstructing plumage evolution in orioles (Icterus): repeated convergence and reversal in patterns. Evolution 54(6):2119–2133

Paganini J, Campan-Fournier A, Da Rocha M, Gouret P, Pontarotti P, Wajnberg E, Abad P, Danchin EGJ (2012) Contribution of lateral gene transfers to the genome composition and parasitic ability of root-knot nematodes. Plos One 7(11):e50875

Pankey MS, Minin VN, Imholte GC, Suchard MA, Oakley TH (2014) Predictable transcriptome evolution in the convergent and complex bioluminescent organs of squid. Proc Natl Acad Sci USA 11(44):E473642

Parker J, Tsagkogeorga G, Cotton JA, Liu Y, Provero P, Stupka E, Rossiter SJ (2013) Genome-wide signatures of convergent evolution in echolocating mammals. Nature 502 (7470):228–231

Pavlicev M, Hiratsuka K, Swaggart KA, Dunn C, Muglia L (2015) Detecting endogenous retrovirus-driven tissue-specific gene transcription. Genome Biol Evol 7(4):1082–1097

Pfenning AR, Hara E, Whitney O, Rivas MV, Wang R, Roulhac PL, Howard JT, Wirthlin M, Lovell PV, Ganapathy G, Mouncastle J, Moseley MA, Thompson JW, Soderblom EJ, Iriki A, Kato M, Gilbert MT, Zhang G, Bakken T, Bongaarts A, Bernard A, Lein E, Mello CV, Hartemink AJ, Jarvis ED (2014) Convergent transcriptional specializations in the brains of humans and song-learning birds. Science 346(6215):1256846

Rosenblum EB, Parent CE, Brandt EE (2014) The molecular basis of phenotypic convergence annual review of ecology. Evol Syst 45:203–226

Royer-Carenzi M, Pontarotti P, Didier G (2013) Choosing the best ancestral character state reconstruction method. Math Biosci 242(1):95–109

Sanderson MJ, Hufford L (eds) (1996) Homoplasy: the recurrence of similarity in evolution. Academic Press, New York

Soria-Carrasco V, Gompert Z, Comeault AA, Farkas TE, Parchman TL, Johnston JS, Buerkle CA, Feder JL, Bast J, Schwander T, Egan SP, Crespi BJ, Nosil P (2014) Stick insect genomes reveal natural selection's role in parallel speciation. Science 344(6185):738–742

Stayton CT (2015) The definition, recognition, and interpretation of convergent evolution, and two new measures for quantifying and assessing the significance of convergence. Evolution 269(8):2140–2153

Stern DL, Orgogozo V (2008) The loci of evolution: how predictable is genetic evolution? Evolution 62(9):2155–2177

Stern DL (2013) The genetic causes of convergent evolution. Nat Rev Genet 14(11):751–764

Stern DL, Frankel N (2013) The structure and evolution of *cis*-regulatory regions: the *shaven baby* story Philos Trans R Soc. Lond B Biol Sci 68(1632):20130028

Ujvari B, Casewell NR, Sunagar K, Arbuckle K, Wüster W, Lo N, O'Meally D, Beckmann C, King GF, Deplazes E, Madsen T (2015) Widespread convergence in toxin resistance by predictable molecular evolution. Proc Natl Acad Sci USA 112(38):11911–11916

Zakon HH, Lu Y, Zwickl DJ, Hillis DM (2006) Sodium channel genes and the evolution of diversity in communication signals of electric fishes: convergent molecular evolution. PNAS 103:3675–3680

Zhang J, Kumar S (1997) Detection of convergent and parallel evolution at the amino acid sequence level. Mol Biol Evol 14(5):527–536

Chapter 2
Analysing Convergent Evolution: A Practical Guide to Methods

Kevin Arbuckle and Michael P. Speed

Abstract Convergent evolution, or the independent evolution of similar traits, has long been investigated and recognised as an important area of research for evolutionary biology. However, as with many areas of comparative biology, new phylogenetic methods that enhance our ability to study convergence have arisen with greater frequency in recent years. Consequently, we now have a wide range of tools at our disposal and a rapidly developing conceptual framework to guide us in our analyses. This chapter aims to provide a practical guide for those interested in convergent evolution that will enable new entrants to the field to quickly develop a well-rounded research agenda. Although some methods can be performed in other pieces of (stand-alone) software, this guide will focus on the R statistical environment.

2.1 Introduction

Convergent evolution is a common phenomenon across the diversity of living organisms. In essence, it refers to the independent evolution of some kind of similarity between two or more organisms, as opposed to any similarity which is a result of inheritance from a common ancestor. Convergent traits may be manifest across a number of levels of biology including both function and form (Losos 2011; Speed and Arbuckle 2016). Convergence can be seen for example in many forms of behaviour, morphology and physiology (McGhee 2011), and in the structure and action of molecules such as toxins or enzymes (Doolittle 1994). For instance, we could consider the phenomenon of mimicry, in which one organism (perhaps a harmless viceroy butterfly, *Limenitis archippus*) evolves to appear like a different

K. Arbuckle (✉) · M.P. Speed
Department of Ecology Evolution and Behaviour Biosciences Building, University of Liverpool, Crown Street, Liverpool L69 7ZB, UK
e-mail: k.arbuckle@liverpool.ac.uk

M.P. Speed
e-mail: speedm@liverpool.ac.uk

P. Pontarotti (ed.), *Evolutionary Biology*, DOI 10.1007/978-3-319-41324-2_2

organism (such as a poisonous monarch butterfly, *Danaus plexippus*) in order to deceive another (e.g. a predator) and gain some advantage as a result (Cott 1940; Ruxton et al. 2004). A contrasting example to highlight the wide reach of convergence is the evolution of myoglobin with similar oxygen binding properties in the muscles of several aquatic mammal lineages, which enables prolonged diving ability (Mirceta et al. 2013).

Convergence has long been considered as an important area of research in evolutionary biology. For example, in Chap. 6 of The Origin of Species, Darwin (1859) discussed several perceived difficulties with his theory of natural selection and invoked evolutionary convergence to counter some of these potential problems. More recently, convergence has been at the heart of modern comparative biology, albeit in an understated way. In particular, the power and applicability of phylogenetic comparative methods is typically limited by the number of independent replicates, which refers to the number of independent origins in these analyses. Hence, the rate of convergence of traits in a given case study can be a limiting constraint on the power of phylogenetic estimations (Maddison and FitzJohn 2015). Indeed, we often use comparative methods to investigate questions of adaptation or, more broadly, the relationship between different organismal traits across a phylogenetic tree, and our evidence for this is typically gained from several independent gains or losses of a set of traits. Stated in this way, we argue, it becomes clear why convergence should occupy a key position in the quantitative evaluation of macroevolutionary patterns.

As with many areas of phylogenetic comparative biology, methods to investigate convergent evolution have undergone a resurgence, with many new approaches having been devised in recent years. This has coincided with renewed interest in the ways in which we analyse convergence and how we think about it as a concept. While some excellent reviews already exist which focus on particular methods (e.g. Mahler and Ingram 2014), the aim of this chapter is to give a broad guideline of how a researcher might undergo a study of evolutionary convergence.

We will first discuss the types of questions we can ask and what we should set as our analytical goals. We will then give a brief survey of the software which is available to implement the methods discussed herein, followed by guidance on actually running the analyses themselves. For clarity and in the interest of making this chapter a practical guide rather than a review per se, we will not attempt to be entirely comprehensive in the methods we suggest, but instead aim to provide a selection that could be combined to generate a strong piece of research (and see review in Speed and Arbuckle 2016). Note that certain methods will be discussed in different sections, but with a focus on different aspects. In particular, methods that can be used both to detect and quantify convergence will be discussed in both sections covering each of those elements. We will end with a brief summary and some future directions for workers in convergent evolution and method developers.

2.2 General Aims in Studies of Convergent Evolution

There are clearly many different possible aims for particular studies of convergent evolution, depending on the system at hand and the interests of particular researchers. Nevertheless, there are some generalities that can be made and it is at this broad level that methodological recommendations can often be directed, which can then be applied to the specific details under investigation at any particular time. In particular, the goal of most studies can be loosely (but not mutually exclusively) divided into the identification or the quantification of evolutionary convergence.

The most common fundamental aim of studies of convergent evolution is to establish whether convergence is present in a particular trait in a particular group of species. For instance, Ujvari et al. (2015) demonstrated that resistance to cardiac glycoside toxins (and the molecular mechanism of this resistance) has independently evolved several times in animals as distantly related as mammals, reptiles, amphibians and insects. Similarly, Westneat et al. (2005) identified frequent convergence in the biomechanical function of jaws in wrasses (Labridae).

Evolutionary convergence can be quantified in two different ways: first by measuring its frequency and second by measuring its 'strength' (e.g. Stayton 2015; Speed and Arbuckle 2016). The aim of both approaches is to provide some way of measuring the influence of evolutionary convergence on the range of traits observed within specified groups, or within specified 'niches'. Enumerating the cases of convergence of phenotypic categories for example allows us to determine the frequency of convergence (Stayton 2015) within a specified set of organisms. Measuring the strength of convergence, by contrast, requires us to ask 'how similar are the trait(s) of organisms within a putative convergent subset?' These two quantitative aspects should be measured separately and considered to be potentially independent. We could, for example, imagine a scenario in which convergence of classes of traits is frequent within a set of organisms, but weak in the sense that the convergent phenotypes are relatively diverse in quantitative terms. Conversely we can imagine rare, but strong convergence.

Examples of both kinds of quantification are becoming increasingly common in the literature. Mahler et al. (2013), for instance, provided several measures of the frequency of morphological convergence (in categorical traits) in Caribbean anoles to provide strong evidence for convergence among different radiations on different islands. More recently, Vidal-García and Keogh (2015) incorporated measures of the strength of convergent evolution into their analysis of (quantitative) morphological convergence in Australian frogs occupying different niches. Both studies benefit from a finer-grained understanding of the questions they were trying to address.

An advantage of the recent move towards quantifying convergence rather than simply documenting it is that it opens up new potential areas of research (Arbuckle et al. 2014; Speed and Arbuckle 2016). For instance, the ability to compare different cases of convergence allows us to explore differences in its frequency or strength when different types of traits or groups of organisms are considered. Asking, for

example, whether some traits are consistently more convergent than others, or whether a particular ecological niche favours stronger convergence than others will provide insights into the creation and limits of biodiversity. Such work will eventually permit a nuanced view of general patterns of convergence, and perhaps the process that generates these patterns, without being limited to a particular model system.

Quantification of convergence will also allow us to assess convergence within a particular system (or multiple systems) at different 'levels of life' (Losos 2011; Speed and Arbuckle 2016). By this, we mean levels such as function, form, development and genetics of a particular trait. This will perhaps provide information on precisely how convergence typically evolves, and therefore identify common constraints on the ability of organisms to evolve similarity over a particular period of time. Comparing the attributes of convergence across different levels of life will also point towards traits that are most remarkably convergent: those traits which have independently become similar at many levels (Arbuckle and Speed 2016).

2.3 Software Available for Analysing Convergent Evolution

There has been a recent surge of methodological development for the study of convergence, and a related increase in available software for such analyses. This section is not intended to give a comprehensive overview of these, but rather a selection with a focus on R packages. We do this because R has several advantages including (1) many new methods are implemented within it, often the only implementation is in an R package, (2) many biologists will already be familiar with it and (3) data handling and other analyses can be managed within the same environment as convergence methods, which reduces the need for data reformatting. See Table 2.1 for a summary of the software covered below and the following two sections for more details about particular methods and how they can be used.

Although many of the newer quantitative methods specifically designed for analyses of convergence are implemented in R, a simple form of detecting convergence is (of course) via ancestral state reconstruction (ASR). Because ASR is been a long-established technique in comparative biology for more reasons than looking for convergence, there are a number of good stand-alone programs to do this. For instance, BayesTraits (Pagel et al. 2004) reconstructs ancestral states by maximum likelihood (ML) or Bayesian Markov chain Monte Carlo (MCMC) methods. RASP (Yu et al. 2015) was primarily developed for reconstructing geographical ranges of ancestors, but can also be used for other traits and can consider polymorphic ancestral states. RASP uses an MCMC approach for ASR. Finally, Mesquite (Maddison and Maddison 2015) also has functions for reconstructing ancestral states by maximum parsimony (MP) or ML.

Table 2.1 A selection of software that are useful for analyses of convergent evolution

Platform	Software	Type of measure	Method(s) implemented
Stand-alone	BayesTraits	Identification	ASR (ML and MCMC)
	Mesquite	Identification	ASR (MP and ML)
	RASP	Identification	ASR (MCMC)
R packages	ape	Identification	ASR (MP and ML)
	ouch	Identification	OU models on custom regimes
	phytools	Identification	ASR (ML, MCMC and stochastic character mapping)
	corHMM	Identification	ASR (ML; including hidden rates models)
	surface	Identification and quantification	OU models over the phylogeny with tests for convergence
	convevol	Identification and quantification	Phylomorphospace and distance-based measures ('C-metrics')
	windex	Quantification	Wheatsheaf index

Additional details of software and methods can be found in the main text. Abbreviations are as follows: *ASR* = Ancestral state reconstruction, *ML* = Maximum likelihood, *MCMC* = Markov chain Monte Carlo, *MP* = Maximum parsimony, *OU* = Ornstein–Uhlenbeck

Within R, several options are available for identifying cases of convergence. This is once again mostly using ASR methods. For instance, ape (Paradis et al. 2004) can reconstruct ancestral states using MP or ML, while corHMM (Beaulieu et al. 2013) uses ML to reconstruct categorical traits and can do so using either standard transition rate models or hidden rates models. The phytools package (Revell 2012) has numerous functions for estimating ancestral states using ML, MCMC and stochastic character mapping and has a particularly wide range of models from which to perform the ASR, including the incorporation of trends and threshold models. We note that most R functions for ASR, including those listed above, make use of a numerical optimisation algorithm rather than calculating an exact solution (for cases where this is possible). Nevertheless, the latter can be implemented using the 'reconstruct' function in ape, which calculates ancestral states using exact calculus. Finally, the package surface (Ingram and Mahler 2013) implements the method of the same name by acting as a wrapper for functions from ouch (Butler and King 2004).

The most simple way of quantifying convergent evolution in a trait is simply to identify it via a method such as ASR or SURFACE (note that the method is written in block capitals whereas the package is written in lower case) and then count the number of independent origins of the trait (e.g. Foote et al. 2015). This gives one measure of the frequency of convergence in the data set. However, a few software packages have been designed to provide additional options for measuring convergence. We have already mentioned the surface package in the context of its identification of convergent regimes, but it also provides additional information which can be used as a form of quantification. Specifically, in addition to the number of

convergent regimes found in the SURFACE analysis, other values such as the proportion of total regime shifts that are convergent and the reduction in the complexity of phenotypic evolution once convergence is accounted for can also be extracted (Ingram and Mahler 2013). Therefore, surface provides several measures of the frequency of convergence.

Several methods recently developed by Stayton (2015), herein referred to as 'C-metrics', are implemented in the convevol package (Stayton 2015) and provide measures of both frequency and strength of convergent evolution in a particular data set. This package also enables significance testing of whether the amount of convergence is 'surprising' (sensu Stayton 2008) using simulations under a Brownian motion model. Finally, another recently developed method, the Wheatsheaf index (Arbuckle et al. 2014), aims to quantify the relative strength of convergence in a subset of species (defined a priori) within a larger data set and is implemented in the windex package (Arbuckle and Minter 2015).

2.4 Detecting Convergent Evolution

The most basic and fundamental question we can ask about convergence is whether or not it occurs in a data set, or at least whether or not we have evidence for it in a data set. Not only is this an important question in itself, but it is of course vital to know prior to (or at least simultaneously with) quantification whether there is anything to quantify in the first place. It should also be noted that in the case of categorical traits, quantification is limited to frequency of convergence, and so its detection becomes proportionately more important in these cases.

Ancestral state reconstruction—The most basic way of testing for the presence of convergence (and again the only way for categorical traits) is to perform an ASR. As alluded to in the previous section, there are a multitude of methods of estimating ancestral states and even more options of software for their implementation (Table 2.1). Once you have completed your ASR there are many options for visualisation (e.g. Revell 2014) and the interpretation is very straightforward for categorical characters; the trait is convergent if it has arrived at the same state more than once across the phylogeny. For continuous traits, the interpretation can often be a little more subjective; how similar do such traits have to be before there is good evidence of convergence? In most cases, this will not be a problem (particularly where convergence is strong) as there will often be a clear shift in trait values that is shared between convergent lineages but is notably different from intermediate estimates. However, where convergence is not particularly strong, detection from ASR will be more difficult and inference will typically be informed by knowledge of what level of variation in the trait is biologically important in the system under study.

Where inference from ASR is not clear but higher certainty is wanted, then quantitative methods such as cluster analysis or dendrograms of trait values (also known as 'phenograms') may be useful. In these scenarios, evidence of

convergence would be gained from the clustering of phylogenetically disparate lineages occupying distinct clusters, or 'clades' in a phenogram (see examples in Speed and Arbuckle 2016). Although evolutionary stasis could produce similar patterns to convergence here, the distribution of other species within clusters may provide some idea of whether this is likely to be a good explanation or not.

An additional consideration when using ASR is which type of method to use. This can be considered first on two levels: maximum parsimony (MP) and model-based methods. Although contrasting views still exist in the literature, we recommend against using MP for this purpose. General criticisms of MP for comparative analyses apply here (e.g. Cunningham et al. 1998; Currie and Meade 2014). For instance, MP ignores information contained in branch lengths and therefore essentially assumes that all branch lengths are equal; an assumption that is unlikely to apply to most phylogenies. It also assumes that evolutionary rates have been slow, which may be the case but this can be explicitly estimated in model-based approaches and accounted for as part of model fitting. MP is also poor at reflecting the level of uncertainty in the reconstruction compared to model-based approaches which can provide a probability for each state at each node in the tree. Finally, specifically in the context of convergence, MP explicitly tries to minimise changes in the trait and therefore is likely to be biased against recovering convergent evolution since this necessitates additional changes.

Having decided to use a model-based method for ASR, the next choice is which model you want to fit. There are a range of options here depending firstly on whether the trait in question is categorical or continuous. Categorical trait models are described in terms of a transition rate matrix, and this could involve each rate parameter in the matrix being different or include various constraints such as certain rates forced to be equal or forced to be zero (such that a particular change can't happen). There are also recent but increasingly used categorical trait models such as threshold or hidden rates models that might be appropriate for some systems. In terms of continuous trait models, there are again a range of options including Brownian motion, Ornstein–Uhlenbeck (OU), and trend models. Whatever range of models you feel is plausible, it is usually worth comparing the fit of each model in some way to give an empirical justification for a particular model choice (e.g. using AIC, likelihood ratio tests for nested models or Bayes factors).

Finally, there is a choice of how to fit the model. For example, you can fit many models in either a frequentist or a Bayesian framework, although the implementation of some models may only be currently available in one or the other. However, where there is a choice, this decision essentially comes down to informed personal preference since there are pros and cons to both approaches and strong and convincing advocates of each.

Ornstein–Uhlenbeck (OU) models—OU models essentially represent evolution of a continuous trait which is changing around an 'optimum' value (Felsenstein 1988; Hansen 1997). As the trait changes away from the optimum value, it experiences a 'pull' back to the optimum, the strength of which increases as the trait evolves further from the optimum. This is arguably a better representation of adaptive convergent evolution than other models because the optimum in the model

can be considered to represent an adaptive peak. Ingram and Mahler (2013) used OU modelling combined with stepwise model selection via AIC to test for convergence in one or (ideally) multiple continuous traits; a method they called SURFACE.

Briefly, SURFACE fits OU models over a phylogeny and uses AIC to identify where a change in the parameters of the model (or a 'regime shift') has occurred. This stage identifies where on the tree, for example, a trait has started to evolve around a new optimum/adaptive peak. In the next stage, SURFACE combines the different regimes estimated in the first stage and tests whether different combinations improve the model fit (again with AIC); in other words, it tests whether any of the regimes have evolved independently in different lineages. Interpretation of whether convergence has been detected is made simple by plotting the estimated regimes on the phylogeny in different colours, also implemented in the R package surface (Ingram and Mahler 2013), and looking to see whether the same regime appears multiple times.

It should be noted that a method very similar to SURFACE has recently been developed by Khabbazian et al. (2016) and implemented in the R package l1ou. The latter method uses a faster algorithm, making it more suitable for very large phylogenies, and can use a newly devised and more conservative information criteria for model selection. For practical purposes, this method can be substituted for SURFACE in the discussions herein, but as it is only very recently been released it has not yet been used as often as SURFACE.

Distance-contrast plots—Muschick et al. (2012) further developed a method to identify convergence that was proposed by Winemiller (1991): distance-contrast plots. The method makes use of the idea that convergent evolution will lead to short phenotypic distances (more similar phenotypes) relative to the phylogenetic distance between a convergent pair of species. Therefore, if pairwise distances (scaled between 0 and 1) are plotted with phylogenetic distance on the x-axis and phenotypic distances on the y-axis, convergence is indicated by species pairs which fall below the line of unity (where x = y) and especially species pairs which fall in the lower right-hand side. Note that the same pattern would also arise from evolutionary stasis, not only convergence, but this also applies to many methods and can be assessed using a method such as ASR to ensure independent origins.

The biggest problem with distance-contrast plots as a means of detecting convergence in a data set is which cut-off to use. In other words, when is a data point far enough below the line of unity to be considered evidence for convergence? Muschick et al. (2012) addressed this by conducting simulations of a trait under Brownian motion (to represent a 'neutral' trait). These simulated traits were used to generate a null expectation, and the distribution of pairwise distances from the putatively convergent trait is compared to this null expectation. Interpretation of convergence (or stasis) is therefore possible when many more species pairs fall into parts of the plot below the line of unity than can be explained by the neutral trait simulations. This method is not fully implemented within R as a separate function, but each part (deriving the required distances, simulations and plotting functions) is available either using standard R functions or those from ape (Paradis et al. 2004).

Position and movement in phylomorphospace—The use of the so-called phylomorphospace plots provides an intuitive way to identify convergent evolution, albeit one which is difficult to decide a threshold for considering a given pattern as evidence for convergence. Briefly, a phylomorphospace plot is a standard plot of at least two continuous traits measured in different species, but with a phylogenetic tree of the species superimposed over the plot and linking the data points. Furthermore, using an ASR method allows the plotting of estimated ancestral states on the same plot via the position of internal nodes of the plotted tree. This therefore allows tracing of the trait's evolution in phenotypic space by following the phylogeny. Phylomorphospace plots can be generated in phytools (Revell 2012), and an interpretation of convergent evolution is obtained from instances of multiple branches independently arriving in the same area of the plot. A specified area of the phylomorphospace plot can be highlighted using convevol (Stayton 2015) which helps to bring some level of objectivity to whether two species are similar enough to be considered convergent (although the initial specification of the area is still arbitrary).

2.5 Quantifying Convergent Evolution

Although the identification of convergence is a laudable and interesting goal in its own right, in most cases we can derive a better understanding of the system in question if we can quantify the convergence in some way (Arbuckle et al. 2014; Speed and Arbuckle 2016). As mentioned earlier, this quantification can relate to one of two aspects of convergence: its frequency and its strength. It is worth highlighting that for categorical traits, only the frequency can be measured and this can be done simply by counting the number of independent gains of such a trait from an ASR used to identify convergence. In fact, this also applies to all methods of identifying convergence above.

More broadly, Stayton (2008) highlighted that convergence can occur simply by chance when even a randomly evolving trait may become similar in multiple unrelated species. Depending on the research question at hand, we may be interested in convergence as a broad pattern (however it is generated) or we may be interested only in convergence that is more frequent (or stronger) than expected by chance. Of course, both of these measures could be interesting but the emphasis put on each one (total versus 'unexpected' convergence) will likely vary with the aim of the study.

For most methods of quantifying convergent evolution, either frequency or strength, simulations can be used to estimate how much convergence we would expect at random and use this for comparison with the observed amount of convergence in the data set to evaluate how much of the total convergence is 'unexpected'. Traits can be simulated in a wide range of R packages (e.g. ape) under many different models, but Brownian motion is commonly used as a model to simulate suitable 'neutral' traits for generating null expectations. In practice, many

(perhaps 1000 or more) traits could be simulated under your null model of choice, the method for measuring convergence applied to each one and the desired value (e.g. number of convergent events) extracted and used to plot a histogram. The actual values from the data can then be denoted on the histogram and used to judge how unexpected the amount of convergence is. Similarly, a P-value could be generated by taking the proportion of the simulated traits that showed the same amount or more convergence than observed in the data.

SURFACE—Although SURFACE is primarily a tool for detecting convergence (see previous section), the summary function for the method in surface allows more quantitative information to be gleaned on the frequency of convergence than simply counting the number of convergent regimes (Ingram and Mahler 2013). This was well illustrated in a study by the same authors that developed the method (and others) which investigated convergence in anole lizards (Mahler et al. 2013). The output from SURFACE can be summarised to generate additional metrics such as the number of shifts to convergent regimes (as above), but also the number of different convergent regimes, and the proportion of all regimes which are convergent. Therefore, SURFACE provides several measures of the frequency of convergence over and above simply counting the number of convergent events.

C-metrics—Stayton (2015) developed five measures of convergent evolution, termed C_1–C_5 and collectively referred to here as 'C-metrics'. All five C-metrics are related to the phylomorphospace concept outlined above, which allows the tracing of changes in similarity over time and across the tree. C_5 was considered separately from the other four metrics by Stayton (2015) as it deals with measuring the frequency of convergence rather than its strength. In essence, C_5 counts the number of lineages that cross into a prespecified area of phylomorphospace (the area denotes a minimum level of similarity to be considered convergent). The area acts as a boundary, and since only crossing the boundary is considered in the C_5 metric, it only counts independent origins of similarity. Furthermore, because estimates of ancestral states are considered in phylomorphospace, the value of C_5 can potentially include instances of convergent ancestors that have since diverged again.

The remaining four C-metrics (C_1–C_4) are also related to the phylomorphospace concept and so also consider the change in similarity through time. C_1 measures the change in similarity of two species as a proportion of the maximum distance the lineages have experienced. C_2 is conceptually similar to C_1 but is measured on an absolute scale in contrast to being relative to the maximum phenotypic distance. C_3 and C_4 are based on standardising C_2 for the total amount of evolutionary change in the clade of interest (defined in different ways), which allows comparison between data sets. Consequently, C_1–C_4 are quantitative measures of the strength of convergent evolution rather than simply its frequency, and can therefore separate convergent changes of different magnitudes, even when the actual number of times convergence has arisen remains the same. The interpretation of all of the C-metrics is based on higher values indicating stronger convergence, or a greater frequency of convergence in the case of C_5.

Wheatsheaf index—The Wheatsheaf index was developed as a way to quantify convergent evolution in a particular subset of species within a clade (Arbuckle et al.

2014). This subset is termed the 'focal group' and may be one of two types. The focal group can be species occupying a particular 'niche' (used loosely), or exploiting the environment in similar ways, for which convergence is expected—for instance burrowing animals when considering body shape, which will give a measure of how strongly the convergence is in a particular trait (or several traits) for the 'niche' of interest. Alternatively, the focal group can be a set of species previously identified as convergent using one of the methods for detecting convergence covered above. In this case, the Wheatsheaf index quantifies how strong the convergence is within the known convergent species without reference to potential drivers of that convergence. In either case, the Wheatsheaf index is designed to be implemented after convergence has been identified in the data set—strictly a means of quantifying convergence rather than detecting it. The interpretation is simple in that stronger convergence will produce larger values of the index.

The Wheatsheaf index combines two aspects of trait evolution. It considers convergence to be stronger when species are more phenotypically similar to one another, but also as the focal group become more different from other species (as this implies a greater distance traversed across an adaptive landscape to achieve the similarity). The windex package (Arbuckle and Minter 2015) also provides a test of whether the Wheatsheaf index is higher (i.e. convergence is stronger) than expected by chance given the topology of the underlying phylogeny. Note that because it assumes convergence is present in the data set, it is not designed to detect convergence per se, but simply whether it is substantially stronger than expected.

As a final point related to the Wheatsheaf index, we can consider it as also providing a measure of divergent selection—in that very low values of the index would tend to indicate more diversification in the traits in the focal subgroup compared to the overall set of species in the tree in question. As with its use to study convergence, we can use the significance tests implemented in windex to determine whether the level of phenotypic disparity is greater than expected (using $P > 0.95$ in that case). Hence, with this one measure, we can ask whether evolution has followed either a convergent or a divergent path.

2.6 Summary and Future Directions

In this chapter, we have attempted to provide an introduction to some of the practical aspects of studies of convergent evolution. Our hope is that readers with an interest in convergence can use this chapter as a starting point and as a set of guidelines to begin their research programmes. At the very least, we hope to have encouraged serious thought about how such a project might be structured and what types of methods are most desirable.

Although we have, to some degree, dealt with different facets of convergence separately here, a combined strategy is usually the best option. By this, we mean that unless a research question is deliberately narrow and specific, a program aimed at understanding convergent evolution in a particular system should incorporate

both the detection *and* quantification of convergence (Speed and Arbuckle 2016). Moreover, because different measures of convergence address different features of convergence, we recommend that a thorough investigation at least quantifies both the frequency and strength of convergence, and preferably uses multiple methods to do each of these.

In relation to using a multitude of methods, we wish to highlight a prime example of the type of approach we advocate. A recent paper by Friedman et al. (2016) investigated convergent evolution in surgeonfishes with respect to feeding niche. They employed a range of methods to detect convergence in morphology as well as to quantify both its frequency and strength, and the pluralistic approach taken provided them with a particularly rich source of information. This in turn is reflected in their interpretation of the results, which displays remarkable nuance and detail that would have been impossible to give with a more limited methodological approach.

Although we have come a long way in recent years in the development of a methodological toolbox for convergent evolution, we are likely to see additions continue to appear. This is partly due to an increasing emphasis on the conceptual framework within which we consider convergence (e.g. Arbuckle et al. 2014; Stayton 2015; Speed and Arbuckle 2016), as new ways of thinking about convergence are likely to generate new ways to study it. However, all methods are fraught with caveats and limitations, and attempts to address some of these will almost certainly result in the modification and extension of current methods as well as the design of entirely new methods. This has been a frequent pattern in the development of phylogenetic comparative methods more generally (e.g. Rabosky and Huang 2015), and there is no reason to think that the subset of those which deal specifically with convergent evolution will not show the same trend. Each method will have its own particular set of attributes which might be developed, but many are likely to benefit from modifications to allow fossil data to be incorporated where available as this typically improves inference from comparative methods (Slater et al. 2012). Nevertheless, we now work at a time when the potential for understanding fundamental evolutionary phenomena such as convergence is greater than ever before, and we just have to make good use of the tools available.

References

Arbuckle K, Minter A (2015) Windex: analyzing convergent evolution using the Wheatsheaf index in R. Evol Bioinform 11:11–14

Arbuckle K, Bennett CM, Speed MP (2014) A simple measure of the strength of convergent evolution. Methods Ecol Evol 5:685–693

Beaulieu JM, O'Meara BC, Donoghue MJ (2013) Identifying hidden rate changes in the evolution of a binary morphological character: the evolution of plant habitat in campanulid angiosperms. Syst Biol 62:725–737

Butler MA, King AA (2004) Phylogenetic comparative analysis: a modeling approach for adaptive evolution. American Naturalist 164:683–695

Cott HB (1940) Adaptive coloration in animals. Methuen and Company, London, UK
Cunningham CW, Omland KE, Oakley TH (1998) Reconstructing ancestral character states: a critical reappraisal. Trends Ecol Evol 13:361–366
Currie TE, Meade A (2014) Keeping yourself updated: bayesian approaches in phylogenetic comparative methods with a focus on Markov chain models of discrete character evolution. In Garamszegi LZ (ed) Modern phylogenetic comparative methods and their application in evolutionary biology: concepts and practice, Springer-Verlag, Berlin, Heidelberg
Darwin CR (1859) On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life. John Murray Publishers, London, UK
Doolittle RF (1994) Convergent evolution: the need to be explicit. Trends Biochem Sci 19:15–18
Felsenstein J (1988) Phylogenies and quantitative characters. Annu Rev Ecol Syst 19:445–471
Foote AD, Liu Y, Thomas GWC, Vinař T, Alföldi J, Deng J, Dugan S, van Elk CE, Hunter ME, Joshi V, Khan Z, Kovar C, Lee SL, Lindblad-Toh K, Mancia A et al (2015) Convergent evolution of the genomes of marine mammals. Nat Genet 47:272–275
Friedman ST, Price SA, Hoey AS Wainwright, PC (2016); Ecomorphological convergence in planktivorous surgeonfishes. J Evol Biology 29:965–978
Hansen TF (1997) Stabilizing selection and the comparative analysis of adaptation. Evolution 51:1341–1351
Ingram T, Mahler DL (2013) SURFACE: detecting convergent evolution from comparative data by fitting Ornstein-Uhlenbeck models with stepwise AIC. Methods Ecol Evol 4:416–425
Khabbazian M, Kriebel R, Rohe K, Ané C (2016) Fast and accurate detection of evolutionary shifts in Ornstein-Uhlenbeck models. Methods in Ecol Evolution (Early View)
Losos JB (2011) Convergence, adaptation, and constraint. Evolution 65:1827–1840
Maddison WP, FitzJohn RG (2015) The unsolved challenge to phylogenetic correlation tests for categorical characters. Syst Biol 64:127–136
Maddison WP, Maddison DR (2015) Mesquite: a modular system for evolutionary analysis. Version 3.04. http://mesquiteproject.org
Mahler DL, Ingram T (2014) Phylogenetic comparative methods for studying clade-wide convergence. In Garamszegi LZ (ed.) Modern phylogenetic comparative methods and their application in evolutionary biology: concepts and practice, Springer-Verlag. Berlin, Heidelberg
Mahler DL, Ingram T, Revell LJ, Losos JB (2013) Exceptional convergence on the macroevolutionary landscape in island lizard radiations. Science 341:292–295
McGhee G (2011) Convergent evolution: limited forms most beautiful. Massachusetts Institute of Technology Press, Cambridge, Massachusetts
Mirceta S, Signore AV, Burns JM, Cossins AR, Campbell KL, Berenbrink M (2013) Evolution of mammalian diving capacity traced by myoglobin net surface charge. Science 340:1234192
Muschick M, Indermaur A, Salzburger W (2012) Convergent evolution within an adaptive radiation of cichlid fishes. Curr Biol 22:2362–2368
Pagel M, Meade A, Barker D (2004) Bayesian estimation of ancestral character states on phylogenies. Syst Biol 53:673–684
Paradis E, Claude J, Strimmer K (2004) APE: analyses of phylogenetics and evolution in R language. Bioinformatics 20:289–290
Rabosky DL, Huang H (2015) A robust semi-parametric test for detecting trait-dependent diversification. Syst Biol 2015:syv066
Revell LJ (2012) phytools: an R package for phylogenetic comparative biology (and other things). Methods Ecol Evol 3:217–223
Revell LJ (2014) Graphical methods for visualizing comparative data on phylogenies. In Garamszegi LZ (ed.) Modern phylogenetic comparative methods and their application in evolutionary biology: concepts and practice, Springer-Verlag. Berlin, Heidelberg
Ruxton GD, Sherratt TN, Speed MP (2004) Avoiding attack: the evolutionary ecology of crypsis. Warning Signals and Mimicry. Oxford University Press, Oxford, UK
Slater GJ, Harmon LJ, Alfaro ME (2012) Integrating fossils with molecular phylogenies improves inference of trait evolution. Evolution 66:3931–3944

Speed MP Arbuckle K (2016) Quantification provides a conceptual basis for convergent evolution. Biol Rev (Early View)

Stayton CT (2008) Is convergence surprising? An examination of the frequency of convergence in simulated datasets. J Theor Biol 252:1–14

Stayton CT (2015) The definition, recognition, and interpretation of convergent evolution, and two new measures for quantifying and assessing the significance of convergence. Evolution 69:2140–2153

Ujvari B, Casewell NR, Sunagar K, Arbuckle K, Wüster W, Lo N, O'Meally D, Beckmann C, King GF, Deplazes E, Madsen T (2015) Widespread convergence in toxin resistance by predictable molecular evolution. Proc Natl Acad Sci USA 112:11911–11916

Vidal-García M, Keogh JS (2015) Convergent evolution across the Australian continent: ecotype diversification drives morphological convergence in two distantly related clades of Australian frogs. J Evol Biol 2015:2136–2151

Westneat MW, Alfaro ME, Wainwright PC, Bellwood DR, Grubich JR, Fessler JL, Clements KD, Smith LL (2005) Local phylogenetic divergence and global evolutionary convergence of skull function in reef fishes of the family Labridae. Proc R Soc B 272:993–1000

Winemiller KO (1991) Ecomorphological diversification in lowland freshwater fish assemblages from five biotic regions. Ecol Monogr 61:343–365

Yu Y, Harris AJ, Blair C, He X (2015) RASP (Reconstruct Ancestral State in Phylogenies): a tool for historical biogeography. Mol Phylogenet Evol 87:46–49

Chapter 3
Convergent Evolution Within CEA Gene Families in Mammals: Hints for Species-Specific Selection Pressures

Robert Kammerer, Florian Herse and Wolfgang Zimmermann

Abstract At the genetic level, one of the fastest means to adapt to environmental cues is by gene duplication. Gene duplication is the core process of gene family evolution, which is described by a model called birth-and-death evolution. According to this model, the differences between species are more likely to be detected by comparing gene family expansion and contraction than by comparing sequences of orthologous genes. Consequently, analyzing the structure of gene families may provide deeper insight into selective pressures driving the evolution of a given species. However, tools to analyze the evolution of gene families at a larger scale are not well developed. Nevertheless, recent advances in genome sequencing provide new possibilities to characterize the evolution of gene families more comprehensively and in greater detail. Here, we describe the evolution of the carcinoembryonic antigen (CEA) gene family, which is composed of the CEA-related cell adhesion molecule (CEACAM) and the pregnancy-specific glycoprotein (PSG) gene families. We found that glycosylphosphatidylinositol (GPI)-anchored CEACAMs, paired receptors, and PSG evolved independently at different time points during mammalian evolution. More specifically, we identified several features of the CEACAM/PSG gene family that are the result of convergent evolution in various mammalian species. Possible selection pressures responsible for convergent evolution and their hints toward the function of CEACAM/PSG family members are discussed.

R. Kammerer (✉)
Institute of Immunology, Friedrich-Loeffler-Institut, Federal Research Institute for Animal Health, Greifswald-Insel, Riems, Germany
e-mail: Robert.Kammerer@fli.bund.de

F. Herse
Experimental and Clinical Research Center, The Max-Delbrueck Center for Molecular Medicine and the Charité Medical Faculty, Berlin, Germany

W. Zimmermann
Tumor Immunology Laboratory, LIFE Center, University Clinic of the LMU Munich, Munich, Germany

P. Pontarotti (ed.), *Evolutionary Biology*, DOI 10.1007/978-3-319-41324-2_3

3.1 Introduction

Most multigene families were found to evolve according to a model called birth-and-death evolution. This model was first proposed in 1992 by Nei and Hughes (Nei 1992). In this model, new genes are created by gene duplication: Some of these genes are maintained and fixed in the genome, while others are deleted or become nonfunctional by deleterious mutations (Nei and Rooney 2005). Once a gene is duplicated, it may diversify to gain new functions or may only increase the gene dosage. In most cases, a functional diversification, including change in spatiotemporal expression pattern [e.g., Hox genes (Lonfat and Duboule 2015)], variation in ligand binding specificity (e.g., chemoreceptor genes (Benton 2015), or a change in signaling capacity, may occur (Sanderson et al. 2014). Understanding why some genes are fixed while other are deleted means to understand natural selection shaping the structure and composition of the gene family (Eirin-Lopez et al. 2012). Four-fifths of the proteins in eukaryotes are multidomain proteins underlining the importance of combinatorial variation for the creation of protein diversity (Buljan and Bateman 2009). At the genetic level, the basic mechanism for such a combinatorial diversification is exon shuffling (Patthy 1999). The immunoglobulin (Ig) superfamily is one of the largest and most diverse groups of proteins in humans, which thus is especially well suited for productive exon shuffling. This is due to the fact that exons coding for immunoglobulin domains have the same reading frame and, therefore, can be freely combined in various ways. In addition, intron sequences in genes coding for members of the immunoglobulin superfamily are relatively large compared with the protein-coding exon sequences favoring recombination without destroying protein-coding sequences (Buljan and Bateman 2009; Patthy 1999).

The carcinoembryonic antigen (CEA) gene family, which belongs to the immunoglobulin superfamily, comprises the CEA-related cell adhesion molecule (CEACAM) and the pregnancy-specific glycoprotein (PSG) gene families. Since CEACAMs and PSGs are involved in immunity and reproduction, it is not surprising that this gene family is a hot spot for genetic variation in the genome of vertebrates (Chuong et al. 2010; Mouse Genome Sequencing et al. 2002; Trowsdale and Parham 2004). Analyzing the evolution of such hot spot genes may give new insights into the mechanisms of gene family evolution. Furthermore, we will focus in this article on evolutionary convergence of the CEA gene family in distantly related linages in order to get hints for selective pressures acting on the CEA gene family evolution. Detecting multiple origins of similar adaptations within a gene family will provide exceptional insights into functional constraints and evolutionary pressures during adaptation (Christin et al. 2010).

3.2 The CEA Gene Family

The CEA gene family is located in the expanded leukocyte receptor complex (LRC), which is found in humans on chromosome 19. The expanded LRC contains besides the LRC, the CEA/PSG gene family, the SIGLEC gene family, and genes for adaptor signaling proteins (Barrow and Trowsdale 2008). The CEA family is a member of the immunoglobulin superfamily with an extracellular Ig domain architecture usually containing N-terminal Ig variable-like (IgV-like) domains followed by Ig constant-like (IgC-like) domains. Historically, the CEA family is divided into the CEA-related cell adhesion molecule (CEACAM) and the pregnancy-specific glycoprotein (PSG) subgroups (Von Kleist 1992). From an evolutionary point of view, the CEA gene family consists of up to five genes (*CEACAM1*, *CEACAM16*, *CEACAM18*, *CEACAM19*, and *CEACAM20*) for which orthologs can be identified in other vertebrates (Zebhauser et al. 2005). While *CEACAM19* can even be identified in reptiles by sequence similarities, *CEACAM16*, *CEACAM18*, and *CEACAM20* are not found outside of the mammalian order (Pavlopoulou and Scorilas 2014, Kammerer unpublished). With respect to domain composition and the presence of signaling motifs, the most ancestral member of the CEA gene family is most likely an CEACAM1-like gene (Kammerer and Zimmermann 2010; Pavlopoulou and Scorilas 2014). Therefore, an ancestor of the CEACAM1 gene is thought to represent the founder gene of the CEA/PSG gene family. According to sequence similarities of the Ig domains, amplification of *CEACAM1* leads to the modern CEA gene families composed of the previously mentioned genes and a different number of CEACAM1-related paralogs which includes CEACAMs and PSGs (Kammerer and Zimmermann 2010). Figure 3.1 exemplarily depicts the composition and nomenclature of the human, mouse, and dog CEA gene families.

3.3 Expansion and Contraction of the CEA Gene Family in Mammals

According to the birth-and-death model of gene family evolution, gene duplication may happen randomly but the gain of new functions or pseudogenization of the multiplied genes is driven by selection. This implies that a given gene family may expand in one species, but not in others at a certain time point (Demuth and Hahn 2009). Change in gene family size is favored when selection pressures are high. Gene family expansion may lead to gene diversification or increase in gene dosage. If diversification is required, positive selection favors nonsynonymous nucleotide mutations that lead to amino acid changes in the encoded protein. Thus, selection of diversification can be identified by detecting positive selection that is characterized by a ratio of the rate of nonsynonymous to the rate of synonymous mutations >1 (Ellegren 2008). However, selection for diversification can also be achieved at

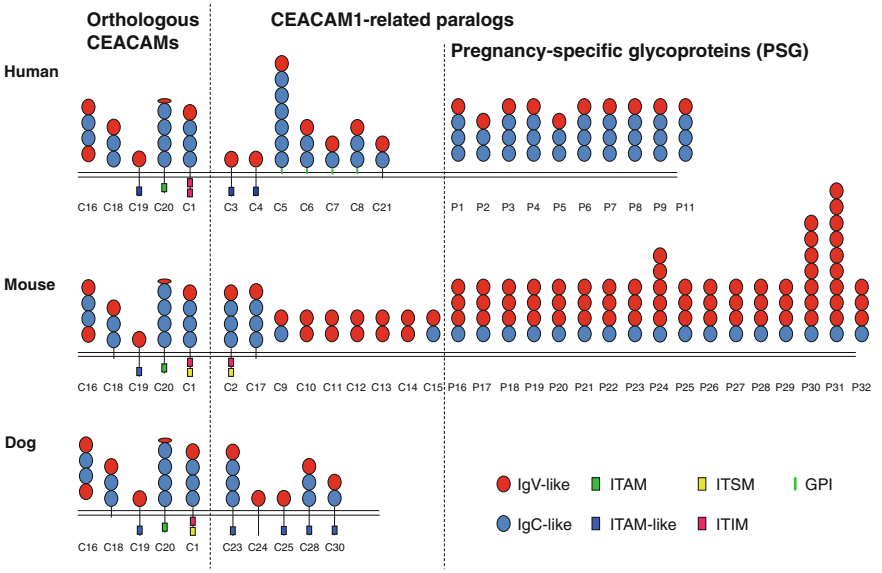


Fig. 3.1 Domain organization of mammalian CEA family members. The domain organization of CEA family members from selected species was retrieved from public databases at NCBI. The conserved orthologous members and the members expressed predominantly in trophoblast cells of the placenta are separated by *dashed lines*. The predicted signaling motifs in the cytoplasmic domains are schematically shown as *green* (ITAM), blue (ITAM-like motif, no acidic amino acid present at position -1 to -3 from first Y in consensus motif E/Dx0 2YxxL/Ix6 8YxxL/I), *red* (ITIM), and *yellow boxes* (ITSM). GPI anchorage is depicted as a *green line*. C, CEACAM; P, PSG

several other levels of gene function, including distinct spatiotemporal expression, membrane anchorage, and signaling capacity of the duplicated gene products. On the other hand, in times where the selective pressure is low, dispensable genes and nonfunctional genes such as pseudogenes may be deleted from the entire genome. In some cases, dispensable genes may also acquire new functions and may be fixed in the genome. However, in general, weak natural selection is most likely accompanied by reduction in gene family size. Such rounds of expansion and contraction of a gene family may have occurred several times during evolution in individual species. Whether a gene family actually expands or whether it is in a phase of purifying contraction may be revealed by comparing the number of functional genes with the number of pseudogenes within the gene family. High numbers of pseudogenes are indicative for a recent expansion. If enough time had passed since expansion, most of the pseudogenes that were coamplified with functional genes would have been already eliminated from the genome. One would assume that periodical or time-limited selection rather than continuous selection may lead to an enormous variability of a gene family within a group of organisms. Indeed, comparison of the CEA gene families in vertebrates and mammals demonstrates that the size of the CEA gene families in mammals differs

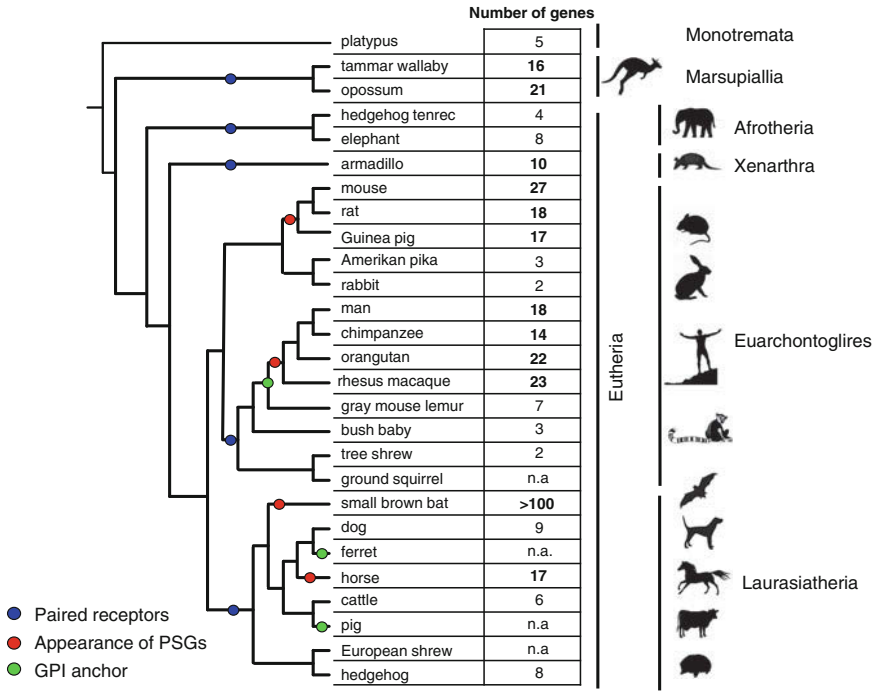


Fig. 3.2 Evolution of the CEACAM gene family in mammals. The phylogenetic and taxonomic relationship of selected mammalian species is shown schematically. The appearance of various features during radiation is marked by colored dots: *blue* paired receptors; *red* PSGs; *green* GPI anchor. The number of CEACAM genes identified and/or predicted is indicated

considerably from less than five genes in rabbits to more than hundred in some bat species (Fig. 3.2; (Kammerer and Zimmermann 2010; Zhang et al. 2013). Zhang et al. observed that the CEACAM genes in microbats expanded after separation from megabats about 68 million years ago (Zhang et al. 2013). Our own unpublished analyses of the size of the CEA gene family in various microbat species indicate that this CEACAM expansion in microbats is still ongoing. These expanded CEACAM1-related genes exhibit some similarities to PSGs in primates and rodents; however, a large fraction of these genes seems to represent pseudogenes (Kammerer unpublished). Taken together, the recent expansion of CEACAMs in microbats supports the view that the CEA gene family is subject of multiple rounds of selection in certain if not most vertebrate species in which this gene family exists.

3.4 CEACAM1 Paralogs

CEACAM1, the ancestor of which is thought to be the founder gene of the CEA/PSG gene family, is composed of an N-terminal IgV-like domain exon, two (cattle and pigs) or three IgC-like domain exons in the order (A1-A2, or A1-B-A2), a transmembrane domain exon, and three exons encoding a cytoplasmic tail that contains two immunoreceptor tyrosine-based motifs; at least one of them is an immunoreceptor tyrosine-based inhibition motif (ITIM), which transduces inhibitory signals (Huang et al. 2015; Kammerer et al. 2004). In humans and some other species, both motifs represent inhibitory motifs. However, in most species, the second, membrane distal motif is a so-called switch motif (Kammerer et al. 2007; Kammerer and Zimmermann 2010), which can transduce both inhibitory and activation signals depending on the cellular context. Remarkably, duplication of CEACAM1 did not lead to amplification of these "inhibitor receptors"; only in a few species, two ITIM containing CEACAMs were found (Kammerer and Zimmermann 2010). In all other species, a single inhibitory receptor is maintained within the CEA family. All other paralogs of CEACAM1 were subject to various modifications leading to molecules that have different signaling properties compared to CEACAM1 (Chang et al. 2013; Kammerer and Zimmermann 2010; Pavlopoulou and Scorilas 2014). The extracellular parts of all CEACAM1 paralogs, so far found in mammals, are composed of structural units already present in CEACAM1, namely IgV-like domains and IgC-like domains of the A and B subtype (Chang et al. 2013; Kammerer and Zimmermann 2010). The extracellular parts of CEACAM1 paralogs were found to interact with various ligands of cellular, bacterial, and viral origin (Dveksler et al. 1993; Sadarangani et al. 2011; Singer et al. 2014; Tchoupa et al. 2014). However, CEACAM1 paralogs may considerably differ from CEACAM1 in respect to their membrane anchorage and signaling capacity. For example, PSGs are secreted proteins missing a transmembrane domain. Some paralogs were combined with mutated transmembrane exons leading to a loss of cytoplasmic signaling domains or to GPI anchorage. Furthermore, some CEACAM1 paralogs have gained completely new transmembrane domains either together with or without cytoplasmic signaling domains (Kammerer unpublished). Most remarkably, the newly gained cytoplasmic signaling units encode immunoreceptor tyrosine-based activation motifs (ITAM). Thus, the CEACAM1 paralogs are composed of similar extracellular, ligand-binding domains as found in CEACAM1. Furthermore, CEACAM1 shows an extraordinary broad expression pattern, including immune cells, endothelial cells, and epithelial cells (Hammarstrom 1999). In contrast, CEACAM paralogs show a restricted expression pattern often limited to a specific cell type. For example, human ITAM-harboring CEACAM3 is exclusively expressed by granulocytes (Nagel et al. 1993; Schmitter et al. 2007), and PSGs are specifically expressed by trophoblast cells (Kromer et al. 1996). Thus, the regulation of CEACAM1 paralog expression adds an additional level of complexity to the CEACAM/PSG system with its structural diversity.

3.5 GPI-Anchored CEACAMs

In humans, several *CEACAM1* paralogs originally generated by duplication of an ancestral CEACAM1 contain the CEACAM1-typical transmembrane exon, however, slightly modified. This changed the transmembrane domain to a signal peptide that conveys after cleavage a post-translational modification resulting in anchorage of the receptor to the cell surface via a GPI anchor (Thompson et al. 1994). GPI anchorage of surface molecules is very common, but the signaling capacities associated with this kind of membrane fixation are not well understood (Maeda and Kinoshita 2011). Nevertheless, GPI anchors associate with membrane microdomains leading to distinct signaling capacities (Nicholson and Stanners 2006). Interestingly, the microdomain association is determined by sequences within the signal peptide; therefore, specific signal peptides may result in different signaling mechanisms into the cell (Nicholson and Stanners 2007). On the other hand, GPI anchors can be cut by phospholipases, and therefore, GPI-linked proteins can be released from the cell surface in a controlled manner. Thus, one can look at the GPI-anchored molecules as secreted molecules that are retained at the cell surface for a controlled period of time. This may have profound consequences for the cells expressing these molecules. For example, expression of GPI-linked CEACAMs may lead to anchorage-independent growth of tumor cells, to the regulation of tissue architecture, and to control of commensal bacteria. For a long time, it was a mystery why GPI-anchored CEACAMs were only found in primates. Interestingly, Stanner's group found that in New World monkeys, a set of mutations within the ancestral CEACAM1 transmembrane domain, which are different from those found in Old World monkeys, led to GPI-anchored CEACAMs, indicating an independent convergent evolution of the GPI anchor in these two primate groups (Naghibalhossaini et al. 2007). Surprisingly, we identified GPI-anchored CEACAMs in pigs, but not in other artiodactyls. More recently, we identified CEACAM sequences derived from carnivores such as mustelids and seals that strongly indicate that the corresponding proteins are GPI-linked. Such sequences are not present in the genome of dogs. Most remarkably, the GPI anchorage of these CEACAMs was generated by mutation of the ITAM-typical transmembrane exon. Further investigations are needed to correlate special functional requirements with the appearance of GPI-anchored CEACAMs.

3.6 Appearance of Paired Receptors Within the CEA Gene Family

Various families of immune receptors such as natural killer (NK) inhibitory receptor families contain so-called paired receptors (Akkaya and Barclay 2013). Paired receptors are receptors that have similar extracellular ligand-binding domains but transduce contrary signals into the cell, i.e., inhibitory and activating signals. The stimulatory signals are transmitted by adaptor proteins that harbor

activating motifs and associate with the immune receptors. Remarkably, receptors of different receptor families compete for the same adaptor proteins, thereby influencing and most likely regulating each other and the signal transmitted by them into the cell (Kuroki et al. 2012). This is not the case in the CEA gene family. Receptors of the CEA family carry the signaling motifs, inhibitory as well as the activating motifs, within their cytoplasmic tails and, therefore, can signal independently of coexpressed (competing) receptors of other receptor families (Kammerer et al. 2007; Kammerer and Zimmermann 2010). While the CEA gene family in all mammals investigated so far contains receptors with inhibitory motifs, receptors with activating signals are found in most but not all mammals (http://www.carcinoembryonic-antigen.de/). This again points to CEACAM1 being the founder gene of the CEA family. It is very intriguing that in species where activating receptors exist, the extracellular domains of the ITAM-bearing CEACAMs are very similar to the extracellular part of the inhibitor receptor, indicating that these receptors may function as paired receptors (Kammerer et al. 2007; Kammerer and Zimmermann 2010). The evolutionary pressure that leads to the fixation of activating receptors in mammals is thought to be due to the usage of the inhibitor receptor by pathogens to infect their hosts and at the same time suppress their immune response. As a counteraction of the host, receptors evolved that bind the same pathogens but then activate the immune response of the host. That this is indeed the case is supported by investigations in humans (Gray-Owen and Blumberg 2006) where several pathogenic bacteria can bind to CEACAM1 on immune cells and thereby are able to suppress the immune response of the host (Slevogt et al. 2008). On the other hand, a large body of evidence has accumulated, indicating that the activating receptor CEACAM3 binds to the same pathogens. In fact, CEACAM3 is found to be specifically expressed on granulocytes and upon binding to Neisseria activates granulocytic pathways involving pathogen uptake and destruction (Schmitter et al. 2007). Since in various groups of mammals activating receptors exist, the first receptor pair evolved presumably very early during mammalian evolution. Tracking the amplification of CEACAM1-related paralogs based on the involved transmembrane domain exons revealed that in some mammalian linages, inhibitory receptor genes were amplified, while in others activating receptor genes were multiplied.

3.7 Evolution of Pregnancy-Specific Glycoproteins (PSGs)

In humans, a group of closely related glycoproteins were identified as the most abundant fetal proteins in the maternal blood at late pregnancy (Bohn 1971; Lin et al. 1974; Tatarino and Masyukev 1970). Subsequently, they were named pregnancy-specific glycoproteins (PSGs). Expression of PSGs could be localized in the syncytiotrophoblast (Horne et al. 1976). Later, it was found that PSG genes belong to the CEA gene family and that they form a subgroup within the CEACAM1-related CEACAMs (Kammerer and Zimmermann 2010; Watanabe and

Chou 1988). Originally, PSG genes were only identified in humans, since orthologous genes could not be unambiguously identified in other species. However, in rodents like rats and mice, a structurally different group of genes was identified within these CEA gene families, which showed also a trophoblast-specific expression pattern (Rebstock et al. 1990; Rudert et al. 1992). However, rodent PSG genes were found to be localized at a locus in the genome different from the one previously found for human PSG genes (McLellan et al. 2005a). The human and mouse PSG genes are clustered on chromosome 19q13 and proximal chromosome 7, respectively. There are eleven human protein-coding PSG genes named PSG1-PSG11 (PSG10 is in some humans a nonprotein-coding pseudogene), while in the mouse 17 PSGs named PSG16-PSG32 exist. The structure of PSG proteins varies considerably among species. Human PSGs consist of one N-terminal IgV-like domain (N domain) followed by one or two IgC-like domains of the A-type (A1 and A2) and one B-type IgC-like domain. The C-terminal end of the protein is built of a relatively short hydrophilic tail (Teglund et al. 1995). Rodent PSGs contain typically three or more N domains at the N-terminus and a single IgC-like domain (A-type) at the C-terminus (McLellan et al. 2005a). The rat has eight PSGs; seven of them have the N1-N2-N3-A domain arrangement, and only PSG36 contains five N domains (McLellan et al. 2005a). Fourteen mouse PSGs have three N domains and one A domain. The remaining three PSGs have, instead of three N domains, 5, 7, or 8 N domains (McLellan et al. 2005a). Interestingly, a recent search in the hamster genome revealed that at least three PSGs exist. All of them contain three N domains, thus indicating that rodent PSG genes are derived from a common ancestor (Kammerer unpublished). On the other hand, the different structural properties of rodent and human PSGs indicate that PSGs evolved independently in rodents and primates. Thus, the question arose whether PSGs are not only expressed by the same tissue, i.e., trophoblast, but also have similar functional properties. One way to identify the function of a protein family may involve the analysis of pathologies associated with abnormal protein expression. Several reports indicate that abnormal PSG levels may be associated with growth restrictions of the fetus (Grudzinskas et al. 1983; Pihl et al. 2009; Tatra et al. 1974; Towler et al. 1977; Zhao et al. 2012). In addition, a correlation between low PSG levels and pre-eclampsia was found in some studies, but not in others. Indeed, in most studies, the identification of individual PSGs may be problematic since the PSGs are very closely related and several tools used to investigate PSG levels may not be specific for individual PSGs. More recently, a correlation between PSG9 and putatively PSG5 and PSG2 in the blood of women with early-onset pre-eclampsia was described (Blankley et al. 2013). In this study, a method called selected reaction monitoring (SRM) mass spectroscopy was used. This method has the advantage that the peptide sequence could be clearly identified as a specific gene product. Remarkably, it was found that peptides from other PSGs, such as PSG1, 2, 3, 6, and 11, did not differ between cases and controls. These findings for individual PSGs may explain earlier contradictory results about the relation between PSG levels and pre-eclampsia cases. In addition, recently, it was found that low levels of PSG4 correlate with gestational diabetes mellitus (Bari et al. 2016; Zhao et al. 2015).

These reports do not clearly point to a certain function of PSGs, but underline the importance of PSGs for a successful pregnancy. The complexity of the PSG family in mice has hindered the creation of PSG-knockout mice that would have provided additional hints to the function of PSG. Interestingly, knockout mice in which single PSG-like CEACAM was deleted did not show a severe phenotype, indicating that certain PSG-like CEACAMs can functionally replace each other (Finkenzeller et al. 2003; Finkenzeller et al. 2000).

3.8 Similarities of the Placenta in Primates and Rodents

Humans and rodents have a hemochorial placenta in which fetal cells have a direct contact with maternal blood. Such an intimate contact with the mother's immune system may require potent mechanisms to control the immune system of the mother in order to protect the allogenic fetus. It has been proposed that immune cells can recognize pregnancy and that their reactivity is switched to a less harmful phenotype. Indeed in recent years, multiple mechanisms of immune modulation have been described. One protein family that may contribute to the immune modulation that is required for a successful pregnancy is the PSG family. Early investigations have demonstrated that PSGs can induce the secretion of anti-inflammatory cytokines from monocytes and macrophages; in particular, it was demonstrated that mouse PSG18 induces the secretion of IL-10 by murine peritoneal macrophages (Wessells et al. 2000). Thereafter, it was shown that PSGs exhibit cross-species activity; i.e., human PSG1, PSG6, and PSG11 can induce the secretion of interleukin (IL)-10, IL-6, and transforming growth factor (TGF)-beta 1 from both human monocytes and murine RAW 264.7 cells (Snyder et al. 2001). Furthermore, the N-terminal IgV-like domain of both murine PSG18 and human PSG6 was sufficient to induce the cytokine release from the stimulated cells (Snyder et al. 2001; Wessells et al. 2000). Subsequently, Motran and colleagues found that human PSG1a induces an alternative activation of human and mouse monocytes in vitro as well as in vivo to enhance the Th2-type immune response (Motran et al. 2002; Motran et al. 2003). A molecular basis of the observed effects of PSGs on macrophages was discovered by the identification of CD9 as a receptor for PSG17 and PSG19 (Waterhouse et al. 2002), and it was indeed confirmed that the interaction of PSG17 with CD9 was necessary for cytokine induction (Ha et al. 2005). In the meantime, additional receptors for PSGs were identified, such as glycosaminoglycans and integrins (Moore and Dveksler 2014; Sulkowski et al. 2011), which may in part explain the similar functions of murine and human PSG on both human and murine cells despite the fact that PSG17 and PSG19 are the only PSGs that were so far found to interact with CD9. More recently, it was observed that human PSGs interact with dendritic cells (DC). PSG1a drastically shifts the phenotype of DC to semi-mature DC, which promotes the enrichment of T cells secreting Th2 cytokines and IL-17 as well as of Treg cells (Martinez et al. 2013; Martinez et al.

2012). These findings demonstrate that PSGs can indirectly regulate T-cell activity via regulating the DC phenotype.

Extensive and highly regulated tissue remodeling is required to establish intimate contact between mother and fetus without destroying the function of the maternal tissue which is needed to guarantee nutrient supply. An essential part of tissue remodeling comprises the generation of new blood vessels. First, indications that PSGs may play a role in neovascularization of the placenta were reported by Wynne and colleagues. They found that mouse PSG is associated with, but not expressed by, the maternal endothelium (Wynne et al. 2006). This finding implies that endothelial cells express a ligand for PSG. In addition, Wu from Dveksler's laboratory observed that recombinant PSG23 induced the secretion of proangiogenic factors such as transforming growth factor-beta (TGF-β) and vascular endothelial growth factor (VEGF) in various cell types, including macrophages dendritic cells, endothelial cells, and trophoblasts. Remarkably, they observed even some cross-reactivity with human monocytes and trophoblast cells (Wu et al. 2008). More recently, it was found that human PSG1 as well as murine PSG22 and PSG23 induces endothelial tube formation on Matrigel and type I collagen (Blois et al. 2012; Lisboa et al. 2011). Together, these findings strongly suggest that despite the independent evolution and expansion of human and rodent PSGs, members of both families have a conserved function in regulating angiogenesis and tissue remodeling.

Infiltration of fetal tissues into maternal blood vessels may bear the risk of an enhanced thrombotic activity. The presence of the tripeptide motif Arg-Gly-Asp (RGD) in the N domain of most human PSGs and a related motif (KGD) on murine PSGs led to the hypothesis that PSGs may have a similar function like disintegrin proteins of snake venoms. Disintegrins are able to disrupt the interaction of integrins with extracellular matrix components or blood clotting mechanisms. Therefore, it was suggested that PSGs may function as soluble integrin ligands, thereby facilitating the invasion of fetal tissue and presumably reducing the thrombotic activity (McLellan et al. 2005b; Zhou and Hammarstrom 2001). Indeed, Shanley and colleagues recently reported that human PSG1, human PSG9, and murine PSG23, but not soluble CEACAM1, bind to αIIbβ3 integrin. This integrin is expressed by platelets, and consequently, the aforementioned PSGs inhibited platelet–fibrinogen interaction (Shanley et al. 2013).

3.9 Concluding Remarks

Convergent evolution of gene families in distantly related linages may be due to adaptation to similar environmental cues. Here, we show that convergent evolution of the CEA family could be observed at multiple structural levels, i.e., expansion of activating receptors, GPI linkage to the membrane, and the generation of secreted glycoproteins. The most striking example for convergent evolution is observed within the PSG subgroup. As discussed above, PSGs evolved independently in

primates and rodents. Nevertheless, many functions are shared by human and murine PSGs. Previous studies using published genomes did not identify PSGs in species other than primates and rodents (Chang et al. 2013; Pavlopoulou and Scorilas 2014). However, due to the considerable difference in PSG structure and genome localization, it is of particular importance to define which CEACAMs should be classified as PSGs. The main criterion is the trophoblast-specific expression, but in some species, e.g., protected wildlife species from which tissue samples are difficult to obtain, this criterion is not applicable. Therefore, new strategies to identify PSGs have to be defined. Since all PSGs are secreted proteins, this is a prerequisite to classify a CEACAM as PSG (Fig. 3.1 and 3.2). Secreted CEACAMs in the bat are composed of a single N domain followed or not by an A domain. Furthermore, secreted equine CEACAMs are built of a single N domain followed by a truncated A domain (Kammerer unpublished). Indeed, our own preliminary results indicate that secreted horse CEACAMs are expressed by trophoblast cells, suggesting that these CEACAMs are PSGs. Horse PSGs are structurally different form both human PSGs and rodent PSGs again pointing to independent evolution. The reason why PSGs are recruited in different species for the aforementioned functions is unknown. However, a striking characteristic of PSGs is that PSGs are the most rapidly evolving proteins expressed in a trophoblast-specific manner (Chuong et al. 2010). In addition, the human PSG gene cluster is a hot spot of sequence variation with an unusual number of SNPs as well as with extensive structural variation, such as deletions, insertions, duplications, and inversions (Chang et al. 2013; Sudmant et al. 2015). Therefore, one may hypothesize that PSGs represent a highly variable system of interaction molecules prone to rapid adaption to environmental cues.

References

Akkaya M, Barclay AN (2013) How do pathogens drive the evolution of paired receptors? Eur J Immunol 43:303–313

Bari MF, Ngo S, Bastie CC, Sheppard AM, Vatish M (2016) Gestational diabetic transcriptomic profiling of micro-dissected human trophoblast. J Endocrinol 229(1):47–59

Barrow AD, Trowsdale J (2008) The extended human leukocyte receptor complex: diverse ways of modulating immune responses. Immunol Rev 224:98–123

Benton R (2015) Multigene family evolution: perspectives from insect chemoreceptors. Trends Ecol Evol 30:590–600

Blankley RT, Fisher C, Westwood M, North R, Baker PN, Walker MJ, Williamson A, Whetton AD, Lin W, McCowan L, Roberts CT, Cooper GJ, Unwin RD, Myers JE (2013) A label-free selected reaction monitoring workflow identifies a subset of pregnancy specific glycoproteins as potential predictive markers of early-onset pre-eclampsia. Mol Cell proteomics 12:3148–3159

Blois SM, Tirado-Gonzalez I, Wu J, Barrientos G, Johnson B, Warren J, Freitag N, Klapp BF, Irmak S, Ergun S, Dveskler GS (2012) Early expression of pregnancy-specific glycoprotein 22 (PSG22) by trophoblast cells modulates angiogenesis in mice. Biol Reprod 86:191

Bohn H (1971) Detection and characterization of pregnancy proteins in the human placenta and their quantitative immunochemical determination in sera from pregnant women. Archiv fur Gynakologie 210:440–457

Buljan M, Bateman A (2009) The evolution of protein domain families. Biochem Soc Trans 37:751–755

Chang CL, Semyonov J, Cheng PJ, Huang SY, Park JI, Tsai HJ, Lin CY, Grutzner F, Soong YK, Cai JJ, Hsu SY (2013) Widespread divergence of the CEACAM/PSG genes in vertebrates and humans suggests sensitivity to selection. PLoS ONE 8:e61701

Christin PA, Weinreich DM, Besnard G (2010) Causes and evolutionary significance of genetic convergence. Trends in genetics: TIG 26:400–405

Chuong EB, Tong W, Hoekstra HE (2010) Maternal-fetal conflict: rapidly evolving proteins in the rodent placenta. Mol Biol Evol 27:1221–1225

Demuth JP, Hahn MW (2009) The life and death of gene families. BioEssays: news and reviews in molecular, cellular and developmental biology 31:29–39

Dveksler GS, Pensiero MN, Dieffenbach CW, Cardellichio CB, Basile AA, Elia PE, Holmes KV (1993) Mouse hepatitis virus strain A59 and blocking antireceptor monoclonal antibody bind to the N-terminal domain of cellular receptor. Proc Natl Acad Sci USA 90:1716–1720

Eirin-Lopez JM, Rebordinos L, Rooney AP, Rozas J (2012) The birth-and-death evolution of multigene families revisited. Genome Dyn 7:170–196

Ellegren H (2008) Comparative genomics and the study of evolution by natural selection. Mol Ecol 17:4586–4596

Finkenzeller D, Fischer B, McLaughlin J, Schrewe H, Ledermann B, Zimmermann W (2000) Trophoblast cell-specific carcinoembryonic antigen cell adhesion molecule 9 is not required for placental development or a positive outcome of allotypic pregnancies. Mol Cell Biol 20:7140–7145

Finkenzeller D, Fischer B, Lutz S, Schrewe H, Shimizu T, Zimmermann W (2003) Carcinoembryonic antigen-related cell adhesion molecule 10 expressed specifically early in pregnancy in the decidua is dispensable for normal murine development. Mol Cell Biol 23:272–279

Gray-Owen SD, Blumberg RS (2006) CEACAM1: contact-dependent control of immunity. Nat Rev Immunol 6:433–446

Grudzinskas JG, Gordon YB, Menabawey M, Lee JN, Wadsworth J, Chard T (1983) Identification of high-risk pregnancy by the routine measurement of pregnancy-specific beta 1-glycoprotein. Am J Obstet Gynecol 147:10–12

Ha CT, Waterhouse R, Wessells J, Wu JA, Dveksler GS (2005) Binding of pregnancy-specific glycoprotein 17 to CD9 on macrophages induces secretion of IL-10, IL-6, PGE2, and TGF-beta1. J Leukoc Biol 77:948–957

Hammarstrom S (1999) The carcinoembryonic antigen (CEA) family: structures, suggested functions and expression in normal and malignant tissues. Semin Cancer Biol 9:67–81

Horne CH, Towler CM, Pugh-Humphreys RG, Thomson AW, Bohn H (1976) Pregnancy specific beta1-glycoprotein–a product of the syncytiotrophoblast. Experientia 32:1197

Huang YH, Zhu C, Kondo Y, Anderson AC, Gandhi A, Russell A, Dougan SK, Petersen BS, Melum E, Pertel T, Clayton KL, Raab M, Chen Q, Beauchemin N, Yazaki PJ, Pyzik M, Ostrowski MA, Glickman JN, Rudd CE, Ploegh HL, Franke A, Petsko GA, Kuchroo VK, Blumberg RS (2015) CEACAM1 regulates TIM-3-mediated tolerance and exhaustion. Nature 517:386–390

Kammerer R, Zimmermann W (2010) Coevolution of activating and inhibitory receptors within mammalian carcinoembryonic antigen families. BMC Biol 8:12

Kammerer R, Popp T, Singer BB, Schlender J, Zimmermann W (2004) Identification of allelic variants of the bovine immune regulatory molecule CEACAM1 implies a pathogen-driven evolution. Gene 339:99–109

Kammerer R, Popp T, Hartle S, Singer BB, Zimmermann W (2007) Species-specific evolution of immune receptor tyrosine based activation motif-containing CEACAM1-related immune receptors in the dog. BMC Evol Biol 7:196

Kromer B, Finkenzeller D, Wessels J, Dveksler G, Thompson J, Zimmermann W (1996) Coordinate expression of splice variants of the murine pregnancy-specific glycoprotein (PSG) gene family during placental development. Eur J Biochem/FEBS 242:280–287

Kuroki K, Furukawa A, Maenaka K (2012) Molecular recognition of paired receptors in the immune system. Frontiers Microbiol 3:429

Lin TM, Halbert SP, Spellacy WN (1974) Measurement of pregnancy-associated plasma proteins during human gestation. J Clin Investig 54:576–582

Lisboa FA, Warren J, Sulkowski G, Aparicio M, David G, Zudaire E, Dveksler GS (2011) Pregnancy-specific glycoprotein 1 induces endothelial tubulogenesis through interaction with cell surface proteoglycans. J Biol Chem 286:7577–7586

Lonfat N, Duboule D (2015) Structure, function and evolution of topologically associating domains (TADs) at HOX loci. FEBS Lett 589:2869–2876

Maeda Y, Kinoshita T (2011) Structural remodeling, trafficking and functions of glycosylphosphatidylinositol-anchored proteins. Prog Lipid Res 50:411–424

Martinez FF, Knubel CP, Sanchez MC, Cervi L, Motran CC (2012) Pregnancy-specific glycoprotein 1a activates dendritic cells to provide signals for Th17-, Th2-, and Treg-cell polarization. Eur J Immunol 42:1573–1584

Martinez FF, Cervi L, Knubel CP, Panzetta-Dutari GM, Motran CC (2013) The role of pregnancy-specific glycoprotein 1a (PSG1a) in regulating the innate and adaptive immune response. Am J Reprod Immunol 69:383–394

McLellan AS, Fischer B, Dveksler G, Hori T, Wynne F, Ball M, Okumura K, Moore T, Zimmermann W (2005a) Structure and evolution of the mouse pregnancy-specific glycoprotein (Psg) gene locus. BMC Genom 6:4

McLellan AS, Zimmermann W, Moore T (2005b) Conservation of pregnancy-specific glycoprotein (PSG) N domains following independent expansions of the gene families in rodents and primates. BMC Evol Biol 5:39

Moore T, Dveksler GS (2014) Pregnancy-specific glycoproteins: complex gene families regulating maternal-fetal interactions. Int J Dev Biology 58:273–280

Motran CC, Diaz FL, Gruppi A, Slavin D, Chatton B, Bocco JL (2002) Human pregnancy-specific glycoprotein 1a (PSG1a) induces alternative activation in human and mouse monocytes and suppresses the accessory cell-dependent T cell proliferation. J Leukoc Biol 72:512–521

Motran CC, Diaz FL, Montes CL, Bocco JL, Gruppi A (2003) In vivo expression of recombinant pregnancy-specific glycoprotein 1a induces alternative activation of monocytes and enhances Th2-type immune response. Eur J Immunol 33:3007–3016

Nagel G, Grunert F, Kuijpers TW, Watt SM, Thompson J, Zimmermann W (1993) Genomic organization, splice variants and expression of CGM1, a CD66-related member of the carcinoembryonic antigen gene family. Euro J Biochem/FEBS 214:27–35

Naghibalhossaini F, Yoder AD, Tobi M, Stanners CP (2007) Evolution of a tumorigenic property conferred by glycophosphatidyl-inositol membrane anchors of carcinoembryonic antigen gene family members during the primate radiation. Mol Biol Cell 18:1366–1374

Nei M, Hughes AL (1992) Balanced polymorphism and evolution by the birth-and-death process in the MHC loci. In: Tsuji MA, Sasazuki T (eds) 11th histocompatibility workshop and conference. Oxford University Press, Oxford, pp 27–38

Nei M, Rooney AP (2005) Concerted and birth-and-death evolution of multigene families. Annu Rev Genet 39:121–152

Nicholson TB, Stanners CP (2006) Specific inhibition of GPI-anchored protein function by homing and self-association of specific GPI anchors. J Cell Biology 175:647–659

Nicholson TB, Stanners CP (2007) Identification of a novel functional specificity signal within the GPI anchor signal sequence of carcinoembryonic antigen. J Cell Biology 177:211–218

Patthy L (1999) Genome evolution and the evolution of exon-shuffling–a review. Gene 238:103–114

Pavlopoulou A, Scorilas A (2014) A comprehensive phylogenetic and structural analysis of the carcinoembryonic antigen (CEA) gene family. Genome Biol Evol 6:1314–1326

Pihl K, Larsen T, Laursen I, Krebs L, Christiansen M (2009) First trimester maternal serum pregnancy-specific beta-1-glycoprotein (SP1) as a marker of adverse pregnancy outcome. Prenat Diagn 29:1256–1261

Rebstock S, Lucas K, Thompson JA, Zimmermann W (1990) cDNA and gene analyses imply a novel structure for a rat carcinoembryonic antigen-related protein. J Biol Chem 265:7872–7879

Rudert F, Saunders AM, Rebstock S, Thompson JA, Zimmermann W (1992) Characterization of murine carcinoembryonic antigen gene family members. Mamm Genome: Off\J Int Mamm Genome Soc 3:262–273

Sadarangani M, Pollard AJ, Gray-Owen SD (2011) Opa proteins and CEACAMs: pathways of immune engagement for pathogenic Neisseria. FEMS Microbiol Rev 35:498–514

Sanderson ND, Norman PJ, Guethlein LA, Ellis SA, Williams C, Breen M, Park SD, Magee DA, Babrzadeh F, Warry A, Watson M, Bradley DG, MacHugh DE, Parham P, Hammond JA (2014) Definition of the cattle killer cell Ig-like receptor gene family: comparison with aurochs and human counterparts. J Immunol 193:6016–6030

Schmitter T, Pils S, Sakk V, Frank R, Fischer KD, Hauck CR (2007) The granulocyte receptor carcinoembryonic antigen-related cell adhesion molecule 3 (CEACAM3) directly associates with Vav to promote phagocytosis of human pathogens. J Immunol 178:3797–3805

Shanley DK, Kiely PA, Golla K, Allen S, Martin K, O'Riordan RT, Ball M, Aplin JD, Singer BB, Caplice N, Moran N, Moore T (2013) Pregnancy-specific glycoproteins bind integrin alphaIIbbeta3 and inhibit the platelet-fibrinogen interaction. PLoS ONE 8:e57491

Singer BB, Opp L, Heinrich A, Schreiber F, Binding-Liermann R, Berrocal-Almanza LC, Heyl KA, Muller MM, Weimann A, Zweigner J, Slevogt H (2014) Soluble CEACAM8 interacts with CEACAM1 inhibiting TLR2-triggered immune responses. PLoS ONE 9:e94106

Slevogt H, Zabel S, Opitz B, Hocke A, Eitel J, N'Guessan P D, Lucka L, Riesbeck K, Zimmermann W, Zweigner J, Temmesfeld-Wollbrueck B, Suttorp N, Singer BB (2008) CEACAM1 inhibits toll-like receptor 2-triggered antibacterial responses of human pulmonary epithelial cells. Nature Immunology 9, 1270–1278

Snyder SK, Wessner DH, Wessells JL, Waterhouse RM, Wahl LM, Zimmermann W, Dveksler GS (2001) Pregnancy-specific glycoproteins function as immunomodulators by inducing secretion of IL-10, IL-6 and TGF-beta1 by human monocytes. Am J Reprod Immunol 45:205–216

Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, Zhang Y, Ye K, Jun G, Hsi-Yang Fritz M, Konkel MK, Malhotra A, Stutz AM, Shi X, Paolo Casale F, Chen J, Hormozdiari F, Dayama G, Chen K, Malig M, Chaisson MJ, Walter K, Meiers S, Kashin S, Garrison E, Auton A, Lam HY, Jasmine Mu X, Alkan C, Antaki D, Bae T, Cerveira, E, Chines P, Chong Z, Clarke L, Dal E, Ding L, Emery S, Fan X, Gujral M, Kahveci F, Kidd JM, Kong Y, Lameijer EW, McCarthy S, Flicek P, Gibbs RA, Marth G, Mason CE, Menelaou A, Muzny DM, Nelson BJ, Noor A, Parrish NF, Pendleton M, Quitadamo A, Raeder B, Schadt EE, Romanovitch M, Schlattl A, Sebra R, Shabalin AA, Untergasser A, Walker JA, Wang M, Yu F, Zhang C, Zhang J, Zheng-Bradley X, Zhou W, Zichner T, Sebat J, Batzer MA, McCarroll SA, Genomes Project C, Mills RE, Gerstein MB, Bashir A, Stegle O, Devine SE, Lee C, Eichler EE, Korbel JO (2015) An integrated map of structural variation in 2,504 human genomes. Nature 526:75–81

Sulkowski GN, Warren J, Ha CT, Dveksler GS (2011) Characterization of receptors for murine pregnancy specific glycoproteins 17 and 23. Placenta 32:603–610

Tatarino YS, Masyukev VN, (1970) Immunochemical identification of a new Beta-1-Globulin in blood serum of pregnant women. B Exp Biol Med-Ussr 69, 666

Tatra G, Breitenecker G, Gruber W (1974) Serum concentration of pregnancy-specific beta-1-glycoprotein (sp-1) in normal and pathologic pregnancies. Archiv fur Gynakologie 217:383–390

Tchoupa AK, Schuhmacher T, Hauck CR (2014) Signaling by epithelial members of the CEACAM family—mucosal docking sites for pathogenic bacteria. Cell Commun Signal 12:27

Teglund S, Zhou GQ, Hammarstrom S (1995) Characterization of cDNA encoding novel pregnancy-specific glycoprotein variants. Biochem Biophys Res Commun 211:656–664

Thompson J, Zimmermann W, Nollau P, Neumaier M, Weber-Arden J, Schrewe H, Craig I, Willcocks T (1994) CGM2, a member of the carcinoembryonic antigen gene family is down-regulated in colorectal carcinomas. J Biol Chem 269:32924–32931

Towler CM, Horne CH, Jandial V, Campbell DM, MacGillivray I (1977) Plasma levels of pregnancy-specific beta 1-glycoprotein in complicated pregnancies. Br J Obstet Gynaecol 84:258–263

Trowsdale J, Parham P (2004) Mini-review: defense strategies and immunity-related genes. Eur J Immunol 34:7–17

Von Kleist S (1992) Introduction to the CEA family: structure, function and secretion. Int J Biolo Markers 7:132–136

Watanabe S, Chou JY (1988) Human pregnancy-specific beta 1-glycoprotein: a new member of the carcinoembryonic antigen gene family. Biochem Biophys Res Comm 152:762–768

Waterhouse R, Ha C, Dveksler GS (2002) Murine CD9 is the receptor for pregnancy-specific glycoprotein 17. J Exp Med 195:277–282

Sequencing CMG, Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, An P, Antonarakis SE, Attwood J, Baertsch R, Bailey J, Barlow K, Beck S, Berry E, Birren B, Bloom T, Bork P, Botcherby M, Bray N, Brent MR, Brown DG, Brown SD, Bult C, Burton J, Butler J, Campbell RD, Carninci P, Cawley S, Chiaromonte F, Chinwalla AT, Church DM, Clamp M, Clee C, Collins FS, Cook LL, Copley RR, Coulson A, Couronne O, Cuff J, Curwen V, Cutts T, Daly M, David R, Davies J, Delehaunty KD, Deri J, Dermitzakis ET, Dewey C, Dickens NJ, Diekhans M, Dodge S, Dubchak I, Dunn DM, Eddy SR, Elnitski L, Emes RD, Eswara P, Eyras E, Felsenfeld A, Fewell GA, Flicek P, Foley K, Frankel WN, Fulton LA, Fulton RS, Furey TS, Gage D, Gibbs RA, Glusman G, Gnerre S, Goldman N, Goodstadt L, Grafham D, Graves TA, Green ED, Gregory S, Guigo R, Guyer M, Hardison RC, Haussler D, Hayashizaki Y, Hillier LW, Hinrichs A, Hlavina W, Holzer T, Hsu F, Hua A, Hubbard T, Hunt A, Jackson I, Jaffe DB, Johnson LS, Jones M, Jones TA, Joy A, Kamal M, Karlsson EK, Karolchik D, Kasprzyk A, Kawai J, Keibler E, Kells C, Kent WJ, Kirby A, Kolbe DL, Korf I, Kucherlapati RS, Kulbokas EJ, Kulp D, Landers T, Leger JP, Leonard S, Letunic I, Levine R, Li J, Li M, Lloyd C, Lucas S, Ma B, Maglott DR, Mardis ER, Matthews L, Mauceli E, Mayer JH, McCarthy M, McCombie WR, McLaren S, McLay K, McPherson JD, Meldrim J, Meredith B, Mesirov JP, Miller W, Miner TL, Mongin E, Montgomery KT, Morgan M, Mott R, Mullikin JC, Muzny DM, Nash WE, Nelson JO, Nhan MN, Nicol R, Ning Z, Nusbaum C, O'Connor MJ, Okazaki Y, Oliver K, Overton-Larty E, Pachter L, Parra G, Pepin KH, Peterson J, Pevzner P, Plumb R, Pohl CS, Poliakov A, Ponce TC, Ponting CP, Potter S, Quail M, Reymond A, Roe BA, Roskin KM, Rubin EM, Rust AG, Santos R, Sapojnikov V, Schultz B, Schultz J, Schwartz MS, Schwartz S, Scott C, Seaman S, Searle S, Sharpe T, Sheridan A, Shownkeen R, Sims S, Singer JB, Slater G, Smit A, Smith DR, Spencer B, Stabenau A, Stange-Thomann N, Sugnet C, Suyama M, Tesler G, Thompson J, Torrents D, Trevaskis E, Tromp J, Ucla C, Ureta-Vidal A, Vinson JP, Von Niederhausern AC, Wade CM, Wall M, Weber RJ, Weiss RB, Wendl MC, West AP, Wetterstrand K, Wheeler R, Whelan S, Wierzbowski J, Willey D, Williams S, Wilson RK, Winter E, Worley KC, Wyman D, Yang S, Yang SP, Zdobnov EM, Zody MC, Lander ES (2002) Initial sequencing and comparative analysis of the mouse genome. Nature 420:520–562

Wessells J, Wessner D, Parsells R, White K, Finkenzeller D, Zimmermann W, Dveksler G (2000) Pregnancy specific glycoprotein 18 induces IL-10 expression in murine macrophages. Eur J Immunol 30:1830–1840

Wu JA, Johnson BL, Chen Y, Ha CT, Dveksler GS (2008) Murine pregnancy-specific glycoprotein 23 induces the proangiogenic factors transforming-growth factor beta 1 and vascular endothelial growth factor a in cell types involved in vascular remodeling in pregnancy. Biol Reprod 79:1054–1061

Wynne F, Ball M, McLellan AS, Dockery P, Zimmermann W, Moore T (2006) Mouse pregnancy-specific glycoproteins: tissue-specific expression and evidence of association with maternal vasculature. Reproduction 131:721–732

Zebhauser R, Kammerer R, Eisenried A, McLellan A, Moore T, Zimmermann W (2005) Identification of a novel group of evolutionarily conserved members within the rapidly diverging murine Cea family. Genomics 86:566–580

Zhang G, Cowled C, Shi Z, Huang Z, Bishop-Lilly KA, Fang X, Wynne JW, Xiong Z, Baker ML, Zhao W, Tachedjian M, Zhu Y, Zhou P, Jiang X, Ng J, Yang L, Wu L, Xiao J, Feng Y, Chen Y, Sun X, Zhang Y, Marsh GA, Crameri G, Broder CC, Frey KG, Wang LF, Wang J (2013) Comparative analysis of bat genomes provides insight into the evolution of flight and immunity. Science 339:456–460

Zhao C, Wang F, Wang P, Ding H, Huang X, Shi Z (2015) Early second-trimester plasma protein profiling using multiplexed isobaric tandem mass tag (TMT) labeling predicts gestational diabetes mellitus. Acta Diabetol 52:1103–1112

Zhao L, Triche EW, Walsh KM, Bracken MB, Saftlas AF, Hoh J, Dewan AT (2012) Genome-wide association study identifies a maternal copy-number deletion in PSG11 enriched among preeclampsia patients. BMC pregnancy and childbirth 12:61

Zhou GQ, Hammarstrom S (2001) Pregnancy-specific glycoprotein (PSG) in baboon (Papio hamadryas): family size, domain structure, and prediction of a functional region in primate PSGs. Biol Reprod 64:90–99

Chapter 4
Convergent Evolution of Starch Metabolism in Cyanobacteria and Archaeplastida

Christophe Colleoni and Ugo Cenci

Abstract It is widely accepted that Archaeplastida phylum comprising Glaucophyta, Rhodophyta, and Chloroplastida originates from a unique endosymbiosis event, called primary plastid endosymbiosis, between a cyanobacterium and a eukaryotic cell. In addition to acquiring oxygenic photosynthesis, the three sister lineages gained the ability to synthesize a novel semi-crystalline storage polysaccharide: starch. In Archaeplastida, several lines of evidence reveal that the transition from glycogen synthesis to starch accumulation results in the recruitment of an isoamylase (ISA)-type debranching enzyme. The latter removes short-branched glucan chains, which prevent amylopectin crystallization. Recently, a small group of unicellular diazotrophic cyanobacteria, possibly the closest relative of the ancestral plastid, have been reported accumulating starch-like granules composed of both amylose and amylopectin fractions instead of glycogen particles. In order to understand starch metabolism in this particular group of cyanobacteria, a random mutagenesis was carried out on the unicellular starch-accumulating *Cyanobacterium* sp. CLg1. Throughout iodine crystal vapors screening, fourteen mutant strains have substituted starch granules by that of glycogen particles. Interestingly, such as in plants, all mutant strains were impaired in an isoamylase-type debranching enzyme activity. However, phylogenetic analyses point out that the critical step for starch crystallization in Archaeplastida did not evolve from the cyanobacterial isoamylase/*glgX* gene, but from another pathogenic bacteria. Based on this work, it appears that the transition from glycogen to starch has evolved independently in both cyanobacteria and Archaeplastida by following a common glucan trimming mechanism.

C. Colleoni (✉) · U. Cenci
UMR8576—Université de Lille-Sciences et Technologies,
Bât C9-R59655 Villeneuve d'Ascq, France
e-mail: Christophe.colleoni@univ-lille1.fr

P. Pontarotti (ed.), *Evolutionary Biology*, DOI 10.1007/978-3-319-41324-2_4

4.1 Introduction

Plants/green algae, red algae, and glaucophytes derive from a common ancestor, which emerged around 1.5 billion years ago, when a phagotrophic heterotroph eukaryote engulfed a cyanobacterium (Eme et al. 2014; Yoon et al. 2004). Through endosymbiotic gene transfer (Martin et al. 2002), the cyanobacterial ancestor evolved to a photosynthetic organelle (i.e., plast), named chloroplasts in plants and green algae (Chloroplastida), rhodoplast in red algae (Rhodophyta), and cyanelle in Glaucophyta lineage. Amazingly, the acquisition of photosynthetic apparatus correlates with the appearance of new dense, water-insoluble semi-crystalline storage polysaccharide named starch. The latter is localized in the chloroplasts of green algae and plants, while similar granules, named floridean starch, are observed in the cytosol of red algae (Viola et al. 2001) and glaucophytes (Plancke et al. 2008). Like hydrosoluble glycogen particles, the oldest and widespread form of storage polysaccharide in many various eukaryotic and prokaryotic cells, starch granules are composed of linear glucan chains made of glucose residues linked in α-1,4 and branched in α-1,6. Nevertheless, the particular organization of glucan chains in both polysaccharides results in opposite physicochemical properties such as solubility in water, size, and crystallinity. It should be stressed out that starch granules are also reported in red algal-derived organisms such as dinoflagellates (Deschamps et al. 2008b), apicomplexan (Coppin et al. 2004), and cryptophytes (Deschamps et al. 2006). Those organisms are grouped into CASH lineages (Cryptophyta, Alveolata (dinoflagellates, apicomplexan), Stramenopiles, and Haptophytes) derived for most of them from a secondary endosymbiosis between red algae and eukaryotic cell. However, despite the increased number of sequence data, the relationship among CASH lineages remains unclear and subjects to discussion (Cavalier-Smith 1999; Petersen et al. 2014).

Because glycogen and starch metabolism pathways share the same enzymatic steps for biosynthesis: (i) synthesis of nucleotide-sugar (ii) formation of α-1,4 linkages and (iii) formation of α-1,6 linkages and two biochemical reactions for metabolizing α-1,4 and α-1,6 linkages, an important question arises for a while in the scientific community: How the related enzyme systems generate two architecturally distinct but chemically identical types of polymers? Over the last two decades, this question has been tackled in different plants models (rice, maize, Arabidopsis) and green alga (*Chlamydomonas reinhardtii*). The general agreement proposes that a glucan trimming step is an essential prerequisite for the transition from glycogen to starch (Ball et al. 1996). Interestingly, a recent investigation extends this mandatory step to a starch-accumulating *Cyanobacterium* sp. CLg1. Thus, we could assume that Archaeplastida inherited this process from cyanobacteria. Nevertheless, phylogeny analysis reveals that starch debranching activities found in Archaeplastida are not related to cyanobacteria, and thus, a convergent evolution of starch crystallization process occurs independently in both *Cyanobacterium* sp. CLg1 and Archaeplastida.

The aim of this chapter was to give an overview of storage metabolism pathways in plants and cyanobacteria, and particularly the crystallization step. Hence, we think that a short description of two major storage polysaccharides, starch and glycogen, will be helpful with regard to their function in the living cells. We will also discuss about unexpected outcome on the study of *Cyanobacterium* sp. CLg1, which offers a new perspective in our knowledge of primary endosymbiosis.

4.2 The Physicochemical Properties of Alpha-Glucan Polysaccharides: Glycogen and Starch

Glycogen, the common storage polysaccharide in Eukaryota, Bacteria, and Archaea, is localized into the cytosol of prokaryotes and eukaryotic organisms (i.e., animal cells and fungi) as tiny hydrosoluble particles of 40–60 nm (for review (Wilson et al. 2010). Mathematical modeling, based on the structural characterization of diverse sources of glycogen, predicts a size of 42 nm in diameter and justifies this self-limitation due to homogenous distribution of branching points. This particular organization leads quickly to an exponential increase in the number of non-reducing ends at the surface of polysaccharide (Fig. 4.1a) and then will limit the access to catalytic site of biosynthetic enzymes and per se the size of glycogen particles (Melendez et al. 1998; Melendez-Hevia et al. 1993). Amazingly, further mathematical modeling infers that glycogen particles are fractal objects (i.e., repetition of a single motif: two branches per glucan chain), which allow the release of a large amount of glucose in a short time by catabolism enzymes (Melendez et al. 1999). Following this mathematical modeling, glucan chains are organized in such manner that 19,000 residues of glucose molecules are available in 20 s (Melendez et al. 1997). Hence, glycogen particles are dynamic polysaccharides, which can meet the immediate need for eukaryotic and prokaryotic cells or can be used as temporally sink of carbon. Because glycogen particles are optimized to fill up the needs of the cell, any structural alteration of glycogen can modify its properties. For instance, the Anderson's disease also called amylopectinosis (Glycogen disease type IV) is due to a mutation in the glycogen branching gene. This enzyme shapes glycogen particles by cleaving an α-1,4 linkages and transferring the oligoglucosyl residue to α-1,6 position onto neighbor glucan chain. In this rare autosomal recessive metabolic disorder, the modification of branching points pattern leads to an accumulation of insoluble glucan similar to starch, termed polyglucosan bodies, which is responsible for lethal hepatic cirrhosis in the first few years (Li et al. 2010; Moses and Parvari 2002). Hence, mathematical modeling and biology evidence highlight the tight relationship between structure and function of glycogen particles.

On the contrary, starch appears as a non-aqueous semi-crystalline polysaccharide with variable size between 0.1 and 100 μm in the chloroplasts of leaves or in the cytosol of red and glaucophytes algae. Those osmotically inert granules represent the final product of dioxide carbon fixation through Calvin cycle during the day.

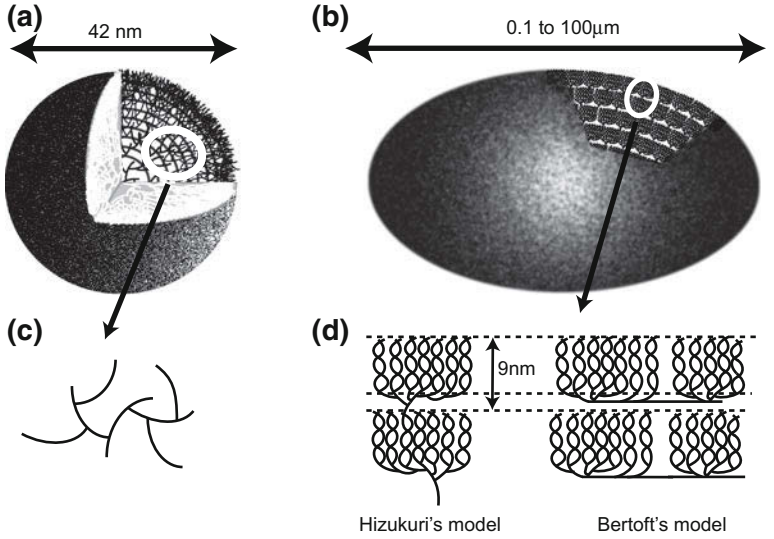


Fig. 4.1 Structural comparison between amylopectin and glycogen. Both polysaccharides are composed of glucan chains made of α-1,4 glucose residues and branched in α-1,6 position. **a** In glycogen, the uniform distribution of branches leads to an exponential increase that self-limit glycogen particle size to 42 nm. **c** Mathematical modeling suggests that glycogen particles are fractal objects made of repeating patterns of glucan chains (*black lines*) harboring two glucan chains (*intersection lines*). **b** The size of semi-crystalline starch granule is variable depending on the source. **c** Clusters of amylopectin are generated through the asymmetric distribution of branches, localized in the amorphous lamellae, while the intertwine glucan chains define crystalline lamellae. The sum of one amorphous and crystalline lamellae amounts to 9 nm independently of amylopectin clusters examined. **c** In the model proposed by Hizukuri (1986), a long glucan chain interconnects two clusters, whereas in the alternate model of Bertoft (2010), the clusters are anchored to a backbone consisting of a long glucan chain

They are usually composed of two types of polysaccharides: amylopectin and amylose. Amylopectin, the major fraction, is a branched polysaccharide containing 5 % of branching points. The physicochemical properties of starch granules result exclusively in the clustered organization of amylopectin. According to this view, first proposed by Hizukuri (1986) and modified by Bertoft (2004), branching points are placed at the root of clusters allowing the formation of double helices of glucan responsible for the crystalline properties (Fig. 4.1b). Interestingly, small-angle X-ray scattering analysis highlights a conservation of size clusters at 9 nm through starch granules of different species (Jenkins et al. 1993). As described above, amylopectin is alone responsible for the architecture of starch granules. Indeed, the absence of amylose, the minor glucan fraction composed of linear (around 1000 glucose residues) and slightly branched (less than 1 %) glucan chains, does not affect the physicochemical properties of starch granules (Inouchi et al. 1987). In many plants and green algae, granule-bound starch synthase activity (GBSS) is dedicated to the synthesis of amylose (Delrue et al. 1992; Nelson and Rines 1962).

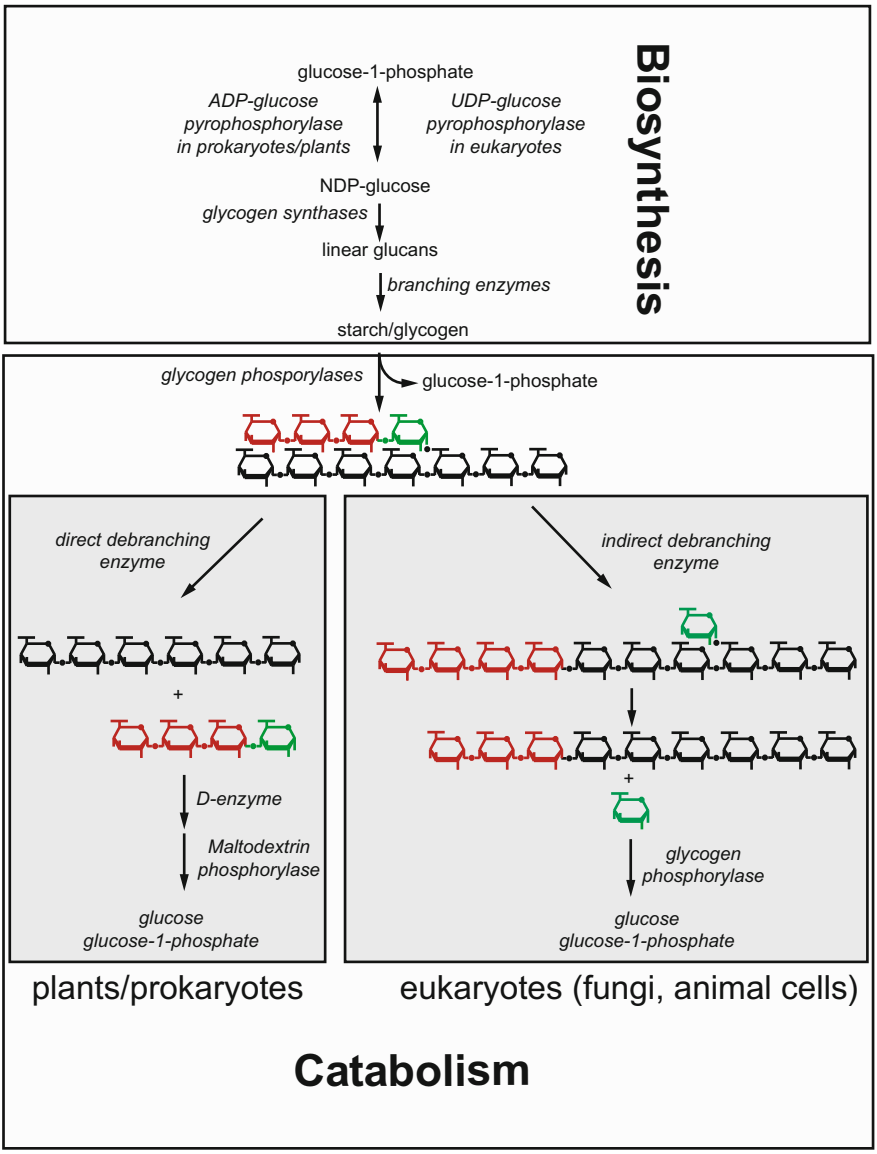
In contrast to soluble starch synthase activities, this enzyme is not found in the cellular extract and displays a unique elongation property, which consists to synthesize long glucan chains exclusively in the presence of semi-crystalline polysaccharide. If the function of GBSS activity is dispensable in higher plants, recent physiological investigations on the effects of carbon dioxide level on *Chlorella* and *Chlamydomonas* green algae have highlighted a significant role of GBSS activity in the morphology and properties of pyrenoid starch at low levels of carbon dioxide (Izumo et al. 2011; Oyama et al. 2006). These results strengthen previous experiments on in vitro amylose synthesis (van de Wal et al. 1998) and genetic interaction in *Chlamydomonas* (Maddelein et al. 1994), which highlight a role of GBSS in the synthesis of long glucan chains of amylopectin.

4.3 The Comparative Biochemistry of Storage Polysaccharide Metabolism in Heterotrophic Eukaryotes and Plants

So far, any heterotrophic eukaryote has been reported accumulating starch as storage polysaccharides. So, it seems reasonable to propose that phagotrophic eukaryote involved in the primary plastid endosymbiosis was probably also a glycogen-accumulating organism and its glycogen metabolism pathway was certainly similar to fungi and animal cells or amoeba. Because starch metabolism arises from the merge of two different storage polysaccharide pathways, it is not quite surprising to observe that starch metabolism is composed of mosaic of genes from different origin sources in Archaeplastida and particularly in plants (Ball et al. 2011). The merge is not so obvious when glycogen and starch metabolism pathways are compared (Fig. 4.2). The first aspect concerns the nucleotide-sugar used. In heterotrophic eukaryotes, UDP-glucose is exclusively used by glycogen synthase, enzyme responsible for the α-1,4 linkages, while in plants, ADP-glucose is preferentially used. The rate-limiting step is on ADP-glucose pyrophosphorylase in plants and on glycogen synthase in heterotrophic eukaryotes (fungi, animal cells). In yeast, the study of regulation of glycogen synthase results in the discovery of phosphorylation/dephosphorylation; a central regulation process in Eukaryote cells (Hardy and Roach 1993). In plants, the flux of photosynthate carbon to starch synthesis is regulated by ADP-glucose pyrophosphorylase. Phylogenetic studies confirm the cyanobacterial origin of this enzyme in plants and green algae (Ball et al. 2011), and like in cyanobacteria, this allosteric enzyme is activated or inhibited by the 3-phosphoglyceraldehyde and orthophosphate, respectively (Iglesias et al. 2006).

In both cases, glycogen and starch catabolism pathways involve two important activities: phosphorylase and debranching enzyme (Ball et al. 2011; Wilson et al. 2010). Phosphorylase activities in plants as well as in heterotrophic eukaryotes release a glucose-1-phosphate residue by the non-reducing end of glucan chains.

The reaction stops 4 to 3 residues of glucose before a branching point. Then, short-branched glucans are cleaved by the debranching enzyme activity. The genome of heterotrophic eukaryotes contains only one gene encoding for one debranching enzyme isoform. This enzyme is a large protein composed of two independent domains (Nakayama et al. 2001). As described in Fig. 4.2 (right side), the N-terminus domain possesses an α-1,4 glucanotransferase activity which



◀ **Fig. 4.2** Storage polysaccharide pathways depicted in plants/prokaryotes and Eukaryotes. NDP-glucose pyrophosphorylases catalyze the first enzymatic step responsible for synthesizing nucleotide-sugars, ADP-glucose, or UDP-glucose in plants/prokaryotes and eukaryotes, respectively. In plants and prokaryotes, various metabolic effectors influence the ADP-glucose pyrophosphorylase activity, which impacts on the amount of glycogen/starch. In contrast to plants/prokaryotes, eukaryotic glycogen synthase activity regulates the level of glycogen through post-translational modifications. Glycogen/starch synthases transfer the glucose residue of NDP-glucose onto the non-reducing end of growing glucan chain, then branching enzymes introduce α-1,6 linkages. During the catabolism pathway, phosphorylase activities release glucose-1-phosphate and stop four glucose residues away from the branching point. In plants/prokaryotes (*on the left*), direct debranching enzyme cleaves off branched maltotetraosyl residues. The latter is completely metabolizes in glucose and G1P by disproportionating and maltodextrin phosphorylase activities. In eukaryotes (*on the right*), indirect debranching enzyme is composed of two domains: One hydrolyzes the α-1,4 linkage next to the branch and transfers maltotriosyl group (*red residues*) onto the non-reducing end of the neighboring chain, and the second domain cleaves off the glucose residue linked in α-1,6 position (*green residue*). The resulting linear glucan is then further hydrolyzed by glycogen phosphorylase

transfers a maltotriosyl group to the non-reducing of neighbor glucan chain, and then the second C-terminus domain cleaves the remaining glucose residue branched in α-1,6 position thanks to an amylo-1,6 glucosidase activity. The result of this debranching step will be the release of one molecule of glucose. This contrasts with plants, where debranching enzymes will cleave the α-1,6 linkages directly and release a short maltooligosaccharide in the stroma. This maltooligosacharide in plants as well as in Bacteria will be disproportionated by an α-1,4 glucanotransferase in order to make it accessible to phosphorylase activities. Through those enzymological differences, the comparison between the debranching enzyme (DBE) activities in plants and in heterotrophic eukaryotes highlights a clear biochemical cutoff in the manner, of which the α-1,6 linkages will be metabolized. Based on biochemical reactions, DBE of heterotrophic eukaryotes are named indirect debranching enzyme (iDBE), while they are called direct debranching enzymes in both plants and prokaryotes.

4.4 Bacteria Debranching Enzyme Activity Is Required for the Transition from Glycogen to Starch in Plants

Until the end of the twentieth centuries, the emergence of starch in the green algae and plants remains a question mark, among the community. The characterization of mutants impaired in the starch biosynthesis in different plant species; then in the green alga *C. reinhardtii* surprisingly revealed a defect in the hydrolytic enzyme capable of cleaving the α-1,6 linkage. Based on the ability to cleave the amylopectin (ISA1; ISA2; and ISA3) or pullulan (Pullulanase), two families of debranching enzymes are distinguished in plants. It must be stressed out that pullulan is not synthesized in plants and green algae. However, this particular

polysaccharide, made of maltotriosyl residues linked in α-1,6, is used to discriminate debranching enzyme activities.

Described in 1901 (Correns 1901), sugary mutants of maize accumulate both a large amount of soluble polysaccharide with a structure similar to glycogen (phytoglycogen) and a high level of sucrose. It is only in 1995 that the Myers group correlated the sugary phenotype with a mutation in the isoamylase or debranching enzyme gene (James et al. 1995). Similar phenotypes were described in the green alga *C. reinhardtii* (Mouille et al. 1996), in rice (Kubo et al. 1999) and *Arabidopsis* (Wattebled et al. 2005). These mutants, which are impaired in the debranching enzyme activity, substitute starch biosynthesis by a large amount of phytoglycogen. This confirms the key role of debranching enzyme on the transition of glycogen to starch. In addition, further heterologous expression experiments in rice and Arabidopsis reveal the universality of this enzyme among plants (Facon et al. 2013; Kubo et al. 2005; Streb and Zeeman 2014). In order to explain this unexpected function for a hydrolytic enzyme in the amylopectin biosynthesis, a trimming model was proposed in 1996 (Ball et al. 1996). This model suggests that an isoamylase-type debranching enzyme trims a highly branched amylopectin, called pre-amylopectin, and allows the formation of double helices of glucan. This implies that DBE is capable of cleaving loosely branched glucans, interfering with the formation of double helices of glucan or cluster organization of amylopectin. In the absence of DBE, a highly branched polysaccharide accumulates in the mutant strains. Moreover, genetic forward experiments in Arabidopsis, maize, and rice indicate the role of ISA3 and pullulanase activities mainly in the catabolism pathway, while ISA1 and ISA2 are involved primarily in the biosynthesis pathway. Despite the identification of orthologous genes in red algae and glaucophytes, such mutants have not been yet characterized due to the complexity to perform mutagenesis and screening on them.

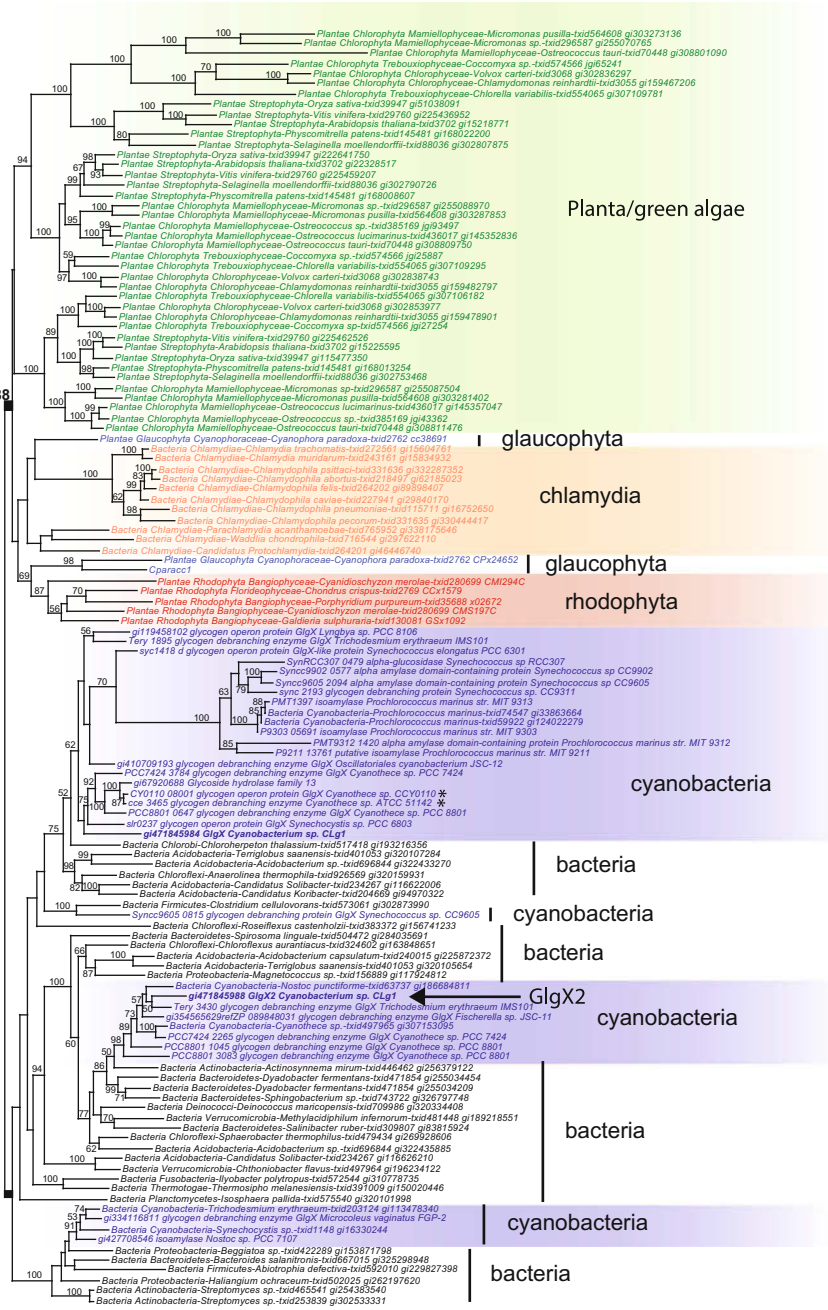
4.5 Evidence of Starch-Accumulating Cyanobacteria Strains

Like Bacteria, glycogen is predominately observed in most of the cyanobacteria as tiny hydrosoluble particles between thylakoid membranes. However in 1994, a time course experiments consisting of the observation of thin sections of nitrogen-fixing cyanobacteria *Cyanothece* ATCC51142 through the day/night cycle revealed an intriguing accumulation of large bodies between thylakoid membranes during the day and a disappearance at night (Schneegurt et al. 1994). Those large bodies were partially characterized and described as abnormal glycogen particles and not as starch-like material (Schneegurt et al. 1997). Later on, a survey of storage polysaccharides in different species of cyanobacteria reported the presence of solid granules in others cyanobacterial species (Nakamura et al. 2005). Detailed characterization indicates now that such carbohydrate granules, including those of

Cyanothece ATCC51142, are composed of a high-molecular weight polysaccharide similar to amylopectin (Suzuki et al. 2013). More recently, *Cyanobacterium* sp. CLg1, a new strain isolated in the tropical North Atlantic Ocean phylogenetically related to *Crocosphaera watsonii,* accumulates starch granules made of both amylopectin and amylose fractions (Deschamps et al. 2008a; Falcon et al. 2002). Interestingly, as in plants, the presence of amylose is correlated with the identification of a polypeptide showing a high similarity in amino acid sequence to GBSS (Granule-Bound Starch Synthase) of plants (Deschamps et al. 2008a). As discussed above, this activity is required for the synthesis of amylose in starch granules of various photosynthetic organisms. So far, the GBSS gene is described only in two cyanobacterial species, *Cyanobacterium* sp. CLg1 and *Crocosphaera watsonii,* belonging to Chroococcales order. Nevertheless, because ADP-glucose-dependent glucan synthases are in essence prokaryotic (never found in heterotrophic eukaryotes) and archaeplastidial GBSS is related to bacterial sequences (Deschamps et al. 2008a), it seems reasonable to assume that this type of sequence has evolved first among starch-accumulating cyanobacteria.

4.6 Convergent Evolution of Glucan Trimming Mechanism in the Starch Biosynthesis Pathway of *Cyanobacterium* sp. CLg1

Over this past decade, the number of completely sequenced genomes and the possibility of applying forward genetic approaches on some cyanobacteria species have provided a great opportunity to better understand the physiology of these unique prokaryotes. In addition, comparative genomic approaches highlight a striking number of isoforms in the Cyanobacterium sp. CLg1 genome in comparison with the famous enterobacteria model *Escherichia coli* (Colleoni and Suzuki 2012). For instance, two genes, *glgX1* and *glgX2,* encoding debranching enzymes (GH13 family) are found in the genome of *Cyanobacterium* sp. CLg1 versus one in the genome of *E.coli*. As described in Fig. 4.2, the characterization of *glgX* mutants in both *E. coli* and *Synechococcus* confirms its function in the catabolism pathway (Dauvillée et al. 2005; Suzuki et al. 2007). The two mutants are not able to metabolize the α-1,6 linkages and thus accumulate glycogen particles enriched with short-branched maltooligosaccharide. Until recently, our understanding of glycogen metabolism pathway in the bacterial world was relied mostly on the characterization of *Escherichia coli* and glycogen-accumulating cyanobacteria models: *Synechocystis* PCC6803 and *Synechoccocus* PCC7942. Unfortunately, deciphering the storage polysaccharide pathway in the starch-accumulating cyanobacteria was an important challenge since all those strains are refractory to any protocol of transformation. Nevertheless, in order to tackle this question, a UV mutagenesis campaign was carried out on *Cyanobacterium* sp. CLg1. Based on iodine staining of cell patches, more than a hundred mutants were identified and subsequently


Planta/green algae
glaucophyta
chlamydia
glaucophyta
rhodophyta
cyanobacteria
bacteria
cyanobacteria
bacteria
GlgX2
cyanobacteria
bacteria
cyanobacteria
bacteria
0.5 substitutions/site

◀ **Fig. 4.3** Maximum likelihood phylogeny of debranching enzymes of Archaeplastida and prokaryotes. DBEs identified in the genome of plants/green algae (*green text*), Glaucophyta (*bleu text*) and Rhodophyta (*red text*) do not share the same phylogenetic origin as the cyanobacterial GlgX (*purple text*) and more particularly GlgX2 of Cyanobacterium sp. Clg1 (*black arrow*). In addition, there is no correlation between the presence of GlgX2 activity and starch among starch-accumulating cyanobacteria (*black stars*). Rather, with a bootstrap value of 88 %, DBEs in Archaeplastida seems closely related to Chlamydiae (*orange text*)

categorized according to the ratio of soluble to insoluble polysaccharide. Among them, a dozen mutants (class A mutant) harboring an increase in the water-soluble glycogen-like fraction and a disappearance of starch granules were further characterized. For all mutants, we identified starch-metabolizing enzymes by using either a specific assay or a gel activity analysis. Among all starch enzymes, gel activity analysis reveals the absence of blue band activity in all mutants CLg1 strain. Both protein purification and molecular characterization point out a defect in direct debranching enzyme (DBE) activity (Cenci et al. 2013), catabolized by a glycosyl hydrolase family 13 according to the CAZy classification (Cantarel et al. 2009). Interestingly, the biochemical characterization reveals that the specificity of this debranching enzyme is similar to the activity involved in the crystallization process of starch in plants. Detailed phylogeny analysis built with debranching enzyme sequences of Archaeplastida and Bacteria suggests that the Archaeplastida sequences do not have a cyanobacterial origin, but derived from Chlamydiales species, known as obligatory intracellular bacteria (Fig. 4.3). Hence, Cenci et al. (2013) point out an example of convergent evolution, based on the trimming step of glucose chains required for starch crystallization in two different systems.

This striking observation opens the door to new research area, which is beyond the scope of this chapter. Nevertheless, it should be stressed out that a model named "ménage à trois" has been proposed to take into account the involvement of Chlamydiales species in the establishment of primary endosymbiosis (Ball et al. 2015b).

4.7 Conclusion

Debranching activities play a critical role in the formation of amylopectin clusters either in starch-accumulating cyanobacteria or in plants. At first glance, we speculated that the common ancestor of Archaeplastida might have inherited the gene encoding for debranching enzyme from the engulfed cyanobacterium. Phylogenetic tree analysis of debranching enzymes reveals that this assumption was wrong. Over around 30 starch-metabolizing enzymes, only ADP-glucose pyrophosphorylase and GBSS activity have been inherited from a cyanobacterium (Ball et al. 2011). Unexpectedly, debranching enzymes of Archaeplastida appear clearly related to debranching enzyme of obligatory intracellular bacteria belonging to Chlamydiales.

Since Cyanobacteria is one of the oldest phyla, it is reasonable to propose that starch granules have emerged first in cyanobacteria using the full suite of genes involved in the glycogen biosynthesis and then reinvented in the Archaeplastida ancestor using a patchwork of genes derived from the Cyanobacteria, host cell, and chlamydia species through iso-convergent evolution. This raises the following question: Why unicellular diazotrophic cyanobacteria synthesize starch granules instead of glycogen? One reasonable explanation would be the permanent rise of oxygen in the atmosphere (Great Oxygenation Event: GOE) due to oxygenic photosynthesis activity of cyanobacteria around 2.4 billion years ago (Kopp et al. 2005). The transition of reductive to oxidative environment was a trigger event for diversification of Cyanobacteria lineages and the appearance of new traits (e.g., morphology, cell size) (Blank and Sanchez-Baracaldo 2010; Latysheva et al. 2012; Sanchez-Baracaldo et al. 2014). A remarkable adaptation was achieved in nitrogen-fixing cyanobacteria. Indeed, the reduction of dinitrogen to ammonium is catalyzed by an extremely oxygen sensitive molybdenum-dependent ATP-hydrolyzing protein complex called nitrogenase (Burris 1991). Recently, phylogenetic reconstructions indicate that the last cyanobacteria common ancestor, which predates GOE, was probably freshwater nitrogen-fixing unicellular cyanobacterium (Larsson et al. 2011; Sanchez-Baracaldo et al. 2014). As oxygen has risen, diazotrophic cyanobacteria have evolved to protect the nitrogenase activity (Bergman et al. 1997). In diazotrophic filamentous cyanobacteria, nitrogenase activity is localized in specialized cells named heterocysts harboring a thick cell wall and an absence of photosystem II, which is responsible for photosynthetic oxygen evolution. The large amount of energy required to fuel the nitrogenase activity (16 ATP/N2) is supplied by neighbor cells performing normal photosynthetic activity. Physical separation of two exclusive biological processes, i.e., photosynthesis and nitrogen fixation, is possible through the invention of dedicated cells. Consistent with this view, it was believed that nitrogen fixation could not occur in unicellular cyanobacteria. However, in 1970, the discovery of two unicellular nitrogen-fixing cyanobacteria strains, *Gloeothece* sp. and *Cyanothece* sp., challenged this paradigm (Singh 1973; Wyatt and Silvey 1969). In contrast to filamentous cyanobacteria that perform nitrogen fixation during the day, unicellular diazotrophic cyanobacteria carry out this activity exclusively at night. In order to solve this problem, unicellular cyanobacteria develops a temporal separation of those two incompatible processes, which takes place for some of them exclusively in micro-aerobic and for a small group of cyanobacteria also aerobic condition. In the latter condition, unicellular nitrogen fixation cyanobacteria face another obstacle: energy costs. Indeed, nitrogen fixation is engaged when the level of oxygen is low enough to allow the nitrogenase activity. To reach this anoxic environment, unicellular aerobic nitrogen-fixing cyanobacteria exhibit high rate of dark respiration, which represents an additional energy cost for cyanobacteria (Compaore and Stal 2010). Thus, unicellular diazotrophic cyanobacteria might evolve to synthesize a more efficient storage polysaccharide allowing an efficient rate of nitrogen fixation in aerobic conditions. It should be stress out that phylogenetic tree based on debranching sequence highlights that there is no obvious

correlation between starch synthesis and any particular type of debranching enzyme, like GlgX2 activity among starch-accumulating cyanobacteria. This suggests that the transition from glycogen to starch occurs in these cyanobacteria by recruiting another debranching enzyme. Recently, the measurement of nitrogen fixation in six *Cyanothece* species under various growth and incubation conditions strengthens this assumption. In this report, starch-accumulating *Cyanothece* spp. exhibit a higher rate of nitrogen fixation than glycogen-accumulating *Cyanothece* species in aerobic growth condition (Bandyopadhyay et al. 2013). This indirect proof reinforces the idea that GOE, 2.4 billion years ago, was probably a driving force for the transition from glycogen to semi-crystalline starch granules.

Another striking observation is the lack of debranching enzyme activity in some starch-accumulating organisms. Genome analyses of starch-accumulating *Cyanobacterium* MIBC10216 (Nakamura et al. 2005) and red algal-derived organisms do not reveal any gene encoding for debranching enzyme (CASH lineages). In such condition, the crystallization of amylopectin may follow a different mechanism. A preliminary answer to this question has been addressed through the characterization of DBE knockout mutant of Arabidopsis (Streb et al. 2008). In the absence of chloroplastic endoamylase, phytoglycogen-accumulating knockout mutant accumulates a significant amount of starch granules. So, it might be possible in this unique condition that the combination of isoforms of branching enzymes alone offers a specific branching pattern that promotes a cluster organization of amylopectin. This could explain why usually three-to-four branching enzymes isoforms are usually found in those genomes (Ball et al. 2015a; Colleoni and Suzuki 2012). Further functional approaches might help us to understand the key processes involved with the crystallization of starch and could be another example of convergent evolution.

Interestingly, the study of starch metabolism pathway in *Cyanobacterium* sp. CLg1 offers a new insight on the ancestral cyanobacterium involved in the primary endosymbiosis. In spite of various attempts, the identity of the closest present cyanobacteria related to the cyanobacteria ancestor remains debated in the community (Criscuolo and Gribaldo 2011; Falcon et al. 2010; Li et al. 2014; Ochoa de Alda et al. 2014). The common agreement on this subject is that the cyanobacterial ancestor was nitrogen-fixing cyanobacterium. Through this study, we propose that ancestral cyanobacterium was a unicellular cyanobacterium, belonging to Chroococcales clade, capable fixing nitrogen and per se starch accumulating as storage polysaccharide. Indeed, a body of argument strengths this hypothesis: (i) Chroococcales order groups some nitrogen-fixing cyanobacteria, (ii) cyanobacterial-derived *gbss* gene in Archaeplastida's genome is exclusively found in both Chroococcales species: *Cyanobacterium* sp. CLg1 and *Crocosphaera watsonii*, (iii) the semi-crystalline environment requires for maintaining GBSS activity, and (iv) the report of cyanobacterial symbiosis between a diatom (*Rhopalodia gibba*) and a cyanobiont, which is close relative to starch-accumulating cyanobacterium (Prechtl et al. 2004), support the Chroococcales order as the putative cyanobacterium ancestor involved in the primary endosymbiosis.

References

Ball S, Colleoni C, Cenci U et al (2011) The evolution of glycogen and starch metabolism in eukaryotes gives molecular clues to understand the establishment of plastid endosymbiosis. J Exp Bot 62:1775–1801

Ball S, Guan HP, James M et al (1996) From glycogen to amylopectin: a model for the biogenesis of the plant starch granule. Cell 86:349–352

Ball SG, Colleoni C, Arias MC (2015a) The transition from glycogen to starch metabolism in cyanobacteria and eukaryotes. In: Nakamura Y (ed) Starch: metabolism and structure. Springer, Berlin, pp 93–158

Ball SG, Colleoni C Kadouche D et al (2015b) Toward an understanding of the function of Chlamydiales in plastid endosymbiosis. Biochim Biophys Acta 1847:495–504

Bandyopadhyay A, Elvitigala T, Liberton M et al (2013) Variations in the rhythms of respiration and nitrogen fixation in members of the unicellular diazotrophic cyanobacterial genus *Cyanothece*. Plant Physiol 161:1334–1346

Bergman B, Gallon JR, Rai AN et al (1997) N2 fixation by non-heterocystous cyanobacteria. FEMS Microbiol Rev 19:139–185

Bertoft E, Laohaphatanalert K, Piyachomkwan K, Sriroth K (2010) The fine structure of cassava starch amylopectin. Part 2: building block structure of clusters. Int J Biol Macromol 47:325–335

Bertoft E (2004) On the nature of categories of chains in amylopectin and their connection to the super helix model. Carbohydr Polym 57:211–224

Blank CE, Sanchez-Baracaldo P (2010) Timing of morphological and ecological innovations in the cyanobacteria—a key to understanding the rise in atmospheric oxygen. Geobiology 8:1–23

Burris RH (1991) Nitrogenase. J Biol Chem 266:9339–9342

Cantarel BL, Coutinho PM, Rancurel C et al (2009) The carbohydrate-active enzymes database (CAZy): an expert resource for Glycogenomics. Nucleic Acids Res 37:D233–D238

Cavalier-Smith T (1999) Principles of protein and lipid targeting in secondary symbiogenesis: euglenoid, dinoflagellate, and sporozoan plastid origins and the eukaryote family tree. J Eukaryot Microbiol 46:347–366

Cenci U, Chabi M, Ducatez M, Tirtiaux C, Nirmal-Raj J, Utsumi Y, Kobayashi D, Sasaki S, Suzuki E, Nakamura Y, Putaux JL, Roussel X, Durand-Terrasson A, Bhattacharya D, Vercoutter-Edouart AS, Maes E, Arias MC, Palcic M, Sim L, Ball SG, Colleoni C (2013) Convergent evolution of polysaccharide debranching defines a common mechanism for starch accumulation in cyanobacteria and plants. Plant Cell 25:3961–3975

Colleoni C, Suzuki E (2012) Storage polysaccharide metabolism in cyanobacteria. Essential reviews in experimental biology: starch: origins, structure and metabolism, vol 5 (Chap. 5)

Compaore J, Stal LJ (2010) Oxygen and the light-dark cycle of nitrogenase activity in two unicellular cyanobacteria. Environ Microbiol 12:54–62

Coppin A, Varre JS, Lienard L et al (2004) Evolution of plant-like crystalline storage polysaccharide in the protozoan parasite Toxoplasma gondii argues for a red alga ancestry. J Mol Evol 60:257–267

Correns C (1901) Bastarde zwischen maisrassen mit besonder berucksichtung der Xenien. Bibl Bot 53:1–161

Criscuolo A, Gribaldo S (2011) Large-scale phylogenomic analyses indicate a deep origin of primary plastids within cyanobacteria. Mol Biol Evol 28:3019–3032

Dauvillée D, Kinderf IS, Li Z et al (2005) Role of the *Escherichia coli glgX* gene in glycogen metabolism. J Bacteriol 187:1465–1473

Delrue B, Fontaine T, Routier F et al (1992) Waxy *Chlamydomonas reinhardtii*: monocellular algal mutants defective in amylose biosynthesis and granule-bound starch synthase activity accumulate a structurally modified amylopectin. J Bacteriol 174:3612–3620

Deschamps P, Colleoni C, Nakamura Y et al (2008a) Metabolic symbiosis and the birth of the plant kingdom. Mol Biol Evol 25:536–548

Deschamps P, Guillebeault D, Devassine J et al (2008b) The heterotrophic dinoflagellate *Crypthecodinium cohnii* defines a model genetic system to investigate cytoplasmic starch synthesis. Eukaryot Cell 7:872–880

Deschamps P, Haferkamp I, Dauvillée D et al (2006) Nature of the periplastidial pathway of starch synthesis in the cryptophyte *Guillardia theta*. Eukaryot Cell 5:954–963

Eme L, Sharpe SC, Brown MW et al (2014) On the age of eukaryotes: evaluating evidence from fossils and molecular clocks. Cold Spring Harb Perspect Biol 6

Facon M, Lin Q, Azzaz AM et al (2013) Distinct functional properties of isoamylase-type starch debranching enzymes in monocot and dicot leaves. Plant Physiol 163:1363–1375

Falcón LI, Cipriano F, Chistoserdov AY et al (2002) Diversity of diazotrophic unicellular cyanobacteria in the tropical North Atlantic Ocean. Appl Environ Microbiol 68:5760–5764

Falcón LI, Magallón S, Castillo A (2010) Dating the cyanobacterial ancestor of the chloroplast. ISME J 4:777–783

Hardy TA, Roach PJ (1993) Control of yeast glycogen synthase-2 by COOH-terminal phosphorylation. J Biol Chem 268:23799–23805

Hizukuri S (1986) Polymodal distribution of the chain lengths of amylopectin and its significance. Carbohydr Res 147:342–347

Iglesias AA, Ballicora MA, Sesma JI et al (2006) Domain swapping between a cyanobacterial and a plant subunit ADP-glucose pyrophosphorylase. Plant Cell Physiol 47:523–530

Inouchi N, Glover DV, Fuwa H (1987) Chain length distribution of amylopectins of several single mutants and the normal counterpart, and *sugary-1* phytoglycogen in maize (*Zea mays*). Starch 39:259–266

Izumo A, Fujiwara S, Sakurai T et al (2011) Effects of granule-bound starch synthase I-defective mutation on the morphology and structure of pyrenoidal starch in Chlamydomonas. Plant Sci 180:238–245

James MG, Robertson DS, Myers AM (1995) Characterization of the maize gene *sugary1*, a determinant of starch composition in kernels. Plant Cell 7:417–429

Jenkins PJ, Cameron RE, Donald AM (1993) A universal feature in the structure of starch granules from different botanical sources. Starch 45:415–420

Kopp RE, Kirschvink JL, Hilburn IA et al (2005) The Paleoproterozoic snowball Earth: a climate disaster triggered by the evolution of oxygenic photosynthesis. Proc Natl Acad Sci USA 102:11131–11136

Kubo A, Fujita N, Harada K et al (1999) The starch-debranching enzymes isoamylase and pullulanase are both involved in amylopectin biosynthesis in rice endosperm. Plant Physiol 121:399–410

Kubo A, Rahman S, Utsumi Y et al (2005) Complementation of *sugary-1* phenotype in rice endosperm with the wheat *isoamylase1* gene supports a direct role for isoamylase1 in amylopectin biosynthesis. Plant Physiol 137:43–56

Larsson J, Nylander JA, Bergman B (2011) Genome fluctuations in cyanobacteria reflect evolutionary, developmental and adaptive traits. BMC Evol Biol 11:187

Latysheva N, Junker VL, Palmer WJ et al (2012) The evolution of nitrogen fixation in cyanobacteria. Bioinformatics 28:603–606

Li B, Lopes JS, Foster PG et al (2014) Compositional biases among synonymous substitutions cause conflict between gene and protein trees for plastid origins. Mol Biol Evol 31:1697–1709

Li SC, Chen CM, Goldstein JL et al (2010) Glycogen storage disease type IV: novel mutations and molecular characterization of a heterogeneous disorder. J Inherit Metab Dis 33(Suppl 3):S83–S90

Maddelein ML, Libessart N, Bellanger F et al (1994) Toward an understanding of the biogenesis of the starch granule. Determination of granule-bound and soluble starch synthase functions in amylopectin synthesis. J Biol Chem 269:25150–25157

Martin W, Rujan T, Richly E et al (2002) Evolutionary analysis of Arabidopsis, cyanobacterial, and chloroplast genomes reveals plastid phylogeny and thousands of cyanobacterial genes in the nucleus. Proc Natl Acad Sci USA 99:12246–12251

Melendez R, Melendez-Hevia E, Canela EI (1999) The fractal structure of glycogen: a clever solution to optimize cell metabolism. Biophys J 77:1327–1332

Melendez R, Melendez-Hevia E, Cascante M (1997) How did glycogen structure evolve to satisfy the requirement for rapid mobilization of glucose? A problem of physical constraints in structure building. J Mol Evol 45:446–455

Melendez R, Melendez-Hevia E, Mas F et al (1998) Physical constraints in the synthesis of glycogen that influence its structural homogeneity: a two-dimensional approach. Biophys J 75:106–114

Melendez-Hevia E, Waddell TG, Shelton ED (1993) Optimization of molecular design in the evolution of metabolism: the glycogen molecule. Biochem J 295(Pt 2):477–483

Moses SW, Parvari R (2002) The variable presentations of glycogen storage disease type IV: a review of clinical, enzymatic and molecular studies. Curr Mol Med 2:177–188

Mouille G, Maddelein ML, Libessart N et al (1996) Preamylopectin processing: a mandatory step for starch biosynthesis in plants. Plant Cell 8:1353–1366

Nakamura Y, Takahashi J, Sakurai A et al (2005) Some Cyanobacteria synthesize semi-amylopectin type alpha-polyglucans instead of glycogen. Plant Cell Physiol 46:539–545

Nakayama A, Yamamoto K, Tabata S (2001) Identification of the catalytic residues of bifunctional glycogen debranching enzyme. J Biol Chem 276:28824–28828

Nelson OE, Rines HW (1962) The enzymatic deficiency in the waxy mutant of maize. Biochem Biophys Res Commun 9:297–300

Ochoa de Alda JA, Esteban R, Diago ML et al (2014) The plastid ancestor originated among one of the major cyanobacterial lineages. Nat Commun 5:4937

Oyama Y, Izumo A, Fujiwara S et al (2006) Granule-bound starch synthase cDNA in *Chlorella kessleri* 11h: cloning and regulation of expression by CO(2) concentration. Planta 224:646–654

Petersen J, Ludewig AK, Michael V et al (2014) *Chromera velia*, endosymbioses and the rhodoplex hypothesis—plastid evolution in cryptophytes, alveolates, stramenopiles, and haptophytes (CASH lineages). Genome Biol Evol 6:666–684

Plancke C, Colleoni C, Deschamps P et al (2008) Pathway of cytosolic starch synthesis in the model glaucophyte *Cyanophora paradoxa*. Eukaryot Cell 7:247–257

Prechtl J, Kneip C, Lockhart P et al (2004) Intracellular spheroid bodies of *Rhopalodia gibba* have nitrogen-fixing apparatus of cyanobacterial origin. Mol Biol Evol 21:1477–1481

Sanchez-Baracaldo P, Ridgwell A, Raven JA (2014) A neoproterozoic transition in the marine nitrogen cycle. Curr Biol 24:652–657

Schneegurt MA, Sherman DM, Nayar S et al (1994) Oscillating behavior of carbohydrate granule formation and dinitrogen fixation in the cyanobacterium *Cyanothece* sp. strain ATCC 51142. J Bacteriol 176:1586–1597

Schneegurt MA, Sherman DM, Sherman LA (1997) Composition of the carbohydrate granules of the cyanobacterium, *Cyanothece* sp. strain ATCC 51142. Arch Microbiol 167:89–98

Singh PK (1973) Nitrogen fixation by the unicellular blue-green alga *Aphanothece*. Arch Mikrobiol 92:59–62

Streb S, Delatte T, Umhang M et al (2008) Starch granule biosynthesis in Arabidopsis is abolished by removal of all debranching enzymes but restored by the subsequent removal of an endoamylase. Plant Cell 20:3448–3466

Streb S, Zeeman SC (2014) Replacement of the endogenous starch debranching enzymes ISA1 and ISA2 of Arabidopsis with the rice orthologs reveals a degree of functional conservation during starch synthesis. PLoS One 9:e92174

Suzuki E, Onoda M, Colleoni C et al (2013) Physicochemical variation of cyanobacterial starch, the insoluble alpha-glucans in cyanobacteria. Plant Cell Physiol 54:465–473

Suzuki E, Umeda K, Nihei S et al (2007) Role of the GlgX protein in glycogen metabolism of the cyanobacterium, *Synechococcus elongatus* PCC 7942. Biochim Biophys Acta 1770:763–773

van de Wal M, D'Hulst C, Vincken JP et al (1998) Amylose is synthesized in vitro by extension of and cleavage from amylopectin. J Biol Chem 273:22232–22240

Viola R, Nyvall P, Pedersen M (2001) The unique features of starch metabolism in red algae. Proc Biol Sci 268:1417–1422

Wattebled F, Dong Y, Dumez S et al (2005) Mutants of Arabidopsis lacking a chloroplastic isoamylase accumulate phytoglycogen and an abnormal form of amylopectin. Plant Physiol 138:184–195

Wilson WA, Roach PJ, Montero M et al (2010) Regulation of glycogen metabolism in yeast and bacteria. FEMS Microbiol Rev 34:952–985

Wyatt JT, Silvey JK (1969) Nitrogen fixation by *gloeocapsa*. Science 165:908–909

Yoon HS, Hackett JD, Ciniglia C et al (2004) A molecular timeline for the origin of photosynthetic eukaryotes. Mol Biol Evol 21:809–818

Chapter 5
The Evolution of Brains and Cognitive Abilities

Christopher Mitchell

Abstract Humans place great significance in intellectual abilities, and because of this, biologists have been interested in the cognitive abilities of animals since Aristotle. The difficulties in defining intelligence in a way that can be applied to taxonomically distant animals have resulted in most studies using relative brain size as a directly measurable metric of intelligence. This approach has received criticism but persists in the literature as it has proved to be informative despite imperfections. Large brain size and, by inference, arguably complex cognition have evolved independently in several lineages including primates, corvids, cetaceans, cephalopods and hymenopteran insects. In these varying taxa, the evolutionary history of intelligence is still hotly debated. Proponents of the social intelligence hypothesis suggest that the cognitive challenges of group living have driven the evolution of large brains whereas other researchers have suggested ecological drivers such as seasonal challenges, dietary differences and energetic constraints. Here I review the study of the evolution of intelligence with particular focus on the widely cited social intelligence hypothesis and cognitive buffering hypothesis. I begin by summarising the study of animal intelligence and the methods employed. I will then go on to briefly review the current state of knowledge concerning cognitive evolution in some of the most heavily studied taxa. Finally, I will summarise and evaluate the social intelligence hypothesis and the cognitive buffering hypothesis. I propose that the weight of evidence suggests that social intelligence is limited to relatively few taxa, such as primates and cetaceans, and that across other taxa, a variety of other factors have driven the evolution of advanced intelligence in animals.

C. Mitchell (✉)
Institute of Integrative Biology, Biosciences Building, University of Liverpool, Crown Street, Liverpool L69 7ZB, UK
e-mail: cmitch@liv.ac.uk

P. Pontarotti (ed.), *Evolutionary Biology*, DOI 10.1007/978-3-319-41324-2_5

5.1 Defining and Studying Intelligence in Animals

Intelligence is defined in humans as an individual's performance on a variety of cognitive tasks, often compiled into a metric such as the intelligence quotient (IQ). Researchers investigating animal intelligence are addressing a slightly different phenomenon than IQ by trying to assess species typical intelligence. Most definitions of intelligence employed in the study of animals are similar in that they emphasise the importance of behavioural flexibility and problem-solving abilities (Reviewed in Roth 2015). Dicke and Roth (2016) define intelligence as "the ability of an organism to solve problems occurring in its natural and social environment, culminating in the appearance of novel solutions that are not part of the animal's normal repertoire". This definition has the advantage that intellectual abilities of animals will be observable in the wild and, if tests can be designed that can be applied to a wide variety of species, testable under laboratory conditions.

In order to study intelligence in animals comparatively, a metric that can be applied across taxa is necessary. The most obvious and easily measurable potential index of intelligence is brain size. Absolute brain size is broadly considered not to be a reliable indicator of cognitive ability as most of the variation in brain mass can be attributed to variation in body mass. However, a recent study in which over 500 individuals from 36 species ranging from pigeons to chimpanzees were tested on problem-solving tasks that required self-control showed that absolute brain size is the best predictor of performance on the tasks (MacLean et al. 2014) raising the possibility that absolute brain size may be informative. The majority of researchers prefer to consider measures of brain size independent of body size and thus attempt to control for the confounding effect of body size on brain size statistically. The vast majority of studies on brain size and intelligence use either encephalization quotient (EQ) or relative brain size. EQ uses the allometric relationship between body size and brain size to derive a relative measure of brain size (Jerison 1973). EQ relies on accurately determining the nature of the allometric relationship between brain and body size. Studies of mammals have variously placed the exponent of this relationship (the slope of the line on a log–log plot of body mass and brain mass) at 0.67 and 0.75 (Boddy et al. 2012). Variation in the scaling relationship between body and brain size between groups can be problematic. In a study of cetacean brain size, Manger (2006) used the scaling parameter of all mammals to calculate EQs for cetaceans. In fact, cetacean brain mass scales quite differently to terrestrial mammals, having a scaling parameter of 0.376 (Manger 2006), possibly due to their aquatic lifestyles (Marino 1998). Improper use of EQs in this manner can drastically alter the results of subsequent statistical analysis as was pointed out in this case by Marino et al. (2008) who advocate the use of relative brain size instead.

Using relative brain size is a slightly different methodology. Relative brain size takes account of the correlation between brain and body size in statistical models and allows researchers to identify any effects down to variation in brain size, independent of body size. There is a vast literature using relative brain size as indicative of intelligence (Reviewed in Healy and Rowe 2007), but this is not

without criticisms. Studying brain size as a metric of cognition assumes that any increase in size results in an increase in function or complexity and increases in relative size of specific brain regions may not be detectable by changes in whole brain size (Healy and Rowe 2007).

Some authors prefer to emphasise behavioural flexibility and thus use reported incidences of cognitively advanced behaviours as a quantitative measure of intelligence (Ducatez et al. 2015; Lefebvre et al. 2004; Reader and Laland 2002). For primates and birds, the behavioural ecology literature contains many examples of innovative behaviour because both groups are well studied and researchers are inclined to report their observations. As a result, primate intelligence has been studied using the reported incidences of innovation, tool use, social learning, extractive foraging and tactical deception (Reader et al. 2011; Reader and Laland 2002; Byrne and Corp 2004), and bird intelligence has been studied using foraging and technical innovations (Ducatez et al. 2015; Overington et al. 2009). Similar observations of apparently cognitively complex behaviours from other groups are sparse. In cetaceans for example, there are numerous behavioural observations of a small number of well-studied species but not enough to allow large-scale comparative analysis. This approach focuses on behaviour and, in doing so, addresses some of the concerns of Healy and Rowe (2007). Most significantly, this approach attempts to study intelligence directly by using complex behaviour as a direct consequence of cognitive complexity and therefore circumvents issues concerning the use of indirect measures such as brain size. Nevertheless, using reports of complex behaviours may be biased towards well-studied species, potentially overestimating the complexity of species such as chimpanzees and bottlenose dolphins which are heavily studied for their behavioural repertoires. In such cases, it may be appropriate to correct for research effort in much the same way as brain size is corrected for body size (Reader and Laland 2002).

5.2 Intelligent Animals

Non-human primates have attracted the most attention in the study of animal intelligence due to their close relation to humans and their relatively large brains compared to other mammals. Primates are known for their capacity for cultural transmission (Yamamoto et al. 2013), tool use (Ottoni and Izar 2008; Boesch and Boesch 1990), behavioural innovation (Reader and Laland 2002), tactical deception (Byrne and Corp 2004) and potentially theory of mind (the ability to attribute mental states to others) (Tomasello et al. 2003). The complexity of the behavioural repertoire of primates correlates well with relative brain size, and this has been used to argue in favour of a general intelligence (Reader et al. 2011).

Amongst birds, two lineages are held up as possessing cognitive abilities comparable to those observed in primates. Corvids (crows, jays and magpies) have been described by researchers as “feathered apes” because of striking similarities in the cognitive abilities of some corvids and those of the great apes (Emery 2006).

Parrots are often held up alongside corvids as extremely intelligent birds and are famous for prodigious vocal learning capacities (Emery 2006). Corvids have been shown to make and use tools (Weir et al. 2002), possess precursors of theory of mind (Bugnyar 2011) and exhibit analogical reasoning (Smirnova et al. 2015) and casual reasoning (Taylor et al. 2010). Scrub-jays use cognitively complex strategies to protect their caches from thieves (Emery et al. 2004), and some have argued that crows and parrots can solve problems by insight (Pepperberg 2004; Bird and Emery 2009). The similar levels of cognitive complexity of corvids, parrots and primates have been attributed to convergent evolution (Emery and Clayton 2004).

Cetaceans (whales, dolphins and porpoises) possess some of the largest brains of any animal. Despite the common perception that cetaceans are some of the most intelligent non-human animals, researchers are in fact deeply divided on the question of cetacean intelligence. Bottlenose dolphins (*Tursiops truncatus*) and killer whales (*Orcinus orca*) in particular are thought to be highly intelligent based on numerous observations in captivity and in the wild of apparently complex behaviours such as male–male alliances similar to those of chimpanzees (Connor 2007), tool use (Krutzen et al. 2014; Smolker et al. 1997) and communication of identity information analogous to names (Janik et al. 2006). Conversely, some authors assert that despite their relatively large brains, claims of advanced cognition in cetacea are overstated (Manger 2013). Roth (2015) describes experiments on dolphin cognition as generating "mixed and often disappointing results" whereas Manger (2013) asserts that the evidence in favour of complex cetacean cognition becomes considerably less impressive when placed within a broader comparative framework as many of the reputedly complex behaviours occur throughout vertebrates and often invertebrates as well. Such critics generate strong responses from cetacean researchers (Marino et al. 2007, 2008) who argue that the weight of evidence in favour of complex cognitive abilities in cetaceans is convincing. Controversy over the intellectual status of cetaceans remains unresolved.

Carnivora, a mammalian order containing approximately 300 terrestrial species and around 30 aquatic species (known as pinnipeds), have been quite well studied, and as a result, data on brain size are available for most species. As a whole, the behavioural repertoires of wild carnivorans have not been assessed in a similar manner to primates or birds, making inferences concerning their behavioural complexity difficult. However, a recent study of captive carnivores has shown that relative brain size reliably predicts the ability of a carnivoran to solve a novel problem (Benson-Amram et al. 2016). This would seem to indicate that relative brain size is a reliable metric of cognitive ability.

The cephalopods (octopuses, cuttlefish, squid and nautiluses) have remarkably large and complex brains, especially octopuses which possess the largest, most complex brain of any invertebrate (Roth 2015). Octopuses have demonstrated abilities such as spatial learning and memory (Boal et al. 2000), observational learning (Fiorito and Scotto 1992) and potentially tool use (Finn et al. 2009).

Rarely considered in studies of intelligence, insects are often thought of as cognitively very simple. However, the neural architecture of insects is relatively well known, and in particular, the structure known as the mushroom body (*corpora*

pedunculata) has been of particular interest. The mushroom body has been shown to play a major role in many of the behavioural markers of complex cognition including associative learning in *Drosophila* (McGuire et al. 2001), spatial memory in cockroaches (Mizunami et al. 1998) and selective attention in *Drosophila* (Xi et al. 2008). Special attention has been given to the apparently structurally complex mushroom bodies of the hymenoptera which are argued to be the seat of complex cognitive behaviours such as spatial orientation and social behaviour (Roth 2015).

5.3 The Evolution of Brains and Cognition

Hypotheses purporting to explain variation in brain size, and hence cognitive complexity, are numerous and have been the subject of much debate in the literature. These competing hypotheses mostly fall into two schools of thought. Ecological theories propose a direct link between cognitive ability and environmental challenges, supposing that given features of the environment favour increased cognitive abilities to deal with the cognitive challenges posed. By contrast, social theories propose that living in large or complex social groups presents cognitive challenges and so selection pressures that favour grouping will also result in increases in cognitive ability. The social and ecological schools of thought differ on one key point. The crux of this debate is which selection pressure is directly responsible for evolutionary increases in cognitive ability.

5.3.1 The Social Intelligence Hypothesis

The most broadly cited hypothesis for the evolution of large brains is the social intelligence hypothesis, which posits that large brains and intelligence are an adaptation to social living (Humphrey 1976). Living as a group provides numerous benefits to animals. Living in a social group is also thought to present a variety of cognitive challenges such as navigating a hierarchy, keeping track of interactions and cooperative behaviour patterns (Dunbar 1998). The social intelligence hypothesis states that these problems are solved or managed by having a larger brain and more advanced cognitive abilities to allow animals to cope with increased competition for food and matings.

The social intelligence hypothesis was initially developed as an explanation for large brains in primates, many of which live in large complex groups. Social group size is limited in primates by the relative size of the neocortex which is thought to be the part of the brain most involved with complex cognition (Dunbar 1992; Kudo and Dunbar 2001). These primate findings have formed the basis of much of the research into the social intelligence hypothesis. However, similar relationships have not been found to be widespread in animals.

Amongst fishes, cichlids have been shown to exhibit cooperative hunting behaviour and other potentially complex social behaviours (Bshary et al. 2002; Roth 2015). In some cases, these behaviours show some resemblance to primates, particularly observations of social learning and traditions in a variety of species (Bshary et al. 2002). Social group size has been linked to brain size in the cichlid species of Lake Tanganyika (Pollen et al. 2007) supporting the social intelligence hypothesis. Some studies of brain size in cartilaginous fish have found larger, more complex brain structures in social species such as carcharhinid and sphyrnid sharks (Yopak et al. 2007).

Studies of carnivores have shown that the social African lion (*Panthera leo*) exhibits sex-specific differences in neocortex size that are absent in the solitary cougar (*Puma concolor*) suggesting a link between sociality and brain organisation (Arsznov and Sakai 2012). In support of the social brain hypothesis, Perez-Barberia et al. (2007) argued for a tight coevolutionary relationship between sociality and relative brain size in carnivorans. However, detailed reconstructions of living and extinct lineages of carnivorans reveal no relationship between sociality and brain size throughout the history of Carnivora (Finarelli and Flynn 2009). In fact, Finarelli and Flynn (2009) determined that the relationship between sociality and brain size in carnivores is limited to the Canidae family and removal of this lineage from their analysis invalidates the claims of Perez-Barbeira et al. (2007). Furthermore, studies of hyaenas have shown that the predictions of the social brain hypothesis (that social species will have larger relative brain sizes) do not apply (Holekamp et al. 2015). Even the highly social spotted hyaena (*Crocuta crocuta*), which has a social system comparable to cercopithecine primates such as baboons, relies much more on relatively simple forms of social learning such as facilitation than primates, seemingly indicating a relatively simple degree of social intelligence (Holekamp et al. 2007). Across Carnivora, relatively large brains are also found in mustelids (weasels, martens, badgers and otters), some of the smaller cats and bears, which all share predominantly solitary lifestyles (Finarelli and Flynn 2009).

It is important to note that the social intelligence hypothesis is not directly concerned with group size but rather with the complexity of social living. The complexity of primate groups is a subject of significant study with considerable variety throughout the clade (Kasper and Voelkl 2009), and the underlying assumption that larger social group sizes imply a more complex social lifestyle has been questioned (Bergman and Beehner 2015). Some efforts have been made to study social network dynamics, and the results are less clear. Lehmann and Dunbar (2009) used network cohesion and found that in primates with larger neocortex ratios, females tend to live in fragmented, smaller grooming clans. To counter this, Lehmann and Dunbar (2009) suggest that the complexities of living within highly fragmented social systems, also known as fission–fusion groups, and maintaining social cohesion drove the evolution of advanced intelligence of primates.

The assumption that larger social groups are more complex seems to hold within primates but not necessarily when we consider other lineages such as ungulates and birds which occasionally gather in herds or flocks numbering in the thousands or more. These very large groups are not always complex as individuals will typically

not engage in complex interactions. In birds, group size shows no relationship to forebrain size (Beauchamp and Fernandez-Juricic 2004). Emery and colleagues (2007) note that birds with long-term pair bonds tend to have the largest brains, possibly supporting the social brain hypothesis. This combined with the observation that bird flocks are much less stable than primate groups suggests that it is the cognitive challenges of forming and maintaining long-term bonds that drove the evolution of cognitive abilities in birds. However, other factors influencing intelligence have also been identified including ecological generalism (Overington et al. 2011) and a resident lifestyle as opposed to migratory (Sol et al. 2005).

A wealth of comparative analyses of brain evolution in bats has revealed a complex picture. One such analysis demonstrated that monogamous bat species have the largest brains with polygynous species also having relatively large brains, but promiscuous species have relatively small brains (Pitnick et al. 2006). Similar results in birds have been used to argue in favour of the social intelligence hypothesis (Shultz and Dunbar 2010), but Pitnick et al. (2006) suggest that mate fidelity in bats and high relative brain size are both the product of an evolutionary trade-off between brain size and testes size and thus the correlation between brain size and mating system is a by-product of sexual selection. This scenario presents a possible confounding factor in the study of sociality and brain size and casts doubt on some interpretations of links between mating system and brain size. There is some debate concerning these results as Shultz and Dunbar (2007) assert that in fact, the mating system–brain size relationship is a direct causal relationship and that the correlation between brain size and testes size is a by-product of both traits being closely related to mating size. Similarly, equivocal evidence comes from ungulates where gregarious species have been shown to have larger brains (Perez-Barberia and Gordon 2005), but other studies have shown that brain size can be predicted by both mating system and ecological factors such as habitat use (Shultz and Dunbar 2006).

Observations of behavioural complexity in hymenopterans (bees, ants and wasps) (Roth 2015) could be assumed to support the social brain hypothesis. However, Farris and Schulmeister (2011) tested the social brain hypothesis in hymenopterans and found that a parasitoid life history rather than a social life history is associated with large, complex mushroom bodies. This leads to their interpretation that the cognitive demands of locating a host drove the enlargement of the mushroom body in this lineage, possibly serving as a pre-adaptation for the subsequent evolution of social living (Farris and Schulmeister 2011).

An extension of the social intelligence hypothesis is the cultural intelligence hypothesis (van Schaik and Burkart 2011). Under this hypothesis, the selective advantages of social learning, such as the rapid spread of novel solutions between individuals, drive increases in behavioural flexibility and general cognitive ability. Thus, non-social cognitive skills such as tool use could be a consequence of general increases in intellectual abilities brought about by social living, particularly the social learning aspect of group living which is the underlying mechanism of culture. Evidence in favour of this hypothesis comes from Reader and colleagues (2011) who used observations of primates reported in the literature and identified a general

intelligence factor with social, ecological and technical intelligence very closely correlated.

The observation that primate intelligence is not modular but instead contains a mix of social and technical skills has also been advanced as evidence against the social brain hypothesis. Based on strong correlations between cerebellum size and extractive foraging and tool use but not group size, Barton (2012) proposes the embodied cognition theory of primate brain evolution. This hypothesis proposes that ecological pressures drove the evolution of complex technical skills in parallel to the established idea that social factors drove evolution of social cognition in the neocortex. However, under this evolutionary scenario, we might expect a degree of modularity in primate cognition with the pre-frontal cortex as the seat of social cognition and the cerebellum as the principal component of technical cognition. This seems to be contradicted by Reader's (2011) general intelligence factor which is closely correlated to the neocortex ratio and includes both extractive foraging and tool use. In fact, the cerebellum has been linked to nest complexity in birds (Hall et al. 2013), and within mammals, the largest cerebellums relative to body size belong to elephants and odontocetes (Maseko et al. 2012) which are thought to be amongst the most intelligent mammals behind some primates. Furthermore, in humans and other great apes, the cerebellum contains four times the neurons of the neocortex and has expanded considerably more rapidly than the neocortex in the evolution of the ape lineage (Barton and Venditti 2014). These observations challenge an exclusively social interpretation of primate cognition.

5.3.2 Ecological Drivers of the Evolution of Intelligence

Ecological explanations for large brains and complex cognition are varied and have attributed observed patterns of variation in brain size to many factors. The expensive tissue hypothesis states that brain tissue is metabolically very expensive to maintain and evolutionary changes in diet must occur to allow the expansion of the brain (Aiello and Wheeler 1995). Studies of diet in primates and small mammals have found that folivores have smaller brains than generalists (Harvey et al. 1980), but further work showed that dietary quality, an index calculated from the relative components of each species' diet, does not account for variation in relative brain size in platyrrhine primates (Allen and Kay 2012). Furthermore, brain size in phytophagous bats is larger than in animalivorous bats (Eisenberg and Wilson 1978) which goes against the expectation that high-energy diets are needed to support large brains (Harvey et al. 1980). The larger brains of phytophagous bats have been argued to be the result of the complexity of the foraging habitat in which for bats foraging in dense vegetation provides much greater sensory challenges (Safi and Dechmann 2005). These findings would suggest a strong influence of ecology on the evolution of brain size in bats. Thus, despite the intuitive appeal of the expensive tissue hypothesis, diet alone is insufficient to explain the variation in brain size.

Melin and colleagues (2014) provide evidence in favour of the hypothesis that seasonality in foraging demands, specifically the seasonal dependence on extractive foraging which requires accessing food embedded within a substrate which may require tool use or planning of behaviour, has been selected for increased "sensorimotor" intelligence in primates. Strong seasonal fluctuations in food abundance require the ability to respond flexibly and thus are argued to explain the observed instances of tool use and innovative problem-solving in primates.

In cetacea, it has been suggested that the relatively large brains of odontocetes (toothed whales) are closely related to their sensory ecology (Jerison 1986). All extant odontocete cetaceans echolocate, as did all known fossil odontocetes but mysticete whales (baleen whales), and other aquatic mammals such as pinnipeds (seals and walruses), do not. The processing demands of echolocation could be argued to explain the high degree of encephalization observed in odontocetes compared to mysticetes. However, it has been pointed out that other echolocating animals such as bats are not highly encephalized (Marino 2007) and therefore echolocation by itself does not explain the large brains of some cetaceans. Another hypothesis proposes that living in cold water has driven the evolution of large brains in cetaceans and large brains in these species have no relation to cognitive abilities at all (Manger 2006). The so-called thermogenesis hypothesis states that the thermal challenges of living in cold waters can in part be solved by expanding the proportion of thermogenic glial cells in the brain and thus generating more heat from the brain without necessarily increasing cognitive function. This hypothesis has been heavily criticised for the dismissal of cetacean species as not exhibiting advanced cognition despite behavioural observations to the contrary (Marino et al. 2008) although these observations are themselves the subject of considerable debate (Manger 2013).

A more general hypothesis is the cognitive buffering hypothesis, which proposes that having a large brain gives an organism the ability to respond flexibly in the face of novel, unpredictable challenges posed by the environment (Sol 2009). At first glance, the cognitive buffer hypothesis may appear to be a very broad hypothesis applicable to almost any animal with a large brain in almost any situation. In fact, the hypothesis makes a clear, testable prediction. The key prediction of this hypothesis is that advanced cognitive abilities have significant survival value and this has been shown in birds where the behavioural flexibility of a species predicts the success of invasion (Sol and Lefebvre 2000). Further support for this idea comes from the fact that large-brained animals have greater longevity (Gonzalez-Lagos et al. 2010). When presented with a model predator, female guppies (*Poecilia reticulata*) with large brains evaluate the risk and habituate faster than those with small brains (Bijl et al. 2015), suggesting that a general cognitive ability is an important factor in anti-predator behavioural responses. Evidence in support of the cognitive buffer hypothesis can even be found in primates, famed for their social intelligence. A study of catarrhine primates showed that species with large relative brain sizes experienced less seasonality in their dietary intake than species in similarly seasonal habitats with smaller brains, suggesting that cognitive

buffering allowed these primates to adjust to changing food availability and maintain their energetic intake (van Woerden et al. 2012).

In marsupial mammals, analysis of brain sizes across the group revealed that species living in the relatively aseasonal environment of New Guinea have larger brains (Weisbecker et al. 2015). The reduced nutritional pressure on these animals is thought to be a factor allowing the development of large brains. The correlation between brain size and litter size in marsupials (Weisbecker and Goswami 2010) is also taken into account in hypotheses of marsupial brain evolution and leads researchers to conclude that marsupials living in environments with reliable food sources invest more in lactation which allows the brains of the young to grow larger. Crucially, no evidence has been found for any behavioural driving force for the evolution of large brains in marsupials (Weisbecker et al. 2015). This stands in stark contrast to the social brain hypothesis and other hypotheses that propose that the challenges of certain lifestyles (social living, unpredictable environments, etc.) drive the evolution of advanced cognitive abilities to deal with such challenges. Instead, in conditions that allowed large brains to grow, animals that grew large brains, and by inference developed greater intelligence/behavioural flexibility, had a fitness advantage. Weisbecker et al. (2015) interpret this as support for the cognitive buffer hypothesis as the general framework for brain evolution in mammals, contradicting claims that the social intelligence hypothesis should be considered as a general hypothesis for mammals (and birds) (Dunbar 2009).

5.4 Conclusion

Studying animal cognition is a difficult task. The consensus of many studies is that despite well-known problems and pitfalls, variation in relative brain size does predict performance on cognitive tasks. Therefore, considerable gains in knowledge can be attained from studying brain size. However, it should be noted that taking relative brain size as a measure of cognitive ability is clearly a flawed approach as whole brain size will also correlate with sensory processing abilities and other non-cognitive tasks. This leads to the conclusion that component parts of the brain may be a more promising avenue of research. This approach has been used very successfully in some groups but can make comparisons between distantly related groups difficult, especially when analogous brain structures may be difficult to identify.

The evidence suggesting that brain size and sociality are causally related, once considered strong and taxonomically widespread, appears to have been weakened by recent research. Although sociality has been strongly linked to cognition in primates and perhaps cetaceans, across mammals, there is only weak support for the social intelligence hypothesis. Uncertainty around the interpretations of different lines of evidence could be resolved by using a reliable measure of social complexity that can be generalised across species such as the number of relationships in a social network. Nevertheless, studies in carnivores, marsupials, bats and social insects

have failed to support the social intelligence hypothesis, suggesting that the relationship between group living and cognition may be limited to certain clades.

The cognitive buffering hypothesis provides a good paradigm in which to consider the evolution of intelligence. Under this paradigm, if there is a survival advantage to having a large brain and complex cognitive abilities, then we will observe greater survival, longevity, invasion success and adaptation to shifting climates in large-brained species. Some studies of birds and mammals have found such patterns as reviewed here and elsewhere. Although the cognitive buffering hypothesis gives an excellent description of how we expect large brains to evolve in circumstances where there are survival advantages, it does not propose a specific factor that may drive the evolution of large brains. This feature of the cognitive buffer hypothesis leaves open the possibility that different factors may have favoured the evolution of advanced cognitive abilities in different lineages. Therefore, the cognitive buffering hypothesis does not stand directly opposed to the social brain hypothesis. Rather, it provides a framework that can be applied across animals to understand the multiple convergent evolutions of large brains and intelligence.

Acknowledgements My funding was provided by a Natural Environment Research Council doctoral training grant. I am grateful to Mike Speed and John Lycett for their valuable comments on an earlier version of this manuscript.

References

Aiello LC, Wheeler P (1995) The expensive-tissue hypothesis—the brain and the digestive-system in human and primate evolution. Curr Anthropol 36(2):199–221. doi:10.1086/204350

Allen KL, Kay RF (2012) Dietary quality and encephalization in platyrrhine primates. Proc R Soc B-Biol Sci 279(1729):715–721. doi:10.1098/rspb.2011.1311

Arsznov BM, Sakai ST (2012) Pride diaries: sex, brain size and sociality in the African Lion (*Panthera leo*) and Cougar (*Puma concolor*). Brain Behav Evol 79(4):275–289. doi:10.1159/000338670

Barton RA (2012) Embodied cognitive evolution and the cerebellum. Philos Trans R Soc B-Biol Sci 367(1599):2097–2107. doi:10.1098/rstb.2012.0112

Barton RA, Venditti C (2014) Rapid evolution of the cerebellum in humans and other great apes. Curr Biol 24(20):2440–2444. doi:10.1016/j.cub.2014.08.056

Beauchamp G, Fernandez-Juricic E (2004) Is there a relationship between forebrain size and group size in birds? Evol Ecol Res 6(6):833–842

Benson-Amram S, Dantzer B, Stricker G, Swanson EM, Holekamp KE (2016) Brain size predicts problem-solving ability in mammalian carnivores. Proc Natl Acad Sci. doi:10.1073/pnas.1505913113

Bergman TJ, Beehner JC (2015) Measuring social complexity. Anim Behav 103:203–209. doi:10.1016/j.anbehav.2015.02.018

Bijl WD, Thyselius M, Kotrschal A, Kolm N (2015) Brain size affects the behavioural response to predators in female guppies (*Poecilia reticulata*). Proc R Soc B-Biol Sci 282(1812):116–124. doi:10.1098/rspb.2015.1132

Bird CD, Emery NJ (2009) Insightful problem solving and creative tool modification by captive nontool-using rooks. Proc Natl Acad Sci USA 106(25):10370–10375

Boal JG, Dunham AW, Williams KT, Hanlon RT (2000) Experimental evidence for spatial learning in octopuses (*Octopus bimaculoides*). J Comp Psychol 114(3):246–252. doi:10.1037//0735-7036.114.3.246

Boddy AM, McGowen MR, Sherwood CC, Grossman LI, Goodman M, Wildman DE (2012) Comparative analysis of encephalization in mammals reveals relaxed constraints on anthropoid primate and cetacean brain scaling. J Evol Biol 25(5):981–994. doi:10.1111/j.1420-9101.2012.02491.x

Boesch C, Boesch H (1990) Tool use and tool making in wild chimpanzees. Folia Primatol 54 (1–2):86–99. doi:10.1159/000156428

Bshary R, Wickler W, Fricke H (2002) Fish cognition: a primate's eye view. Anim Cogn 5(1): 1–13. doi:10.1007/s10071-001-0116-5

Bugnyar T (2011) Knower-guesser differentiation in ravens: others' viewpoints matter. Proc R Soc B-Biol Sci 278(1705):634–640. doi:10.1098/rspb.2010.1514

Byrne RW, Corp N (2004) Neocortex size predicts deception rate in primates. Proc R Soc B-Biol Sci 271(1549):1693–1699. doi:10.1098/rspb.2004.2780

Connor RC (2007) Dolphin social intelligence: complex alliance relationships in bottlenose dolphins and a consideration of selective environments for extreme brain size evolution in mammals. Philos Trans R SocB-Biol Sci 362(1480):587–602. doi:10.1098/rstb.2006.1997

Dicke U, Roth G (2016) Neuronal factors determining high intelligence. Philos Trans R Soc Lond B: Biol Sci 371(1685). doi:10.1098/rstb.2015.0180

Ducatez S, Clavel J, Lefebvre L (2015) Ecological generalism and behavioural innovation in birds: technical intelligence or the simple incorporation of new foods? J Anim Ecol 84(1):79–89. doi:10.1111/1365-2656.12255

Dunbar RIM (1992) Neocortex size as a constraint on group size in primates. J Hum Evol 22(6):469–493. doi:10.1016/0047-2484(92)90081-j

Dunbar RIM (1998) The social brain hypothesis. Evol Anthropol 6(5):178–190. doi:10.1002/(sici)1520-6505(1998)6:5<178::aid-evan5>3.0.co;2-8

Dunbar RIM (2009) The social brain hypothesis and its implications for social evolution. Ann Hum Biol 36(5):562–572. doi:10.1080/03014460902960289

Eisenberg JF, Wilson DE (1978) Relative brain size and feeding strategies in the Chiroptera. Evolution 32(4):740–751. doi:10.2307/2407489

Emery NJ (2006) Cognitive ornithology: the evolution of avian intelligence. Philos Trans R Soc B-Biol Sci 361(1465):23–43. doi:10.1098/rstb.2005.1736

Emery NJ, Clayton NS (2004) The mentality of crows: convergent evolution of intelligence in corvids and apes. Science 306(5703):1903–1907. doi:10.1126/science.1098410

Emery NJ, Dally JM, Clayton NS (2004) Western scrub-jays (*Aphelocoma californica*) use cognitive strategies to protect their caches from thieving conspecifics. Anim Cogn 7(1):37–43. doi:10.1007/s10071-003-0178-7

Emery NJ, Seed AM, von Bayern AMP, Clayton NS (2007) Cognitive adaptations of social bonding in birds. Philos Trans R Soc B-Biol Sci 362(1480):489–505. doi:10.1098/rstb.2006.1991

Farris SM, Schulmeister S (2011) Parasitoidism, not sociality, is associated with the evolution of elaborate mushroom bodies in the brains of hymenopteran insects. Proc R Soc B-Biol Sci 278 (1707):940–951. doi:10.1098/rspb.2010.2161

Finarelli JA, Flynn JJ (2009) Brain-size evolution and sociality in Carnivora. Proc Natl Acad Sci USA 106(23):9345–9349. doi:10.1073/pnas.0901780106

Finn JK, Tregenza T, Norman MD (2009) Defensive tool use in a coconut-carrying octopus. Curr Biol 19(23):R1069–R1070

Fiorito G, Scotto P (1992) Observational learning in Octopus vulgaris. Science 256(5056): 545–547. doi:10.1126/science.256.5056.545

Gonzalez-Lagos C, Sol D, Reader SM (2010) Large-brained mammals live longer. J Evol Biol 23(5):1064–1074. doi:10.1111/j.1420-9101.2010.01976.x

Hall ZJ, Street SE, Healy SD (2013) The evolution of cerebellum structure correlates with nest complexity. Biol Lett 9(6). doi:10.1098/rsbl.2013.0687

Harvey PH, Cluttonbrock TH, Mace GM (1980) Brain size and ecology in small mammals and primates. Proc Natl Acad Sci USA-Biol Sci 77(7):4387–4389. doi:10.1073/pnas.77.7.4387

Healy SD, Rowe C (2007) A critique of comparative studies of brain size. Proc R Soc B-Biol Sci 274(1609):453–464. doi:10.1098/rspb.2006.3748

Holekamp KE, Dantzer B, Stricker G, Shaw Yoshida KC, Benson-Amram S (2015) Brains, brawn and sociality: a hyaena's tale. Anim Behav 103:237–248. doi:10.1016/j.anbehav.2015.01.023

Holekamp KE, Sakai ST, Lundrigan BL (2007) Social intelligence in the spotted hyena (*Crocuta crocuta*). Philos Trans R Soc B-Biol Sci 362(1480):523–538. doi:10.1098/rstb.2006.1993

Humphrey NK (1976) The social function of intellect. Bateson, P. P. G. And R. A. Hinde

Janik VM, Sayigh LS, Wells RS (2006) Signature whistle shape conveys identity information to bottlenose dolphins. Proc Natl Acad Sci USA 103(21):8293–8297. doi:10.1073/pnas.0509918103

Jerison HJ (1973) Evolution of the brain and intelligence. Academic Press, New York, p 1973

Jerison HJ (1986) The perceptual world of dolphins. In: Schusterman RJ, Thomas JA, Wood FG (eds) Dolphin cognition and behaviour: a comparative approach. Lawrence Erlbaum, Mahwah, New Jersey, pp 141–166

Kasper C, Voelkl B (2009) A social network analysis of primate groups. Primates 50(4):343–356. doi:10.1007/s10329-009-0153-2

Krutzen M, Kreicker S, MacLeod CD, Learmonth J, Kopps AM, Walsham P, Allen SJ (2014) Cultural transmission of tool use by Indo-Pacific bottlenose dolphins (*Tursiops* sp.) provides access to a novel foraging niche. Proc R Soc B-Biol Sci 281(1784). doi:10.1098/rspb.2014.0374

Kudo H, Dunbar RIM (2001) Neocortex size and social network size in primates. Anim Behav 62:711–722. doi:10.1006/anbe.2001.1808

Lefebvre L, Reader SM, Sol D (2004) Brains, innovations and evolution in birds and primates. Brain Behav Evol 63(4):233–246. doi:10.1159/000076784

Lehmann J, Dunbar RIM (2009) Network cohesion, group size and neocortex size in female-bonded old world primates. Proc R Soc B-Biol Sci 276(1677):4417–4422. doi:10.1098/rspb.2009.1409

MacLean EL, Hare B, Nunn CL, Addessi E, Amici F, Anderson RC, Aureli F, Baker JM, Bania AE, Barnard AM, Boogert NJ, Brannon EM, Bray EE, Bray J, Brent LJN, Burkart JM, Call J, Cantlon JF, Cheke LG, Clayton NS, Delgado MM, DiVincenti LJ, Fujita K, Herrmann E, Hiramatsu C, Jacobs LF, Jordan KE, Laude JR, Leimgruber KL, Messer EJE, Moura ACdA, Ostojic L, Picard A, Platt ML, Plotnik JM, Range F, Reader SM, Reddy RB, Sandel AA, Santos LR, Schumann K, Seed AM, Sewall KB, Shaw RC, Slocombe KE, Su Y, Takimoto A, Tan J, Tao R, van Schaik CP, Viranyi Z, Visalberghi E, Wade JC, Watanabe A, Widness J, Young JK, Zentall TR, Zhao Y (2014) The evolution of self-control. Proc Natl Acad Sci USA 111(20):E2140–E2148. doi:10.1073/pnas.1323533111

Manger PR (2006) An examination of cetacean brain structure with a novel hypothesis correlating thermogenesis to the evolution of a big brain. Biol Rev 81(2):293–338. doi:10.1017/s1464793106007019

Manger PR (2013) Questioning the interpretations of behavioural observations of cetaceans: is there really support for a special intellectual status for this mammalian order? Neuroscience 250:664–696. doi:10.1016/j.neuroscience.2013.07.041

Marino L (1998) A comparison of encephalization between odontocete cetaceans and anthropoid primates. Brain Behav Evol 51(4):230–238. doi:10.1159/000006540

Marino L (2007) Cetacean brains: how aquatic are they? Anat Record-Adv Integr Anat Evol Biol 290(6):694–700. doi:10.1002/ar.20530

Marino L, Butti C, Connor RC, Fordyce RE, Herman LM, Hof PR, Lefebvre L, Lusseau D, McCowan B, Nimchinsky EA, Pack AA, Reidenberg JS, Reiss D, Rendell L, Uhen MD, Van der Gucht E, Whitehead H (2008) A claim in search of evidence: reply to Manger's thermogenesis hypothesis of cetacean brain structure. Biol Rev 83(4):417–440. doi:10.1111/j.1469-185X.2008.00049.x

Marino L, Connor RC, Fordyce RE, Herman LM, Hof PR, Lefebvre L, Lusseau D, McCowan B, Nimchinsky EA, Pack AA, Rendell L, Reidenberg JS, Reiss D, Uhen MD, Van der Gucht E, Whitehead H (2007) Cetaceans have complex brains for complex cognition. PLoS Biol 5(5):966–972. doi:10.1371/journal.pbio.0050139

Maseko BC, Spocter MA, Haagensen M, Manger PR (2012) Elephants have relatively the largest cerebellum size of mammals. Anat Record-Adv Integr Anat Evol Biol 295(4):661–672. doi:10.1002/ar.22425

McGuire SE, Le PT, Davis RL (2001) The role of Drosophila mushroom body signaling in olfactory memory. Science 293(5533):1330–1333. doi:10.1126/science.1062622

Melin AD, Young HC, Mosdossy KN, Fedigan LM (2014) Seasonality, extractive foraging and the evolution of primate sensorimotor intelligence. J Hum Evol 71:77–86. doi:10.1016/j.jhevol.2014.02.009

Mizunami M, Weibrecht JM, Strausfeld NJ (1998) Mushroom bodies of the cockroach: their participation in place memory. J Comp Neurol 402(4):520–537

Ottoni EB, Izar P (2008) Capuchin monkey tool use: overview and implications. Evol Anthropol 17(4):171–178. doi:10.1002/evan.20185

Overington SE, Griffin AS, Sol D, Lefebvre L (2011) Are innovative species ecological generalists? A test in North American birds. Behav Ecol 22(6):1286–1293. doi:10.1093/beheco/arr130

Overington SE, Morand-Ferron J, Boogert NJ, Lefebvre L (2009) Technical innovations drive the relationship between innovativeness and residual brain size in birds. Anim Behav 78(4):1001–1010. doi:10.1016/j.anbehav.2009.06.033

Pepperberg IM (2004) "Insightful" string-pulling in Grey parrots (*Psittacus erithacus*) is affected by vocal competence. Anim Cogn 7(4):263–266. doi:10.1007/s10071-004-0218-y

Perez-Barberia FJ, Gordon IJ (2005) Gregariousness increases brain size in ungulates. Oecologia 145(1):41–52. doi:10.1007/s00442-005-0067-7

Perez-Barberia FJ, Shultz S, Dunbar RIM (2007) Evidence for coevolution of sociality and relative brain size in three orders of mammals. Evolution 61(12):2811–2821. doi:10.1111/j.1558-5646.2007.00229.x

Pitnick S, Jones KE, Wilkinson GS (2006) Mating system and brain size in bats. Proc R Soc B-Biol Sci 273(1587):719–724. doi:10.1098/rspb.2005.3367

Pollen AA, Dobberfuhl AP, Scace J, Igulu MM, Renn SCP, Shumway CA, Hofmann HA (2007) Environmental complexity and social organization sculpt the brain in Lake Tanganyikan cichlid fish. Brain Behav Evol 70(1):21–39. doi:10.1159/000101067

Reader SM, Hager Y, Laland KN (2011) The evolution of primate general and cultural intelligence. Philos Trans R Soc B-Biol Sci 366(1567):1017–1027. doi:10.1098/rstb.2010.0342

Reader SM, Laland KN (2002) Social intelligence, innovation, and enhanced brain size in primates. Proc Natl Acad Sci USA 99(7):4436–4441. doi:10.1073/pnas.062041299

Roth G (2015) Convergent evolution of complex brains and high intelligence. Philos Trans R Soc Lond Ser B, Biol Sci 370(1684). doi:10.1098/rstb.2015.0049

Safi K, Dechmann DKN (2005) Adaptation of brain regions to habitat complexity: a comparative analysis in bats (Chiroptera). Proc R Soc B-Biol Sci 272(1559):179–186. doi:10.1098/rspb.2004.2924

Shultz S, Dunbar RIM (2006) Both social and ecological factors predict ungulate brain size. Proc R Soc B-Biol Sci 273(1583):207–215. doi:10.1098/rspb.2005.3283

Shultz S, Dunbar RIM (2007) The evolution of the social brain: anthropoid primates contrast with other vertebrates. Proc R Soc B-Biol Sci 274(1624):2429–2436. doi:10.1098/rspb.2007.0693

Shultz S, Dunbar RIM (2010) Social bonds in birds are associated with brain size and contingent on the correlated evolution of life-history and increased parental investment. Biol J Linn Soc 100(1):111–123

Smirnova A, Zorina Z, Obozova T, Wasserman E (2015) Crows spontaneously exhibit analogical reasoning. Curr Biol 25(2):256–260. doi:10.1016/j.cub.2014.11.063

Smolker R, Richards A, Connor R, Mann J, Berggren P (1997) Sponge carrying by dolphins (Delphinidae, *Tursiops* sp.): a foraging specialization involving tool use? Ethology 103(6): 454–465

Sol D (2009) Revisiting the cognitive buffer hypothesis for the evolution of large brains. Biol Lett 5(1):130–133. doi:10.1098/rsbl.2008.0621

Sol D, Lefebvre L (2000) Behavioural flexibility predicts invasion success in birds introduced to New Zealand. Oikos 90(3):599–605. doi:10.1034/j.1600-0706.2000.900317.x

Sol D, Lefebvre L, Rodriguez-Teijeiro JD (2005) Brain size, innovative propensity and migratory behaviour in temperate Palaearctic birds. Proc R Soc B-Biol Sci 272(1571):1433–1441. doi:10.1098/rspb.2005.3099

Taylor AH, Elliffe D, Hunt GR, Gray RD (2010) Complex cognition and behavioural innovation in New Caledonian crows. Proc R Soc B-Biol Sci 277(1694):2637–2643. doi:10.1098/rspb.2010.0285

Tomasello M, Call J, Hare B (2003) Chimpanzees understand psychological states—the question is which ones and to what extent. Trends Cogn Sci 7(4):153–156. doi:10.1016/s1364-6613(03)00035-4

van Schaik CP, Burkart JM (2011) Social learning and evolution: the cultural intelligence hypothesis. Philos Trans R Soc B-Biol Sci 366(1567):1008–1016. doi:10.1098/rstb.2010.0304

van Woerden JT, Willems EP, van Schaik CP, Isler K (2012) Large brains buffer energetic effects of seasonal habitats in catarrhine primates. Evolution 66(1):191–199. doi:10.1111/j.1558-5646.2011.01434.x

Weir AAS, Chappell J, Kacelnik A (2002) Shaping of hooks in new Caledonian crows. Science 297(5583):981. doi:10.1126/science.1073433

Weisbecker V, Blomberg S, Goldizen AW, Brown M, Fisher D (2015) The evolution of relative brain size in Marsupials is energetically constrained but not driven by behavioral complexity. Brain Behav Evol 85(2):125–135. doi:10.1159/000377666

Weisbecker V, Goswami A (2010) Brain size, life history, and metabolism at the marsupial/placental dichotomy. Proc Natl Acad Sci USA 107(37):16216–16221. doi:10.1073/pnas.0906486107

Xi W, Peng Y, Guo J, Ye Y, Zhang K, Yu F, Guo A (2008) Mushroom bodies modulate salience-based selective fixation behavior in Drosophila. Eur J Neurosci 27(6):1441–1451. doi:10.1111/j.1460-9568.2008.06114.x

Yamamoto S, Humle T, Tanaka M (2013) Basis for cumulative cultural evolution in chimpanzees: social learning of a more efficient tool-use technique. Plos One 8(1). doi:10.1371/journal.pone.0055768

Yopak KE, Lisney TJ, Collin SP, Montgomery JC (2007) Variation in brain organization and cerebellar foliation in chondrichthyans: Sharks and holocephalans. Brain Behav Evol 69(4):280–300. doi:10.1159/000100037

Chapter 6
Convergence as an Evolutionary Trade-off in the Evolution of Acoustic Signals: Echolocation in Horseshoe Bats as a Case Study

David S. Jacobs, Gregory L. Mutumi, Tinyiko Maluleke and Paul W. Webala

Abstract The evolution of novel acoustic signals that are optimal for a particular function or habitat may restrict distinct lineages to the same ecological niche resulting in convergence of phenotypic traits. Such convergence could represent an evolutionary trade-off. The evolution of flutter detection may have restricted horseshoe bats (*Rhinolophus*) to similar foraging modes resulting in the convergence of phenotypic traits across different lineages. We investigated convergence in African rhinolophids using several phenotypic features. There was pronounced convergence between distantly related lineages including *R. damarensis* and *R. darlingi*, and between *R. simulator* and *R. blasii*. However, phenotypic divergence, notably in body size and resting frequency, was also evident amongst close relatives of *R. damarensis* and *R. darlingi*. These relatives diverged from both the ancestral character state and *R. damarensis* and *R. darlingi*. Such divergence suggests that an evolutionary trade-off associated with flutter detection is probably not the cause of convergence in these bats.

D.S. Jacobs (✉) · G.L. Mutumi · T. Maluleke
Department of Biological Sciences, University of Cape, Cape Town, South Africa
e-mail: david.jacobs@uct.ac.za

G.L. Mutumi
e-mail: gmutumi@gmail.com

T. Maluleke
e-mail: tinyiko.maluleke2009@gmail.com

P.W. Webala
Department of Forestry and Wildlife Management, Massai Mara University, Narok, Kenya
e-mail: paul.webala@gmail.com

P. Pontarotti (ed.), *Evolutionary Biology*, DOI 10.1007/978-3-319-41324-2_6

6.1 Introduction

Convergent evolution, or simply convergence, is the independent evolution of similar phenotypes amongst genetically distinct lineages (Losos 2011; Jacobs et al. 2013). When it involves fitness-enhancing traits, convergence is often attributed to adaptation to similar ecological niches (Colborn et al. 2001; Losos 2011) and is often used as evidence for natural selection. Some classical examples include the evolution of wings in pterosaurs, bats and birds and convergent evolution of mammals and marsupials (Futuyma 1998; Ridley 1996). However, besides natural selection, other processes such as random genetic drift or biological constraints can also lead to convergence (Stayton 2008). Thus, it has been recommended that pattern be separated from process when investigating convergence (Stayton 2015), allowing one to recognize convergence independently of the process that caused it.

Investigations of convergence have to be done within a phylogenetic framework for two reasons: firstly to ensure that lineages being investigated are in fact distinct in accordance with the definition of convergence and, secondly, to determine the character state of the putative ancestor of the lineages being compared. The ancestral character state is essential in determining the extent to which lineages have diverged from the common ancestor which allows the determination of whether convergence is the result of constraints or selection and/or drift.

Shared biological constraints can cause similarity in the phenotypes of distinct lineages even in the absence of selection or drift (Losos 2011) because of stasis, i.e. the lineages have diverged very little from the common ancestor. Such stasis can result from several factors. For example, the lineages have too little variation to respond to selection pressures, or because they have not experienced any selection pressure or selection for a particular trait prevents them from responding to other selection pressures, e.g. evolutionary trade-offs (Roff and Fairbairn 2007).

Evolutionary trade-offs occur when selection on one trait is opposed by the loss of fitness as a result of a concomitant change in another (Futuyma 1998; Resnick et al. 2000; Roff 2000; Roff and Fairbairn 2007). Such trade-offs may result in neither trait reaching its adaptive optimum. For example, functional trade-offs in the skulls of birds and bats between feeding and signal production has resulted in reduced song producing abilities in birds and reduced masticatory power in bats (Ballantine 2006; Jacobs et al. 2014). If such evolutionary trade-offs are shared amongst lineages, especially as a result of sharing a fitness-enhancing innovation, their phenotypes may converge as a result of the innovation preventing responses to other selection pressures. Sensory traits such as bird and frog song and echolocation may be particular susceptible to such trade-offs because slight deviations may render them ineffective. This could result in a high level of convergence across different lineages in these traits (e.g. Jacobs et al. 2013).

Bat echolocation is a unique sonar system used for orientation, foraging and communication (Thomas et al. 2004; Jones and Siemers 2010). Bat echolocation can be divided into two broad categories: low-duty-cycle (LDC) echolocation in which the period (time elapsed between the starting points of successive calls) is

long relative to the duration of the calls and high-duty-cycle (HDC) echolocation in which the period is short relative to the duration of the calls. Duty cycle is expressed as a percentage and is the ratio of the duration of the call to the period (Fenton 1999).

HDC echolocation is an evolutionary innovation that occurs in three Old World families of bats, the Hipposideridae, the Rhinonycteridae and the Rhinolophidae and one American species *Pteronotus parnellii* (Mormoopidae). HDC allows these bats to overcome the masking effects of clutter (echoes from non-target objects), e.g. background vegetation that interferes with echoes from the target. HDC bats have a uniquely specialized system in which they couple an acoustic fovea with Doppler shift compensation (DSC). The acoustic fovea is a region of the auditory cortex that is sensitive to a very narrow range of frequencies. The frequency of the echolocation calls whilst the bat is at rest, called the resting frequency (RF), falls within the narrow range of frequency of the acoustic fovea. However, the returning echo for a bat in flight returns to the bat's ears at a higher frequency as a result of the Doppler shift of the returning echo due to the bats forward motion. This means that the returning echo would fall outside of the range of the fovea. To compensate for this during flight, hence DSC, the bat emits its calls at a lower frequency than its resting frequency so that the frequency of the returning echo falls within the range of the acoustic fovea (Schnitzler and Flieger 1983; Neuweiler 1984; Schnitzler and Denzinger 2011).

This echolocation system overcomes the masking effects of clutter through the generation of acoustic glints (Neuweiler 1984; Schnitzler and Kalko 2001). HDC bats emit calls of long duration (>30 ms) dominated by a constant-frequency (CF) component with short frequency-modulated (FM) components at the beginning and end of the CF component (Schnitzler and Flieger 1983; Neuweiler 1984). These bats use the long-duration CF components in combination with DSC to generate constant echoes from the background. The bats do so by adjusting the frequency of their emitted call to compensate for the changes in frequency resulting from the Doppler shift in frequency caused by its flight speed relative to the stationary background. The bat is then able to detect flying targets by the acoustic glints superimposed on this constant background echo. Acoustic glints are changes in frequency and amplitude generated when the bat's call is reflected off the fluttering wings of flying insects during the insect's wing beat cycle (Schnitzler and Flieger 1983; Neuweiler 1984). When the long CF signal impinges upon a wing, the amplitude and frequency of the echo are dependent on the position of the wing and whether the wing is moving towards the bat or away from it. The amplitude of an echo is dependent on the size of the object generating it. When the insect wing is perpendicular to the impinging echolocation call at the top or bottom of its wing beat cycle, all of its surface area reflects the call, producing an echo of high amplitude. When the wing is parallel to the bat, only its edge reflects its call and a weak echo is generated. Similarly when the wing is moving down and towards the bat during the first part of the power stroke or up and towards the bat during the recovery stroke, the frequency of the impinging echolocation call is Doppler shifted to a higher frequency. In the same way, when the wing is moving away from the

impinging call towards the top or bottom of its beat cycle, the frequency of the echo is Doppler shifted lower than that of the bat's signal. These changes in the amplitude and frequencies of echoes from insect wings are perceived by the bat as amplitude and frequency glints against the constant echo from the background clutter (Schnitzler and Flieger 1983; Neuweiler 1984), and this allows it to detect flapping insect wings against background clutter.

The specialized echolocation system of HDC bats restricts them to foraging in dense vegetation where distances to background vegetation are short and atmospheric attenuation (Lawrence and Simmons 1982) is reduced allowing detectable acoustic glints. Flying in dense vegetation means that these bats have to fly slowly and manoeuvrably which requires short broad wings that generate lift at low flight speeds (e.g. Figure 6.1; Norberg and Rayner 1987). The wings and echolocation system of bats therefore form an adaptive complex (Aldridge and Rautenbach 1987; Norberg and Rayner 1987) which is constrained by the acoustic and aerodynamic requirements of their specialized ecological niche. This could result in particularly pronounced convergent morphology and echolocation in the family Rhinolophidae.

Morphology and echolocation in rhinolophids are indeed highly convergent throughout their Old World distribution. Despite differences in size, wing shapes are remarkably similar across rhinolophid species (Fig. 6.1). This similarity in morphology has resulted in the 78 or so recognized species being placed in a single genus (*Rhinolophus*) (Csorba et al. 2003). The genus consists of two major phylogenetic clades: an Afro-Palaearctic clade and an Asian clade (Stoffberg et al. 2011; Dool et al. 2016, Foley et al. 2015). Furthermore, these bats are known to use their habitats in the same way and are likely, therefore, to encounter similar prey. This may result in the convergence of skull morphology as well. However, it is evident that, although there might be evolutionary trade-offs that constrain some species causing convergence, given the wide range in body size and call parameters, these organisms are obviously responding to other selection pressures as well. How organisms respond to different selection pressures and the trade-offs involved as a result of responding to a suit of selection pressure is an interesting area of enquiry. African rhinolophids currently comprise 27 species (Happold and Cotterill 2013) with many of them having wide geographic distributions (Monadjem et al. 2010). Rhinolophids are thus ideal for investigating the role of evolutionary constraints on convergence.

Evidence for evolutionary constraints as the cause of convergence has to account for the role of shared ancestry as the cause of phenotypic similarity. We therefore investigated the role of evolutionary constraints in the evolution of phenotypic convergence in the Southern African rhinolophids in a phylogenetic context. We used the recent phylogeny reported in (Dool et al. 2016) to do so. If convergence stems from stasis as a result of evolutionary constraints, then the phenotypes of most or all of these species should converge on each other and the ancestral phenotype. Furthermore, there should be more extant species that echolocate at or above the ancestral frequency than below because higher frequencies facilitate DSC (Waters 2003).

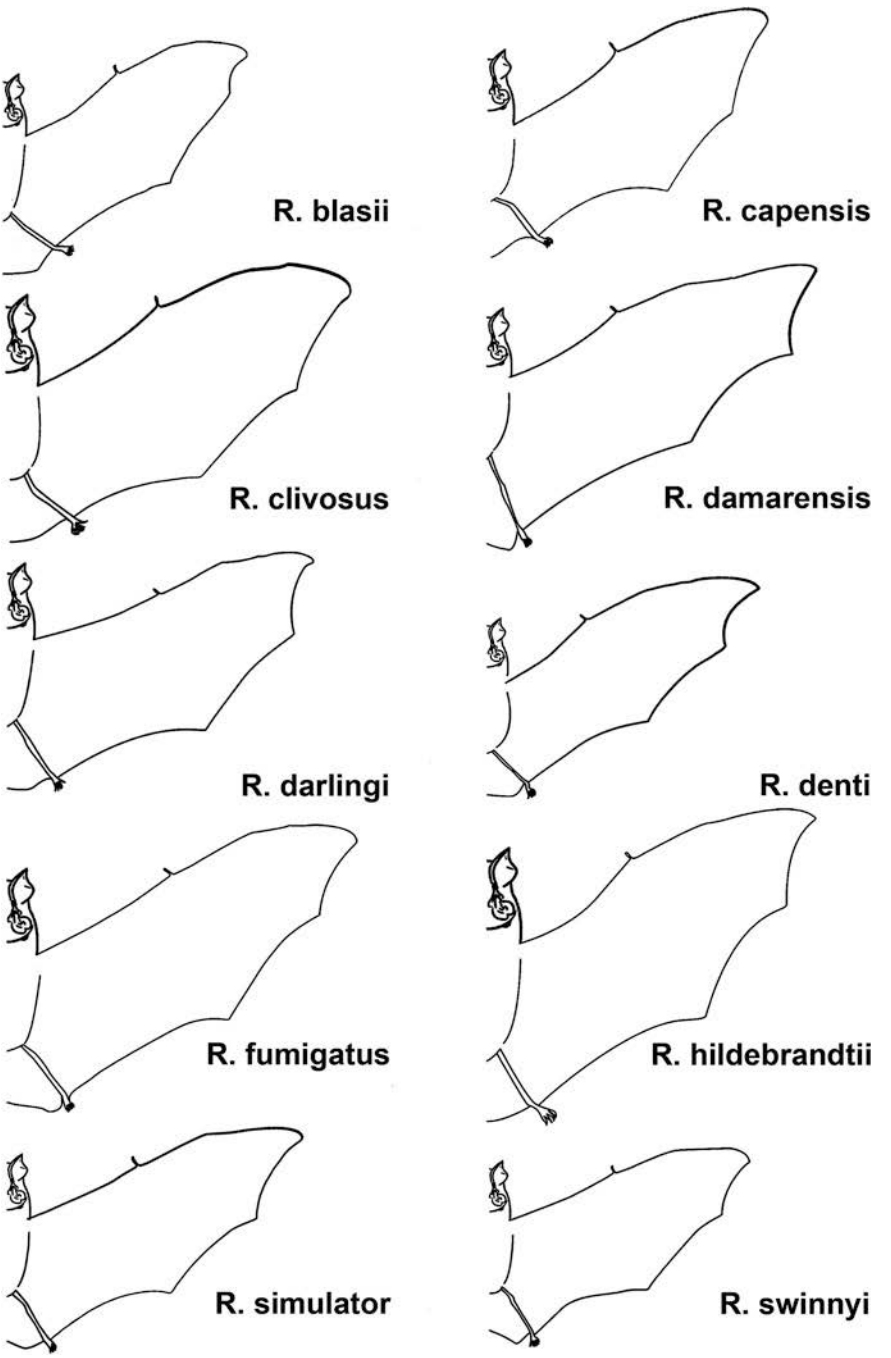


Fig. 6.1 Outline of the wings of several Southern African rhinolophids

6.2 Methods

6.2.1 Taxonomic Notes

We investigated phenotypic convergence in 12 species of African Rhinolophidae (Table 6.1). Lineages within the family Rhinolophidae are poorly resolved and are likely to contain several cryptic species (Dool et al. 2016). Although this may be true for several groups in this family, it is particularly evident in the *fumigatus* group which includes *Rhinolophus fumigatus*, *R. damarensis*, *R. darlingi*, *R. eloquens* and *R. hildebrandtii* (Csorba et al. 2003; Jacobs et al. 2013; Dool et al. 2016). In this study, we used the species designations of Dool et al. (2016). *R. fumigatus* from the western part of Southern Africa appears to be a distinct but sister lineage to *R. fumigatus* from the eastern half of Africa (Dool et al. 2016). We thus treat *R. fumigatus* from the eastern part of the continent as *R. fumigatus* sensu stricto since the type specimen for this species comes from Ethiopia in East Africa. We refer to the lineage from the west as *R. cf. fumigatus*. We place all *R. hildebrandtii* with resting call frequencies between 37 and 39 kHz in *R. cf. mossambicus* (Dool et al. 2016). We treated all other rhinolophids with resting frequencies of between 42 and 46 kHz as belonging to *R. hildebrandtii* for the following reasons: (1) there is currently no genetic data that allow us to place them in the species designations erected by Taylor et al. (2012), (2) the range in resting frequency is small and (3) we found similar ranges within the same roost, e.g. 42–44 at Mushandike (20°7′ S, 30°35′ E), Zimbabwe, and 44–46 kHz at big Baobab tree (22°30′ S, 30°37′ E), South Africa (Table 6.1). Furthermore, we used the designation of *R. cf. simulator* for a lineage that displayed genetic similarity to *R. simulator* (Dool et al. 2016) but echolocated at a much higher frequency (Table 6.1).

6.2.2 Sampling

Bats were caught from caves and disused mine shafts across their distributional ranges in Southern, Eastern and Central Africa using hand nets and continuously monitored harp traps and mist-nets. Captured bats were held individually in soft cotton bags. Sex and reproductive status were checked immediately following capture, and juveniles, pregnant or lactating bats were released immediately. Reproductive status was determined by examination of the nipples and palpation of the abdomen of female bats (Racey 1988). Juveniles were distinguished from adults by the presence of cartilaginous epiphyseal plates in their finger bones detected by trans-illuminating the bat's wings (Anthony 1988).

The forearm length (FA) to the nearest 0.1 mm and body mass to the nearest 0.1 g was measured using dial callipers and a portable electronic balance, respectively. We also measured the upper tooth-row length (CM^3), head height (HH), head width (HW), head length (HL), and nose-leaf width (NLW) to the nearest 0.1 mm using a dial callipers (Table 6.2).

Table 6.1 Means ± SD and ranges (in parentheses below the means) of phenotypic parameters of Southern African rhinolophids. Sample sizes are given below species names

Species	Mass (g)	Forearm (mm)	CM^3 (mm)	HH (mm)	HW (mm)	HL (mm)	NLW (mm)	RF (kHz)
Rbl (15)	9.3 ± 0.5 (8.4–10.5)	45.4 ± 0.7 (43.7–46.5)	6.6 ± 0.4 (6.0–7.0)	6.7 ± 0.2 (6.1–7.0)	9.3 ± 0.5 (8.4–10.1)	18.1 ± 1.3 (16.0–19.7)	8.2 ± 0.7 (7.1–9.3)	90.9 ± 1.1 (87.6–91.8)
Rca (62)	12.0 ± 1.1 (9.1–14.6)	50.4 ± 1.3 (47.4–53.5)	7.8 ± 0.7 (6.5–9.8)	9.2 ± 1.4 (6.1–12.7)	10.0 ± 1.2 (7.5–13.6)	19.8 ± 1.6 (17.2–22.1)	±8.50.5 (7.4–9.3)	81.8 ± 3.7 (75.0–86.8)
Rcl (54)	18.0 ± 2.0 (14.1–21.7)	54.8 ± 1.8 (52.3–59.4)	8.3 ± 0.7 (6.7–10.2)	9.8 ± 1.4 (7.6–13.8)	10.5 ± 1.3 (8.1–13.8)	21.8 ± 1.3 (19.8–24.1)	7.6 ± 0.5 (6.3–8.5)	91.6 ± 1.0 (88.2–93.5)
Rda (19)	10.7 ± 0.6 (9.3–11.5)	49.5 ± 2.0 (47.3–51.3)	7.2 ± 0.5 (6.2–8.2)	7.5 ± 0.4 (6.6–8.1)	9.1 ± 0.6 (7.5–9.8)	19.0 ± 1.3 (17.1–20.7)	8.1 ± 0.4 (7.2–8.8)	86.7 ± 1.4 (84.7–89.1)
Rdr (28)	8.9 ± 1.6 (6.2–13.8)	46.5 ± 1.3 (44.0–49.3)	7.2 ± 0.7 (5.8–8.2)	7.8 ± 0.7 (6.6–9.8)	9.7 ± 0.3 (9.1–10.0)	19.7 ± 1.4 (17.1–21.9)	8.4 ± 0.4 (7.7–8.9)	81.3 ± 2.5 (78.4–84.5)
Rde (20)	6.8 ± 1.2 (5.7–8.8)	42.7 ± 0.9 (41.5–44.3)	6.1 ± 0.4 (5.2–6.5)	6.0 ± 0.7 (5.0–7.2)	7.8 ± 0.7 (6.7–8.9)	15.7 ± 1.2 (13.5–17.5)	6.7 ± 0.4 (6.0–7.3)	111.8 ± 2.7 (107.5–115.6)
Rel (20)	19.8 ± 1.0 (18.0–21.5)	58.7 ± 1.9 (56.0–63.2)	–	–	–	–	12.5 ± 0.5 (12.0–13.0)	45.1 ± 0.6 (43.9–45.7)
Rfu (32)	13.1 ± 1.3 (10.9–16.8)	52.0 ± 1.4 (48.5–54.5)	8.2 ± 0.6 (7.1–8.9)	9.0 ± 0.6 (8.2–10.0)	11.1 ± 0.9 (9.9–12.6)	23.4 ± 2.1 (18.9–27.9)	10.6 ± 0.9 (8.1–12.3)	55.6 ± 1.4 (53.3–57.3)
R. cf. fu (38)	18.8 ± 1.4 (16.2–22.4)	58.6 ± 1.7 (55.1–62.0)	–	–	–	–	–	55.0 ± 0.9 (53.0–56.5)
Rhi (30)	28.7 ± 2.8 (25.0–33.5)	63.8 ± 1.6 (62.1–65.9)	9.5 ± 1.0 (8.3–11.9)	9.5 ± 1.0 (10.0–11.9)	11.3 ± 1.0 (11.1–16.2)	27.4 ± 1.4 (23.9–28.9)	12.3 ± 0.7 (11.2–13.9)	44.8 ± 1.4 (43.3–47.1)
Rmo (22)	27.7 ± 5.1 (23.6–38.8)	62.6 ± 1.5 (60.0–66.9)	10.2 ± 0.7 (8.6–11.3)	12.3 ± 0.9 (11.0–14.9)	13.3 ± 1.2 (11.1–14.6)	26.4 ± 1.4 (25.2–29.1)	13.4 ± 0.6 (12.3–14.4)	38.8 ± 0.7 (36.8–39.7)
Rsi (20)	7.3 ± 1.0 (6.2–10.3)	44.7 ± 1.3 (43.4–47.1)	6.7 ± 0.6 (6.1–7.9)	7.0 ± 0.4 (5.8–8.2)	9.1 ± 0.6 (7.8–9.9)	18.3 ± 1.4 (17.1–21.0)	7.9 ± 0.4 (7.3–8.8)	81.8 ± 2.3 (77.4–84.9)

(continued)

Table 6.1 (continued)

Species	Mass (g)	Forearm (mm)	CM^3 (mm)	HH (mm)	HW (mm)	HL (mm)	NLW (mm)	RF (kHz)
R. cf. si (24)	5.8 ± 0.7 (5.0–7.1)	41.5 ± 1.4 (39.1–44.3)	6.5 ± 0.4 (5.7–7.5)	6.6 ± 0.3 (6.0–7.3)	8.6 ± 0.3 (8.3–9.4)	17.4 ± 1.0 (15.6–19.1)	7.4 ± 0.4 (6.7–7.9)	103.5 ± 1.6 (100.5–106.5)
Rsw (22)	± 7.60.7 (7.0–9.8)	± 43.70.9 (42.5–44.8)	–	–	–	–	–	107.20.6 (106.5–108.1)

Abbreviations: *Rbl* = *R. blasii*; *Rca* = *R. capensis*, *Rcl* = *R. clivosus*; *Rda* = *R. damarensis*, Rdr = *R. darlingi*; *Rde* = *R. denti*; *Rel* = *R. eloquens*; *Rfu* = *R. fumigatus*; *R. cf. fu* = R. cf. fumigatus; *Rhi* = *R. hildebrandtii*; *Rmo* = *R. mossambicus*; *Rsi* = *R. simulator*, *R. cf. si* = *R. cf. simulator*; *Rsw* = *R. swinnyi*; CM^3 = length of the upper tooth row; HH = head height, HW = head width; HL = head length, NLW = nose-leaf width; RF = resting frequency

Table 6.2 Phenotypic parameters measured from live bats in the field

Abbreviation	Name	Description
NLW	Nose-leaf width	The broadest distance across the two leafs measured in millimetres
CM^3	Upper tooth-row length	Upper tooth-row length (measured in millimetres) from the end of the last molar to the front end of the incisor
HH	Head height	Head height measured (in millimetres) from beneath the jaw to the topmost tip of the head
HW	Head width	Head width measured (in millimetres) from behind the two ears
HL	Head length	Condylobasal length (measured in millimetres) from the nose tip to behind the nap

Echolocation calls from hand-held bats were recorded and analysed as in Mutumi et al. (2016). Hand-held calls allow the determination of resting frequency (RF; frequency of maximal energy when at rest) in rhinolophid bats (Siemers et al. 2005) and eliminate variations in frequency as a result of rhinolophid bats compensating for Doppler shifts during flight (Schnitzler 1987).

In addition to the RFs reported here, we also obtained the RFs from other species from Csorba et al. (2003) and Zhou et al. (2009) to compare the number of extant species with RFs equal to, above or below that of the reconstructed ancestral frequency (Dool et al. 2016).

6.2.3 Statistical Analyses

As far as possible, we kept sexes equal to account for potential sexual dimorphism. We conducted a principal component analysis (PCA) on 8 phenotypic variables (Table 6.1) to extract 8 independent and uncorrelated principal components from the original set of variables to meet the assumptions of discriminant function analysis (DFA). DFA was done on the factor scores of the first two principal components to examine instances of multivariate phenotypic convergence within African rhinolophids. Prior to PCA, variables were $\log_{10}$ transformed. All statistical analyses were done in Dell Statistica (version 13, Southern African Analytics Pty Ltd.).

6.3 Results

The first two principal components (PC) recovered from the PCA explained 87.4 % of the variation (PC 1—79.5 %; PC2—6.4 %). The two roots extracted by DFA on these two principal components explained 100 % of the variation (Table 6.3) PC 1

Table 6.3 Results of discriminant function analysis on principal component scores extracted by principal component analyses on 8 phenotypic variables (Table 6.1)

	Root 1	Root 2	Wilks' λ	$F_{(10, 305)}$	P
PC 1	−5.63	−0.21	0.15	968.0	<0.0001
PC2	0.47	−2.52	0.03	176.98	<0.0001
Eigenvalue	31.92	5.62			
Cumulative (%)	85.03	100			
Wilks' λ	0.004	0.15			
χ	1666.40	584.97			
df	20	9			
P	<0.0001	<0.0001			

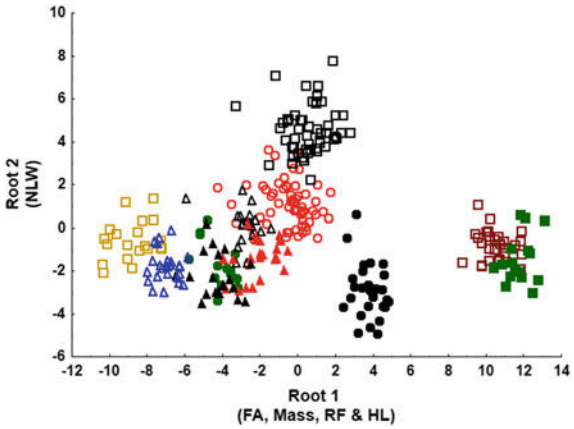


Fig. 6.2 Plot of canonical scores obtained from discriminant function analyses on phenotypic parameters (Table 6.1) for several Southern African rhinolophid species. Some of the phenoptypes with heaviest loadings are given below each root (abbreviations are given in Table 6.1). Key to species: *R. blasii—solid green circles*; *R. capensis*—open *red circles*; *R. clivosus*—open *black squares*; *R. damarensis*—open *black triangles*; *R. darlingi*—solid *red triangles*; *R. denti*—open *gold triangles*;; *R. fumigatus* (east)—solid *black circles*; *R. hildebrandtii*—open *brown squares*; *R. mossambicus*—solid *green squares*; *R. simulator—solid black triangles*; *R. cf. simulator*—open *blue triangles*

(associated mainly with FA, RF, mass and head length) loaded the highest on Root 1 and PC 2 (associated with NLW) loaded the highest on Root 2 (Table 6.3; Fig. 6.2).

All species were tightly clustered together with the exception of the three largest species using the lowest call frequencies, *R. fumigatus*, *R. hildebrandtii* and *R. mossambicus* (Fig. 6.2; Table 6.1). These species loaded highest on Root 1 whilst the rest loaded lowest on Root 1 (Fig. 6.2). However, the squared Mahalanobis distances between all species were significant ($F_{(2, 305)}$s > 14, Ps < 0.00001) with the exception of the mahalonobis distances between *R. blasii* and *R. simulator*

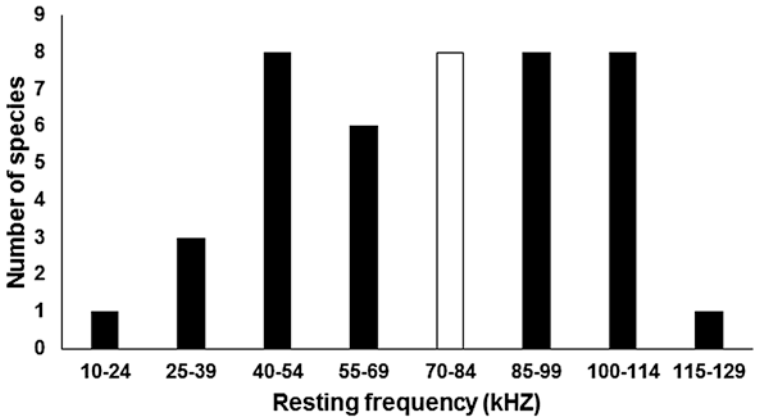


Fig. 6.3 Distribution of resting frequencies of extant rhinolophids (*black bars*) in relation to the ancestral frequency band. The *white bar* represents the ancestral frequency band and the extant bats that have resting frequencies within it

($F_{(2,\ 305)} = 1.14$, P = 0.32; Fig. 6.2). Although the mahalonobis distance between *R. damarensis* and *R. darlingi* (mahalonobis distance = 0.27; $F_{(2,\ 305)} = 14.50$, P < 0.00001; Fig. 6.2) and between *R. hildebrandtii* and *R. mossambicus* (mahalanobis distance = 1.92; $F_{(2,\ 305)} = 10.65$, P < 0.0001; Fig. 6.2) were significant these two pairs of species had the lowest significant mahalanobis distances. These results reflect why *R. hildebrandtii* and *R. mossambicus* were formerly placed in a single species (Taylor et al. 2012). Lastly, there was complete overlap in the RFs of *R. fumigatus* and *R. cf. fumigatus* but almost no overlap in their masses and forearm lengths (Table 6.1).

There was little difference in the number of extant rhinolophid species that echolocated above or below the ancestral frequency band (Fig. 6.3). Similarly, the number of extant species echolocating within or above the ancestral resting frequency was slightly higher (24 vs. 17 species), but not significantly so ($\chi^2 = 1.19$; df = 1, P > 0.1), than those echolocating below the ancestral frequency (Fig. 6.3).

6.3.1 Discussion

Most of the species considered converged on the 95 % confidence intervals of the ancestral character state calculated as 45.6–54.8 mm for forearm length and 72.6–81.8 kHz for resting frequency for all extant rhinolophids (Dool et al. 2016). The phenotypes of the species considered here ranged in forearm length and resting frequency from 39–54 mm and 75–92 kHz (Table 6.1), respectively. Convergence on the ancestral phenotype was particularly evident in the *capensis* group, but also for two species pairs, *R. damarensis*/*R. darlingi* and *R. hildebrandtii*/*R.*

mossambicus in the *fumigatus* group and *R. blasii* which is placed in a basal position to the Afro-Palaearctic clade (Fig. 6.2, Table 6.1; Dool et al. 2016). Mass did not carry any phylogenetic signal (Dool et al. 2016) and is not therefore considered in these comparisons.

Convergence is also evident between species pairs. There is pronounced convergence between *R. damarensis* and *R. darlingi* (Fig. 6.2), which occurs in separate clades within the *fumigatus* group (Jacobs et al. 2013; Dool et al. 2016) in support of the study by Jacobs et al. (2013). Similar phenotypic convergence is evident between *R. hildebrandtii* and *R. mossambicus* in the *fumigatus* group (Fig. 6.3) and between *R. simulator* of the *capensis* group and *R. blasii* (Fig. 6.2). The RFs of the two *fumigatus* lineages also converge but not their body sizes (Table 6.1). Thus, there is convergence in species that have overlapping distributions (*R. hildebrandtii* and *R. mossambicus*; *R. simulator* and *R. blasii*) and in species with disjunct distributions (*R. damarensis* and *R. darlingi*; *R. cf. fumigatus* and *R. fumigatus*) suggesting that in at least some cases local adaptation to the same habitats may not be the cause of the convergence.

The convergence described here may result from one or more of several processes such as inheritance from a common ancestor, adaptation to similar environments, random genetic drift and shared constraints (Losos 2011; Harmon et al. 2005). It is unlikely that the convergence we report here is the result of inheritance from a common ancestor. *R. mossambicus* is a sister lineage to *R. cf. fumigatus* and not to *R. hildebrandtii* (Dool et al. 2016) despite *R. mossambicus* and *R. hildebrandtii* being formerly placed in the same species (see Taylor et al. 2012). Similarly, *R. simulator* and *R. blasii* are also from different clades (Dool et al. 2016).

Furthermore, the fact that there are genetically closely related lineages within the *fumigatus* and *capensis* groups that are phenoptypically divergent probably excludes biological constraints, such as evolutionary trade-offs, as an explanation for convergence between several lineages. Such divergence indicates that despite their highly specialized echolocation system, rhinolophids are nevertheless able to respond evolutionarily to other forces in their environment. This is supported by the wide divergence from the ancestral body size (Table 6.1) and resting frequency (Fig. 6.3) observed amongst species of rhinolophids. The role of other processes (e.g. selection or drift) is supported by the absence of bias in the distribution of RFs amongst extant bats (Fig. 6.3), towards the ancestral RF or higher frequencies. This is so despite more pronounced Doppler shifts at higher frequencies which allow the acoustic fovea of these bats to encompass a wider range of frequencies. A wider range means that these bats can use less precise shifts in frequency (Waters 2003), facilitating DSC. It would appear that shifts towards lower frequencies, which would make DSC more difficult, would require substantial selection pressure.

Local adaptation also appears to be an unlikely explanation for the convergence at least between *R. damarensis* and *R. darlingi* which occur in different biomes. Random genetic drift may offer a better explanation for convergence between these two lineages (Jacobs et al. 2013). However, the two pairs of lineages *R. blasii*/*R. simulator* and *R. hildebrandtii*/*R. mossambicus* occur in the same biomes and some

times in the same roost (DSJ, unpublished data) suggesting that adaptation to similar habitats could offer a valid explanation for the convergence in these pairs of lineages. However, one cannot exclude the possibility that both these species pairs have largely retained the character states of their closest common ancestor which had a FA length of 49.4 mm and a RF of 86.4 kHz (Stoffberg et al. 2011), with some divergence from the ancestral character state due to adaptation (e.g. for discrete frequency bands, see Mutumi et al. 2016) or genetic drift (Jacobs et al. 2013). Stasis may also explain convergence in RF between the two sibling lineages, *R. cf. fumigatus* and *R. fumigatus*. However, the divergence in their body sizes suggests that adaptation to local environments, at least in body size, may play a role in the evolution of these lineages. *R. cf. fumigatus* occurs in the more arid western half of Southern Africa whilst *R. fumigatus* is distributed in the more eastern half of Africa. The larger body size in *R. cf. fumigatus* could therefore be an advantage in the more arid and cooler conditions that prevail over its geographic range (Monadjem et al. 2010) in accordance with James' Rule (James 1970). Why there was not a concomitant allometric response in the RFs (see Jacobs et al. 2007) of these two lineages remains to be determined.

Phenotypic divergence in the African Rhinolophidae is also evident. It is particularly pronounced in the *fumigatus* group (Fig. 6.2). Although there is pronounced convergence between *R. damarensis* and *R. darlingi*, close relatives of these two species in the *fumigatus* group, *R. eloquens*, *R. fumigatus*, *R. cf. fumigatus*, *R. hildebrantii* and *R. mossambicus*, have diverged appreciably from both the ancestral character state and *R. damarensis* and *R. darlingi* (Fig. 6.2, Table 6.1) supporting convergence as a result of other processes besides stasis. Furthermore, there is a pronounced repeated pattern of divergence in RF amongst pairs of sibling species where one member of the pair retains the ancestral frequency and the other member of the pair diverges appreciably with RFs > 100 kHz. For example, the RF of *R. simulator* is similar to that of the ancestral frequency but its sibling lineages, *R. cf. simulator* and *R. denti*, have RFs > 100 kHz. The same is true for *R. capensis* and its sibling species *R. swinnyi* (Table 6.1). This suggests that RF may play an integral role in lineage diversification in this family of bats despite echolocation being primarily involved in orientation and prey detection. Its role in lineage diversification may be mediated by its secondary function in communication (Bastian and Jacobs 2015).

In conclusion, we have identified several instances of convergence amongst African rhinolophids. It is unlikely that the convergences reported here is the result of evolutionary trade-offs or some other kind of biological constraint. However, further investigation of other processes that may be responsible for convergence in the Rhinolophidae is likely to offer great insight into the evolution of convergence in bats in general and in other taxa as well. The success of such investigations is entirely dependent on robust pholygenies and accurate determination of ancestral character states.

References

Aldridge HDJN, Rautenbach IL (1987) Morphology, echolocation and resource partitioning in insectivorous bats. J Anim Ecol 56:763–778

Anthony ELP (1988) Age determination in bats. In: Kunz TH (ed) Ecological and behavioral methods for the study of bats. Smithsonian Institution Press, Washington DC, pp 47–58

Ballentine B (2006) Morphological adaptation influences the evolution of a mating signal. Evolution 60:1936–1944

Bastian A, Jacobs DS (2015) Listening carefully: increased perceptual acuity for species discrimination in multispecies signalling assemblages. Anim Behav 101:141–154

Colborn J, Crabtree RE, Shaklee JB, Pfeiler E, Bowen BW (2001) The evolutionary enigma of bonefishes (*Albula* spp.): Cryptic species and ancient separations in a globally distributed shorefish. Evolution 55(4):807–820

Csorba G, Ujhelyip P, Thomas N (2003) Horseshoe bats of the world (Chiroptera: Rhinolophidae). Alana Books, Shropshire

Dool SE, Puechmaille SJ, Foley NM, Allegrini B, Bastian A, Mutumi GL, Maluleke TG, Odendaal LJ, Teeling EC, Jacobs DS (2016) Nuclear introns outperform mitochondrial DNA in inter-specific phylogenetic reconstruction: lessons from horseshoe bats (Rhinolophidae: Chiroptera). Mol Phylogenet Evol 97:196–212

Fenton MB (1999) Describing the echolocation calls and behaviour of bats. Acta Chiropterologica 1(2):127–136

Foley N, Vu dinh T, Goodman S, Armstrong K, Jacobs D, Puechmaille S, Teeling E (2015) How and why overcome the impediments to resolution: lessons from rhinolophid and hipposiderid bats. Mol Biol Evol 32(2):313–333

Futuyma DJ (1998) Evolutionary Biology. Sinauer Associates Inc, Sunderland, MA

Happold M, Cotterill FPD (2013) Family Rhinolophidae Horseshoe Bats. In: Happold M, Happold DCD (eds) Mammals of Africa: Volume IV: hedgehogs, shrews and bats. Bloomsbury Publishing, London, pp 301–303

Harmon LJ, Kolbe JJ, Cheverud JM, Losos JB (2005) Convergence and the multidimensional niche. Evolution 59:409–421. doi:10.1111/j.0014-3820.2005.tb00999.x. PubMed: 15807425

Jacobs DS, Barclay RMR, Walker MH (2007) The allometery of echolocation call frequencies of insectivorous bats: why do some species deviate from the pattern? Oecologia 152:583–594

Jacobs DS, Babiker H, Bastian A, Kearney T, van Eeden R, Bishop JM (2013) Phenotypic convergence in genetically distinct lineages of a Rhinolophus species complex (Mammalia, Chiroptera). PLoS ONE 8(12):e82614. doi:10.1371/journal.pone.0082614

Jacobs DS, Bastian A, Bam L (2014) The influence of feeding on the evolution of sensory signals: a comparative test of an evolutionary trade-off between masticatory and sensory functions of skulls in southern African Horseshoe bats (Rhinolophidae). J Evol Biol 27:2829–2840. doi:10.1111/jeb.12548

James FC (1970) Geographic size variation in birds and its relationship to climate. Ecology 51:365–390

Jones G, Siemers BM (2010) The communicative potential of bat echolocation pulses. J Comp Physiol A 197(5):447e457

Lawrence BD, Simmons JA (1982) Measurements of atmospheric attenuation at ultrasonic frequencies and the significance for echolocation by bats. J Acoust Soc Am 71(3):585–590

Losos JB (2011) Convergence, adaptation and constraint. Evolution 65:1927–1840. doi:10.1111/j.1558-5646.2011.01263.x. PubMed: 21729048

Monadjem A, Taylor PJ, Cotterill FPD, Schoeman MC (2010) Bats of Southern and Central Africa: a biogeographic and taxonomic synthesis. Wits University Press, Johannesburg

Mutumi GL, Jacobs DS, Henning W (2016) Sensory drive mediated by climatic gradients partially explains divergence in acoustic signals in two horseshoe bat species, Rhinolophus swinnyi and Rhinolophus simulator. PLoS ONE 11(1):e0148053. doi:10.1371/journal.pone.0148053

Neuweiler G (1984) Foraging, echolocation and audition in bats. Naturwissenschaften 71(9):446–455

Norberg UM, Rayner JMV (1987) Ecological morphology and flight in bats (Mammalia; Chiroptera): wing adaptations, flight performance, foraging strategy and echolocation. Philos Trans R Soc Lond B 316:335–427

Racey PA (1988) Reproductive assessment in bats. In: Kunz TH (ed) Ecological and behavioural methods for the study of bats. Smithsonian Institution Press, Washington, DC, pp 31–45

Reznick D, Nunney L, Tessier A (2000) Big houses, big cars, superfleas and the costs of reproduction. Trends Ecol Evol 15:421–425

Ridley M (1996) Evolution. Blackwell Science, Cambridge

Roff DA (2000) Trade-offs between growth and reproduction: an analysis of the quantitative genetic evidence. J Evol Biol 13:434–445

Roff DA, Fairbairn DJ (2007) The evolution of trade-offs: where are we? J Evol Biol 20(2):433–447

Schnitzler H-U (1987) Echoes of fluttering insects: information from echolocating bats. In: Fenton MB, Racey PA, Rayner JMV (eds) Recent advances in the study of bats. Cambridge University Press, Cambridge

Schnitzler H-U, Denzinger A (2011) Auditory fovea and Doppler shift compensation: adaptations for flutter detection in echolocating bats using CF-FM signals. J Comp Physiol A: Neuroethol Sensory Neural Behav Physiol 197(5):541–559

Schnitzler H-U, Flieger E (1983) Detection of oscillating target movements by echolocation in the greater horseshoe bat. J Comp Physiol 153(3):385–391

Schnitzler H-U, Kalko EKV (2001) Echolocation by insect-eating bats. Bioscience 51(7):557–569

Siemers BM, Beedholm K, Dietz C, Dietz I, Ivanova T (2005) Is species identity, sex, age or individual quality conveyed by echolocation call frequency in European horseshoe bats? Acta Chiropterologica 7:259–274

Stayton CT (2008) Is convergence surprising? An examination of the frequency of convergence in simulated datasets. J Theor Biol 252:1–14. doi:10.1016/j.jtbi.2008.01.008

Stayton CT (2015) The definition, recognition, and interpretation of convergent evolution, and two new measures for quantifying and assessing the significance of convergence. Evolution 69:2140–2153

Stoffberg S, Jacobs DS, Matthee CA (2011) The divergence of echolocation frequency in Horseshoe Bats: moth hearing, body size or habitat? J Mamm Evol 18(2):117–129

Taylor PJ, Stoffberg S, Monadjem A, Schoeman MC, Bayliss J, Cotterill FPD (2012) Four new bat species Rhinolophus hildebrandtii complex reflect plio-pleistocene divergence of dwarfs and giants across an afromontane archipelago. PLOS ONE 7(9): e41744. doi:10.1371/journal.pone.0041744. PubMed: 22984399

Thomas JA, Moss CF, Vater M (2004) Echolocation in bats and dolphins. University of Chicago Press, Chicago, IL

Waters DA (2003) Bats and moths: what is there left to learn? Physiol Entomol 28(4):237–250

Zhou ZM, Guillen-Servent A, Lim BK, Eger JL, Wang YX, Jang X-L (2009) A new species from southwestern China in the Afro-Palearctic lineage of the Horseshoe bats (*Rhinolophus*). J Mammal 90:57–73

Chapter 7
Convergence and Parallelism in *Astyanax* Cave-Dwelling Fish

Joshua B. Gross

Abstract The blind Mexican cavefish, *Astyanax mexicanus*, has emerged as a powerful model system for informing evolutionary and biomedical problems. A number of fascinating features evolve in cave-dwelling lineages, irrespective of phylogeny, or geography. For instance, cave-dwelling animals frequently lose vision and pigmentation, but evolve substantial gains in non-visual sensation. *Astyanax* cavefish are interesting models for cave evolution because they do not exist as a solitary group or population. Rather, there are 29 different *Astyanax* cave populations distributed across the Sierra de El Abra region of Mexico. Until recently, our understanding of how different cave populations relate to one another, and to surface-dwelling forms, was limited. We now know that there were two principal invasions over the past several millions of years—an 'older' stock invaded ~2–5 My ago and seeded the southern El Abra caves. The descendants of the surface-dwelling fish from this older stock have gone locally extinct. A second, 'younger' stock invaded the region ~1–2 My ago and seeded the northern Guatemala caves and the western Micos cave network. Descendants of the second invasion persist as extant surface-dwelling fish in the surrounding rivers. Recent population genetic evidence indicates that substantial 'mixing' has occurred between many caves populations, as well as between cave and surface-dwelling forms. This highly complex biogeography has raised many questions about the nature of cave-associated traits. Do these traits evolve once and spread throughout multiple cave systems? Or, alternatively, do cave-related traits evolve repeatedly in different cave complexes? This chapter discusses experimental approaches that have shed light on this question—complementation crosses and candidate gene analyses. Interestingly, complementation analyses show that vision loss and pigmentation loss evolve through similar and different genomic regions, respectively. Vision loss occurs throughout the Sierra de El Abra, but the extent to which the

J.B. Gross (✉)
Department of Biological Sciences, University of Cincinnati, Rieveschl Hall Room 711B, 312 Clifton Court, Cincinnati, OH 45221, USA
e-mail: grossja@ucmail.uc.edu

P. Pontarotti (ed.), *Evolutionary Biology*, DOI 10.1007/978-3-319-41324-2_7

same genes are implicated in different eyeless caves decreases with geographic distance between caves. In contrast, pigmentation changes affecting two Mendelian traits (albinism and *brown*) appear to evolve through the same genes, via distinct mutations. Standing genetic variation, present in local surface-dwelling forms, also plays a key role in the evolution of reduced eye sizes, albinism, and appetite control. Thus, evolution of cave-associated phenotypes proceeds through a complex mosaic of both convergent and parallel processes. These patterns reflect the complex evolutionary and geographic origins of cavefish in this region of Mexico. Future studies based on genome-level analyses will provide new insight to the pace and mechanism(s) of cave evolution in this emerging model system.

7.1 Introduction

Owing to the presence of ~29 different cave-dwelling populations, the blind Mexican cavefish is an intriguing system for evaluating the genetic and genomic bases for trait evolution (Mitchell et al. 1977; Gross 2012). Although a great deal of insight has been gained in the past several decades—including isolation of specific genes mediating cave-associated phenotypes—much remains unknown. In particular, the extent to which the same phenotypes are evolving through similar versus diverse mechanisms is just beginning to be elucidated. This is attributed to at least two crucial challenges.

The first relates to the complex historical biogeography of *Astyanax* fish in the Sierra de El Abra region (Mitchell et al. 1977). There have been at least two major 'waves' of migration of surface-dwelling forms into the region over the past several millions of years (Bradic et al. 2012, 2013). Extant cave populations are therefore regarded as having evolved from either the first, older stock of epigean ('surface-dwelling') forms ~2–5 My ago, or the younger stock of epigean fish that invaded the region ~1–2 My ago (Ornelas-García et al. 2008). Older populations comprise the so-called El Abra lineage, and most are centered around the southern El Abra karst complex (Fig. 7.1). The younger populations are set to the north (the 'Guatemala' populations) and to the west (the 'Micos' populations; Fig. 7.1). Evidence from a number of genetic studies reveals gene flow between cave and surface populations, as well as between different cave populations (Bradic et al. 2012). The assignment of convergent versus parallel evolution rests on knowledge of the relatedness of species, or subspecies, evolving the same phenotype. Therefore, the complex history of invasions, reinvasions, hybridization, and gene flow over the past several millions of years has created an important challenge to our understanding of how different cave populations are related to one another.

The second issue relates to the nature of phenotypic change in cave-dwelling animals. Irrespective of climate and geographic position, organisms that invade the cave microenvironment trend toward the same extreme phenotypic changes—loss

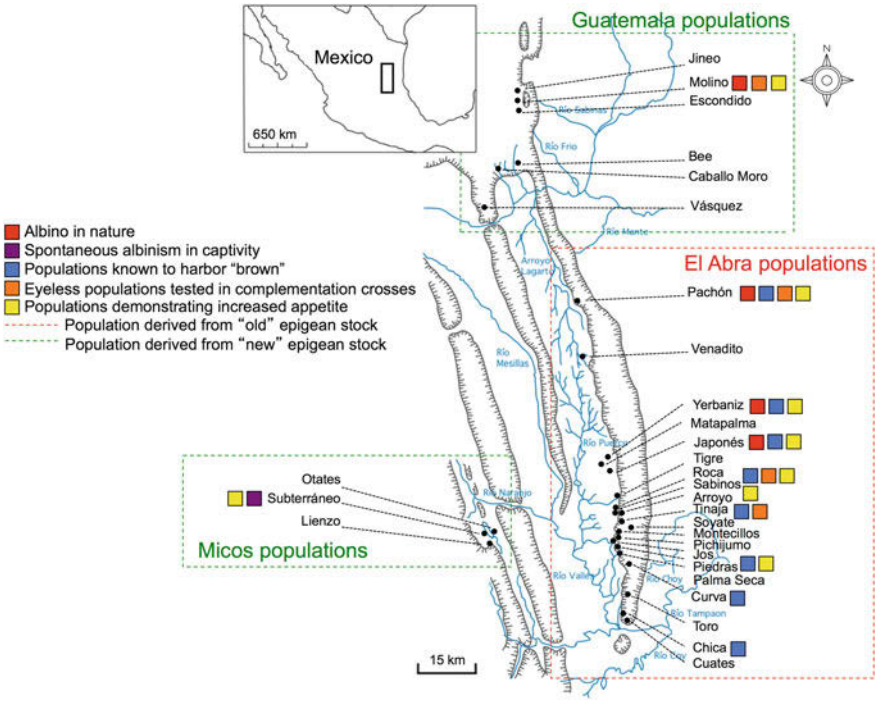


Fig. 7.1 Distribution of cave-associated phenotypes across 29 populations of *Astyanax* cavefish in the Sierra de El Abra. The complex evolutionary and biogeographical history of *Astyanax* cavefish has yielded a complicated distribution of several phenotypes, including albinism, *brown*, eyelessness, and increased appetite. The collective distribution of these phenotypes has been facilitated by gene flow between caves and surface populations, as well as selection for rare alleles present in the standing genetic variation of extant surface lineages

of pigmentation and vision (Protas et al. 2011; Stahl et al. 2015). Thus, convergence on troglomorphic (cave-associated) phenotypes is universal and likely arises more as a consequence of extreme environmental pressures (low nutrients, absence of light) rather than a reflection of local ecological pressures (Culver 1982). This can be problematic since the widespread convergence on similar characters renders the delineation of species, subspecies, and populations challenging for *Astyanax* cavefish (Ornelas-García et al. 2008). It has only been in the last several years, with the implementation of sophisticated population genetic studies incorporating next-generation sequencing, that a clearer picture of the level of relatedness of cavefish populations has emerged (Gross 2012). Given that the determination of evolution through convergence versus parallelism rests on degree of relatedness, the case studies presented in this report reflect the current knowledge of cavefish population relations that have emerged in recent years.

7.2 Complementation Studies—Conflicting Results for Pigmentation Regression and Vision Loss

7.2.1 Complementation of Pigmentation Regression

Wilkens and Strecker (2003) explored the extent to which the 'brown' phenotype, characterized by melanin reduction in melanophores, was present among disparate *Astyanax* cave populations. They reported the presence of 'brown' in four cave populations—including the Piedras, Curva, Tinaja, and Pachón cavefish localities (Wilkens and Strecker 2003). This was based on failure to complement this trait in direct crosses between individuals from each cave population. In their report, Wilkens and Strecker (2003) proposed either of two possibilities. First, if traits were observed in all cave populations, it would support the notion of a single origin that spread to multiple cave populations via subterranean connections. The second possibility they proposed was that each cave population independently evolved '*brown.*' Their results were consistent with the latter interpretation since '*brown*' was not found ubiquitously across the Sierra de El Abra. One of their examined populations was the northern Molino cave, derived from the second (more recent) 'wave' of epigean ancestors, which did not harbor the *brown* phenotype (Wilkens and Strecker 2003). Thus, the sporadic appearance of the *brown* phenotype across the El Abra karst landscape provided support for the notion that this trait has arisen independently through convergence in multiple populations.

An additional possibility, however, is that the trait arose once and spread to other cave populations through direct hybridization. Cave populations are generally believed to be isolated from one another and isolated from surface fish populations. However, Bradic et al. (2012) modeled a substantial amount of gene flow both in and out of the cave networks. Further, certain cave populations including the Chica cave and the Pachón cave have documented reports of hybridization between the surrounding surface-dwelling populations and the cave populations (Gross 2012). However, a second issue is the extent to which there may be a karstic 'connection' that unites two different cave populations (Culver 1982; Gross 2012). For instance, underground aquifers may harbor sizable populations shared by several geographically distinct cave populations. These aquifers would therefore provide a direct route for hybridization among different caves. The size, distribution, and influence of such aquifers on inter-cave gene flow remain unknown. However, more recent analyses of the *brown* phenotype provide some intriguing evidence that gene flow between certain caves may exist.

Gross et al. (2009) mapped the '*brown*' mutation to a genomic interval inclusive of the gene *Mc1r*. When *Mc1r* signaling was abrogated through morpholino injection, morphants displayed a phenocopy of the brown trait—depigmented eyes as juveniles, reduction in melanin content of melanophores (Gross et al. 2009). Interestingly, genetic mapping was carried out using individuals derived from the Pachón cave that harbored a 2-bp deletion (Δ23,24) near the 5′ end of the open reading frame of *Mc1r*, which causes a frameshift and prematurely truncated protein

(Gross et al. 2009). This loss-of-function mutation was evaluated in the other cavefish populations shown to harbor the '*brown*' phenotype through complementation studies and direct observation (Wilkens and Strecker 2003).

Surprisingly, the Δ23,24 mutation was not limited solely to the Pachón cavefish locality (although this was the sole mutation present among these individuals). This mutation was also present in the more southern Japonés and Yerbaniz populations at very low frequency (heterozygous condition in one sampled individual from each cave; Gross et al. 2009). This was particularly interesting since the only other functional mutation observed across 12 cavefish populations was found in the Japonés and Yerbaniz populations. The C490T point mutation was distributed widely among the sampled individuals from the Japonés and Yerbaniz caves (Gross et al. 2009). Interestingly, the remaining cave populations demonstrated no differences in the coding sequences of *Mc1r* compared to surface-dwelling fish. Incidentally, no coding sequence alterations (including Δ23,24 and C490T) were observed in any surface-dwelling fish populations.

The 'brown' trait represents an example of a Mendelian phenotype that has spread widely across the El Abra landscape (Wilkens 1988; Fig. 7.1). Although two coding sequence mutations have evolved which impact function of the causative gene, these mutations are largely restricted to specific caves or cave networks. Interestingly, however, these coding mutations are not *exclusive* to each cave population, but the only cave populations found to harbor the *brown* phenotype were the more ancient El Abra populations. The 'younger' cave populations to the north (Guatemala) and west (Micos) neither harbor the brown phenotype nor demonstrate any coding sequence alterations to *Mc1r* (Gross et al. 2009).

An important caveat of this study was the fact that the *brown* phenotype was present in many caves, but absent in several others (Gross et al. 2009). The complementation analysis inferred that the same gene (or closely linked genes) must govern the same phenotype in different populations. Consistent with this scenario, the authors suggested that the *brown* phenotype is present in other cave populations such as Piedras and Curva as a consequence *cis*-regulatory (as opposed to coding) sequence changes affecting the transcriptional abundance of the Mc1r receptor. Stahl and Gross (2015) evaluated this scenario and found that fish from the Pachón cave indeed demonstrate reduced expression of the *Mc1r* gene in addition to the loss-of-function frameshift mutation present in the open reading frame.

The reduced expression of *Mc1r* was also discovered in a number of other geographically distinct cave populations including the southern Chica and Tinaja caves (Stahl and Gross 2015). Neither of these latter two cave populations demonstrated any coding sequence alterations in *Mc1r*. This raised the possibility that reduction of *Mc1r* expression may be governed by convergent changes to the upstream regulatory sequence of this gene. Although the precise structure of the *Mc1r* regulatory complex is incompletely characterized, prior studies showed that the ~1 kb region 5′ of the open reading frame can function as a minimal promoter (Moro et al. 1999). When this region was sequenced and characterized in 12 cave populations, Stahl and Gross (2015) found a remarkably high level of sequence variation. This was quite surprising given that the coding sequence for *Mc1r* was

largely intact across every cave population evaluated. One locus, however, was highly conserved across these cave forms—a SNP impacting position 2148 bp (upstream of the translational start site). Interestingly, this SNP resides in a region of the minimal promoter that has been shown to reduce expression of the Mc1r receptor protein in other systems (Stahl and Gross 2015). Further, this SNP was absent from two populations that do not demonstrate the *brown* phenotype—surface-dwelling forms from the same geographic vicinity of the El Abra cave network and cavefish from the Molino cave locality (Stahl and Gross 2015).

The presence of this genetic lesion raises the intriguing possibility that this putative *cis*-regulatory mutation arose once during the evolutionary history of *Astyanax* cavefish. This mutation may have been strongly selected in the subterranean habitat (e.g., owing to elimination of the energetic cost associated with melanin production in total darkness). This allele may have spread widely across the El Abra landscape via direct contact among cave-dwelling forms, united throughout the cavernous karst subterranean architecture. Alternatively, this regulatory mutation may be present at very low frequency in surface-dwelling forms (although it has not been directly detected in these populations) and risen to high frequency in cave populations through selection.

Irrespective of the mechanism through which this mutation has become common in cave populations, how do we explain the fact that some *brown* cave populations harbor coding sequence variants in *Mc1r*, but others do not? One explanation may relate to the phylogeography of the region. For instance, it is clear that some cave populations are older than others—the 'El Abra' populations were seeded by epigean fish ~2–5 My ago, while the 'Micos' and 'Guatemala' populations were seeded ~1–2 My ago (Fig. 7.1). The cave populations demonstrating non-synonymous coding sequence alterations are uniformly derived from the older invasion, whereas the younger cave populations never demonstrate differences in the *Mc1r* open reading frame compared to extant surface-dwelling fish (Gross et al. 2009; Gross and Tabin 2010; Stahl and Gross 2015). Thus, it may be that *cis*-regulatory changes are favored and result in reduced expression of the Mc1r receptor. This would explain why putative regulatory impacts are more widely distributed across all cave populations compared to coding sequence changes. Then, for deeper lineages, coding sequence alterations may have arisen in *Mc1r* through accumulation of mutations that 'knocked out' the function of the receptor.

Future studies will clarify if the putative regulatory mutation affecting *Mc1r* is present within surface-dwelling populations. If documented, even at very low frequency, it will provide support for the notion that reduced function of *Mc1r* is negatively selected in natural populations of cavefish. Additional functional analyses will also support the interpretation that Mc1r function poses an energetic cost for fish living in complete darkness.

In sum, *brown* may be adopted widely throughout the Sierra de El Abra through both convergent and parallel processes. Further, the mechanisms explaining the evolution of *brown* cavefish would include a combination of direct (negative)

selection and neutral processes. The latter changes, affecting *Mc1r* coding sequence, may be favored as a consequence of more substantial periods of isolation in the oldest cave populations.

7.2.2 Complementation of Vision Loss

Complementation analyses provide a powerful tool for quickly determining whether the same gene(s) is/are implicated in the convergent similarities across different populations. Given the results of complementation studies of depigmentation (above), one might predict that the results observed for *brown* would be identical for other regressive traits. Interestingly, however, a very different scenario emerges for the evolution of eye loss (Fig. 7.1). The first complementation study, published in 1971, demonstrated that the offspring of cavefish from the Pachón and Sabinos caves (both eyeless populations) had more developed eyes compared to either parent (Wilkens 1971). This landmark study indicated that the causative 'eye loss' genes were different in each of two different cave populations. In addition to the functional implications for the genetic basis of eye loss, this study provided the first direct evidence for the independent colonization of numerous Mexican caves by ancient epigean colonizers (Wilkens 1971).

There are several additional cave populations, however, that demonstrate eye loss beyond the Pachón and Sabinos caves. To determine whether the same complementation of vision can be demonstrated more widely, Borowsky (2008) performed a number of additional breeding analyses. The inter-cave population crosses confirmed and extended the work of Wilkens (1971). However, Borowsky (2008) noted that recovery of a visual system was usually an incomplete phenotype (Borowsky 2008). Indeed, an important and long under-appreciated aspect of this work was the finding of substantial variation in the recovered visual system. Borowsky (2008) formalized the recovery of functional vision through an optokinetic reflex assay, in which retinal movement in response to alternating black-and-white vertical 'stripes' indicates intact vision. This study showed that offspring derived from two eyeless parental stock do not uniformly demonstrate the same level of vision recovery (Borowsky 2008). Rather, only a percentage of the offspring demonstrated vision recovery.

Interestingly, the extent of vision recovery seemed to be associated with the geographic distance between cave populations. For instance, the extent to which complementation is recovered in crosses between Molino cavefish and Tinaja cavefish is 39 %, whereas the Pachón cavefish and Tinaja cavefish demonstrate complementation $\sim$8 % of the time. Interestingly, the distance between the Pachon cave and Tinaja cave is relatively short, whereas the Molino cave is much more distant from the Tinaja cave. Thus, geographically distant eyeless cave populations tend to show more complementation compared to geographically close cave populations (Borowsky 2008).

This work implies that geographically close cave populations may share more genetic variants associated with 'eye loss.' This could explain why complementation is less successful when crossing eyeless adults from these two cave populations. Two geographically close cave populations share more genetic changes associated with eye loss, and thus, the percentage of offspring demonstrating an intact, functional visual system are lower. This is due to a failure to complement between causative vision loss genes shared between populations.

Clearly, complementation studies evaluating pigmentation versus eye loss produce very different results. This is explained by the fact that the two pigmentation phenotypes evaluated through complementation analyses were Mendelian (monogenic) phenotypes—albinism and *brown* (Şadoğlu and McKee 1969; Protas et al. 2006; Bilandzija et al. 2013; Gross and Wilkens 2013). In contrast, vision loss in cavefish—irrespective of the population—is highly complex owing to the participation of several genes in vision regression (Protas et al. 2007, 2008). The variable penetrance of vision loss may reflect varying degrees of shared and novel eye loss genes in geographically close caves. Why would these populations harbor a combination of shared and different genes? Shared 'eye loss' genes may have evolved in one local karst environment and become distributed to other nearby caves through early colonization events. Karst geology is highly flexible—communication between caves can be generated and lost through time owing to the dynamic nature of limestone caverns (Mitchell et al. 1977; Bradic et al. 2012). Alternatively, shared 'eye loss' genes may have secondarily appeared in different cave environments through direct migration or gene flow between cave populations. Indeed, substantial gene flow has occurred both in and out of the cave, across the landscape of *Astyanax* cavefish populations (Bradic et al. 2012).

The evolution of albinism and *brown* provides fascinating examples of how Mendelian phenotypes can arise through a convergence across multiple cave populations. Interestingly, the causative alleles leading to albinism (*Oca2*) and brown (*Mc1r*) are shared between two specific populations. For instance, the principal genetic lesion in *Mc1r* is found in the Pachón cave (high frequency) and the Tinaja/Japonés cave populations (low frequency; Gross et al. 2009). This scenario may indicate a role for parallelism in the evolution of *brown*, since Pachón and Tinaja/Japonés are both descended from the older 'El Abra' surface-dwelling lineage.

The evolution of eye loss, however, is quite different. Owing to the multigenic nature of this phenotype, complementation analyses do not result in phenotypes that can be assessed in a binary manner (i.e., presence vs. absence). Rather, recovery of a functional visual system results in a phenotypic continuum, which maps on to the geographic proximity between populations (Borowsky 2008). Thus, eye loss appears to arise across the cavefish landscape through a combination of shared and unique 'eye loss' genes. This would indicate that both parallelism and convergence appear to be at play in vision loss. Parallel features of eye loss could include shared genes across geographically close eyeless populations, irrespective of the route of

these genes between populations. Convergent features of eye loss would be those genes that are different between cave populations. The extent to which convergence and parallelism participate in eye loss across populations requires deeper sampling across many additional cave populations. At present, however, this mosaic regressive phenotype has been largely adopted in different caves independently, suggesting convergence plays a larger relative role in the evolution of eye loss.

7.3 Evolution Through Standing Genetic Variation: Case Studies in Albinism, Eye Loss, and Metabolism

Similar to cavefish, stickleback fish harbor distinct morphotypes capable of interbreeding to produce experimental F_2 pedigrees. Phenotypic variation reflects the two dramatically different environments in which sticklebacks are found, including a marine form and a freshwater form (Colosimo et al. 2005). One of the most dramatic differences between morphs is the presence of ossified bony 'plates' that protect the flanks of marine forms (these plates are reduced or lost in freshwater forms). Quantitative trait locus analyses revealed that a variant form of the gene *EDA* largely mediates bony plate loss in freshwater forms (Colosimo et al. 2005). Following an exhaustive survey, it was discovered that the causative allele was present in extremely low frequency in the wild marine stickleback population (Colosimo et al. 2005).

This work established that rare alleles in a surrogate 'ancestral' population could provide a source of 'standing genetic variation', which is rapidly selected in the derived population. A similar process appears to operate in *Astyanax* cavefish. It should be noted, however, that the evolutionary history of sticklebacks and Mexican cavefish is very different. Sticklebacks evolved very recently—since the recession of the last ice age, ~10,000–20,000 years ago, depending on the freshwater population (Colosimo et al. 2005). In contrast, Mexican cavefish evolved over the last several millions of years, followed by substantial neutral forces such as migration and drift (Gross 2012). Having said this, at least three different phenotypes provide support for the evolution of 'convergent' similarities: a spontaneous origin for albinism, cryptic variation in eye size, and the evolution of metabolic changes.

7.3.1 A Spontaneous Origin of Albinism

Albinism has arisen convergently across the landscape of blind Mexican cavefish (Fig. 7.1). Four populations are known to express the albinotic phenotype in nature: the northern (younger) Molino cave population, and three (older) El Abra populations—Pachón, Japonés, and Yerbaniz (Gross and Tabin 2010). Japonés and

Yerbaniz are believed to be part of the same population given that the surface entrances to the caves are >5 km from one another. While lesions in the same gene (*Oca2*) appear to mediate albinism in all four naturally albino populations, the phylogenetic distance between Molino and the El Abra caves argues strongly that these populations evolved albinism through convergence (Bradic et al. 2012). Interestingly, within the El Abra caves, it is unclear if albinism arose through parallelism or convergence. All three caves arose from the same 'older' stock of surface-dwelling colonizers, and therefore, they share a more recent origin compared to populations drawn from the northern Guatemala and western Micos caves. Similar to the *brown* trait, albinism fails to complement in experimental breedings between the Pachón and Japonés/Yerbaniz caves (Protas et al. 2006). Pachón and Molino harbor destructive changes to the coding sequence of *Oca2*, while Japonés/Yerbaniz have an intact, normal coding sequence of *Oca2* (Protas et al. 2006). Since these populations fail to complement, it is assumed that the Japónes/Yerbaniz populations demonstrate a loss of expression of *Oca2* through regulatory mutations (Protas et al. 2006). It remains to be seen whether the same causative mutation leading to reduced expression is present within the Pachón cave populations. However, if they are present in Pachón and Japonés/Yerbaniz, then the primary lesion responsible for albinism through *Oca2* evolved once and spread between these three caves. Accordingly, the coding sequence mutation may have arisen later in the Pachón cave lineage—explaining its absence from Japonés/Yerbaniz.

A spontaneous occurrence of albinism is more difficult to explain. In 1969, Horst Wilkens discovered the western Micos cave populations (sensu 'Colmena' populations; Gross and Wilkens 2013). A collection of males and females was retrieved from this population and used to create a breeding colony in Wilkens' laboratory in Hamburg, Germany for several decades. After many generations, a small number of albino individuals spontaneously appeared (Gross and Wilkens 2013). Strikingly, these individuals harbored the same coding sequence lesion as Pachón cavefish! Following a series of phenotypic analyses, it was clear that the same haplotype present in Pachón cavefish was present in albino Micos cavefish—and absent from pigmented Micos cavefish (Gross and Wilkens 2013). The authors argued that the most likely scenario was the origin of the mutated allele in the Pachón cave population. This allele was swept into the local surface-dwelling population via migration out of the Pachón cave and then secondarily deposited into the much younger Micos cave population (Gross and Wilkens 2013). This scenario argues for the presence of the loss-of-function *Oca2* allele at very low frequency in the contemporary surface-dwelling population that surrounds the Micos caves. This allele would have been present in the original founders of Wilkens' stock of fish collected in 1969, and random breeding events among these founders enabled the spontaneous appearance of albinism after several generations (Gross and Wilkens 2013).

7.3.2 Hsp90 *Acts as a Capacitor to Morphological Variation in Surface-Dwelling Populations*

Interestingly, in the context of complementation studies that indicate different genes contribute to eye loss in cave-dwelling populations, certain factors shared widely across the El Abra landscape may also play an important role. Rohner et al. (2013) evaluated this question in the context of a chaperone protein encoded by *heat shock protein 90* (*Hsp90*). This work set out to determine whether natural eye size variation may be present in the putatively 'ancestral' surface-dwelling fish (using extant individuals as surrogates). However, the nature of variation in the eye size they sought to characterize was cryptic, in the sense that it is not phenotypically expressed (Rohner et al. 2013). In other systems, it has been shown that a great deal of phenotypic variation can be masked by the action of robust chaperone molecules (e.g., Hsp90) that maintain a narrow range of conformation for a number of phenotypically relevant protein molecules (Rohner et al. 2013).

Their results demonstrated, indeed, surface-dwelling fish harbor a substantial amount of hidden variation that remains masked in these populations. Environmental stressors can act as a trigger to destabilize chaperone molecules, causing variation in eye size to be unmasked following colonization into the extreme cave microenvironment (Rohner et al. 2013). Therefore, the dark and nutrient-poor cave environment results in inhibition of Hsp90 and revealed eye size variation, which could be selected as an adaptive feature among cave-dwelling populations. This work provided support for the existence of an additional mechanism wherein pre-existing phenotypic variation, resulting from alteration of a single chaperone molecule, provides substrate upon which natural selection can act (Rohner et al. 2013). In the context of convergence versus parallelism, this mechanism is more difficult to interpret. Provided that there are no cave-specific changes to the gene *Hsp90*, changes to eye size through Hsp90 inhibition are likely evolving through convergence in geographically distinct cave populations.

This is especially true for the three distinct cave clusters—El Abra, Guatemala, and Micos cave complexes—since these fish were colonized by separate 'older' and 'younger' stock of epigean forms. This study evaluated the phenotypic response to Hsp90 inhibition in extant forms, which are related to the epigean stock that seeded the younger cave populations. While it is impossible to evaluate Hsp90 inhibition with direct ancestors of the 'older' epigean populations, it would be interesting to determine whether the same buffering exists in other surface-dwelling lineages throughout Mexico. Additionally, since the phenotypic (not genotypic) variation is selected following exposure to the stressful subterranean environment (Rohner et al. 2013), it is possible that subtly different forms of this variation are selected in different cave populations. Thus, this fascinating mechanism provides a direct route toward convergent eye loss through standing variation, though not formally through the same gene(s).

7.3.3 *Evolution of Metabolic Changes Through Standing Variation in the Gene Mc4r*

The nutrient-poor cave environment has led to a number of adaptive changes in *Astyanax* cavefish. Among these changes include resistance to starvation, and over-eating when (unlikely) food sources become available in the cave. Using a candidate gene approach, Aspiras et al. (2015) recently identified genetic variants associated with *melanocortin receptor 4* (*Mc4r*) which can mediate increased appetite (Fig. 7.1). Interestingly, one of the affected amino acid residues in the encoded receptor is linked to obesity in humans (Aspiras et al. 2015).

The *Mc4r* gene variant was originally discovered in the Tinaja cave population and found to harbor three non-synonymous mutations in the receptor. Collectively, these hypomorphic mutations led to lower maximal response and basal activity of the receptor compared to the wild-type surface-dwelling form (Aspiras et al. 2015). Further, based on a feeding/appetite assay, animals harboring the derived *Mc4r* variant consumed significantly more worms compared to surface-dwelling fish (Aspiras et al. 2015).

Interestingly, several of the same mutations affecting *Mc4r* were identified throughout the entire Sierra de El Abra landscape. In a comprehensive survey, the investigators noted the presence of the same variants identified from Tinaja cavefish (G145S, V162I, M259I) in numerous additional cave populations, including Pachón Arroyo, Yerbaniz, Piedras, Micos, Japonés, and Sabinos (Aspiras et al. 2015). The derived allelic form was also present in the northern Molino cave population; however, it was not inclusive of all three variants (present in other caves). Strikingly, the derived *Mc4r* variant was also found in surface-dwelling fish, providing a direct route for selection through standing genetic variation present in the 'ancestral' population (Aspiras et al. 2015). This study therefore represents a striking example of evolution of an adaptive phenotype (increased appetite) through parallelism, based on its adoption in cave populations throughout the El Abra, irrespective of whether they were colonized by the 'younger' or 'older' epigean stock (Aspiras et al. 2015). Presumably, the older cave populations adopted the derived variants earlier than the younger cave populations. However, since the same variants are present widely across many surface and cave populations, the altered *Mc4r* variant is most likely evolving in parallel through selection for the same allele.

7.4 Conclusions

Blind Mexican cavefish provide a powerful set of natural 'replicated' experiments for the evolution of the cave phenotypes. These fish are derived from multiple invasions of the subterranean biome, and therefore, they provide a rich context from which we can evaluate how cave phenotypes are adopted in nature. Cave-dwelling

animals evolve a number of stereotypical phenotypes—visual system loss, pigmentation loss, increased appetite, and increased non-visual sensation. Historically, many assumptions underpinned beliefs around the evolutionary origin of *Astyanax* cave populations. Early theories argued that cave populations evolved once and spread throughout the Sierra de El Abra region via subterranean connections. However, recent population genetic studies revealed a very different picture, inclusive of multiple origins of the cave environment via separate 'waves' of epigean stock. In this context, one can evaluate the origin of specific phenotypes and ask whether these evolved in parallel or through convergence. Based on contemporary evidence, most phenotypes appear to evolve through convergence. Mendelian phenotypes, including *brown* and albinism, show mixed patterns. Among closely related populations, these traits are likely evolving in parallel, sometimes through the same allele. However, more distantly related cave populations may evolve the same traits, through the same genes, but owing to distinct genetic lesions and geographic proximity these traits are evolving through convergence. This suggests that certain genes (e.g., *Mc1r*, *Oca2*) may be favorable 'targets' for phenotypic change owing to a vulnerability to mutation in this species. Other traits, including eye loss through inhibition of chaperone proteins, provide an interesting mechanism of 'buffering' cryptic variation in the ancestral form. This buffering is lost in the extreme cave environment, leading to eye variation, which is selected upon in cave-dwelling populations. Selection for metabolic changes through *Mc4r* provides an additional case study for the adoption of derived alleles through parallelism. The variant alleles are present widely across many cave-dwelling populations. What is remarkable, however, is that (unlike *Mc1r* and *Oca2*) these same alleles are also present in the 'ancestral' surface-dwelling stock. This provides a direct route for evolutionary change through selection for adaptive alleles drawn from 'ancestral' surface-dwelling populations. In the coming years, as the draft genome continues to improve, identification and presence of genetic variants shared across multiple cave and surface populations will also continue to improve. This will, in turn, lead to a much clearer understanding of the pace of evolutionary changes across the cavefish landscape, the genetic contributors to cave-associated phenotypes, and the extent to which convergence and parallelism mediate phenotypic change in these remarkable animals.

Acknowledgements The author is grateful to Pierre Pontarotti and Marie-Hélène Rome for the invitation to participate in the 19th Evolutionary Biology Marseilles Meeting and contribute a chapter to this volume. The author also wishes to thank members of the Gross laboratory for helpful comments and suggestions on an earlier draft of this manuscript. This work was supported by grants from the National Science Foundation (DEB-1457630 to JBG), and the National Institutes of Dental and Craniofacial Research (NIH; DE025033 to JBG).

References

Aspiras AC, Rohner N, Martineau B, Borowsky RL, Tabin CJ (2015) Melanocortin 4 receptor mutations contribute to the adaptation of cavefish to nutrient-poor conditions. Proc Natl Acad Sci USA 112:9668–9673

Bilandzija H, Ma L, Parkhurst A, Jeffery WR (2013) A potential benefit of albinism in *Astyanax* cavefish: Downregulation of the *oca2* gene increases tyrosine and catecholamine levels as an alternative to melanin synthesis. PLoS ONE 8:e80823

Borowsky R (2008) Restoring sight in blind cavefish. Curr Biol 18:R23-R24

Bradic M, Beerli P, Garcia-de Leon FJ, Esquivel-Bobadilla S, Borowsky RL (2012) Gene flow and population structure in the Mexican blind cavefish complex (*Astyanax mexicanus*). BMC Evol Biol 12:9

Bradic M, Teotonio H, Borowsky RL (2013) The population genomics of repeated evolution in the blind cavefish *Astyanax mexicanus*. Mol Biol Evol 30:2383–2400

Colosimo PF, Hosemann KE, Balabhadra S, Villarreal G Jr, Dickson M, Grimwood J, Schmutz J, Myers RM, Schluter D, Kingsley DM (2005) Widespread parallel evolution in sticklebacks by repeated fixation of *ectodysplasin* alleles. Science 307:1928–1933

Culver DC (1982) Cave life: Evolution and ecology. Harvard University Press, Cambridge, p 189

Gross JB (2012) The complex origin of *Astyanax* cavefish. BMC Evol Biol 12:105

Gross JB, Tabin CJ (2010) Evolutionary genetics of pigmentation loss. In: Search of the causes of evolution: from field observations to mechanisms. Princeton University Press, Princeton

Gross J, Wilkens H (2013) Albinism in phylogenetically and geographically distinct populations of *Astyanax* cavefish arises through the same loss-of-function *Oca2* allele. Heredity 111:122–130

Gross JB, Borowsky R, Tabin CJ (2009) A novel role for *Mc1r* in the parallel evolution of depigmentation in independent populations of the cavefish *Astyanax mexicanus*. PLoS Genet 5: e1000326

Mitchell RW, Russell WH, Elliott WR (1977) Mexican eyeless characin fishes, genus *Astyanax*: environment, distribution, and evolution. Texas Tech Press, Lubbock, p 89

Moro O, Ideta R, Ifuku O (1999) Characterization of the promoter region of the human *melanocortin-1 receptor* (*MC1R*) gene. Biochem Biophys Res Commun 262:452–460

Ornelas-García CP, Dominguez-Dominguez O, Doadrio I (2008) Evolutionary history of the fish genus *Astyanax* Baird & Girard (1854) (Actinopterygii, Characidae) in Mesoamerica reveals multiple morphological homoplasies. BMC Evol Biol 8:340

Protas ME, Hersey C, Kochanek D, Zhou Y, Wilkens H, Jeffery WR, Zon LI, Borowsky R, Tabin CJ (2006) Genetic analysis of cavefish reveals molecular convergence in the evolution of albinism. Nat Genet 38:107–111

Protas M, Conrad M, Gross JB, Tabin C, Borowsky R (2007) Regressive evolution in the Mexican cave tetra, *Astyanax mexicanus*. Curr Biol 17:452–454

Protas M, Tabansky I, Conrad M, Gross JB, Vidal O, Tabin CJ, Borowsky R (2008) Multi-trait evolution in a cave fish, *Astyanax mexicanus*. Evol Dev 10:196–209

Protas ME, Trontelj P, Patel NH (2011) Genetic basis of eye and pigment loss in the cave crustacean, *Asellus aquaticus*. Proc Natl Acad Sci USA 108:5702–5707

Rohner N, Jarosz DF, Kowalko JE, Yoshizawa M, Jeffery WR, Borowsky RL, Lindquist S, Tabin CJ (2013) Cryptic variation in morphological evolution: HSP90 as a capacitor for loss of eyes in cavefish. Science 342:1372–1375

Şadoğlu P, McKee A (1969) A second gene that affects eye and body color in Mexican blind cave fish. J Hered 60:10–14

Stahl BA, Gross JB (2015) Alterations in *Mc1r* gene expression are associated with regressive pigmentation in *Astyanax* cavefish. Dev Genes Evol 225:367–375

Stahl BA, Gross JB, Speiser DI, Oakley TH, Patel NH, Gould DB, Protas ME (2015) A transcriptomic analysis of cave, surface, and hybrid isopod crustaceans of the species *Asellus aquaticus*. PLoS ONE 10:e0140484

Wilkens H (1971) Genetic interpretation of regressive evolutionary processes: studies on hybrid eyes of two *Astyanax* cave populations (Characidae, Pisces). Evolution 25:530–544

Wilkens H (1988) Evolution and genetics of epigean and cave *Astyanax fasciatus* (Characidae, Pisces): support for the neutral mutation theory. In: Hecht MK, Wallace B (eds) Evolutionary biology. Plenum Publishing Corporation, New York, pp 271–367

Wilkens H, Strecker U (2003) Convergent evolution of the cavefish *Astyanax* (characidae: Teleostei): genetic evidence from reduced eye-size and pigmentation. Biol J Linnean Soc 80:545–554

Chapter 8
Evolutionary Pathways Maintaining Extreme Female-Biased Sexual Size Dimorphism: Convergent Spider Cases Defy Common Patterns

Matjaž Kuntner and Ren-Chung Cheng

Abstract Several animal and plant lineages exhibit pronounced sexual size dimorphism (SSD). Here, we review the evolution of female-biased, extreme SSD (hereafter eSSD; females at least twice male size) in two model spider clades, Nephilidae and Argiopinae. Although these two clades exhibit comparable levels of eSSD, we show that the phenomenon takes different evolutionary pathways. In nephilids, no correlation between male and female size changes is detected while this correlation is maintained in argiopines. In nephilids, sizes in both sexes increase through evolutionary time, but female sizes rise faster, which maintains eSSD. In contrast, argiopines exhibit no directional size change in either sex, and eSSD slowly declines. Model fitting analyses reveal that in nephilids, female size and eSSD adhere to Brownian motion, but male body size evolves toward an optimum between 3.5 and 5.7 mm. In contrast, no directional trends can be detected in argiopines with Brownian motion as the best-fit model. Finally, phylogenetic allometric analyses reveal no relationships between male and female sizes in nephilids, while argiopine size evolution is isometric. The sole agreement between the clades seems to be falsification of both Rensch's rule and its converse. However, to establish pervasive patterns in spider size evolution, studies on other comparable lineages are essential. We point toward candidate clades and pose open questions in eSSD research.

M. Kuntner (✉) · R.-C. Cheng
Evolutionary Zoology Laboratory, Biological Institute ZRC SAZU,
Novi trg 2, 1000 Ljubljana, Slovenia
e-mail: kuntner@gmail.com

R.-C. Cheng
e-mail: bolasargiope@gmail.com

M. Kuntner
National Museum of Natural History, Smithsonian Institution,
Washington, D.C., USA

P. Pontarotti (ed.), *Evolutionary Biology*, DOI 10.1007/978-3-319-41324-2_8

8.1 Introduction

Sexual size dimorphism (SSD) describes a morphological condition where male and female sizes differ significantly within a species. Vertebrate cases with males as the larger sex are well known (e.g., humanoids, elephant seals, or elephants) and readily explained by male–male competition mechanisms (Trivers 1972; Isaac 2005). However, female-biased sexual size dimorphism may account for more dramatic size differences (Blanckenhorn 2005; Fairbairn et al. 2007) and its evolution is also more difficult to interpret. Notable examples of large females and small males include widow spiders, barnacles, anglerfish, queen ants, and marine echiuran worms. In spiders, females may be over 100 times the male's weight (Kuntner et al. 2012), but the record-holding animal is the octopus where females outweigh males by more than 10,000 fold (Norman et al. 2002). Because cases of extreme, female-biased SSD (hereafter eSSD, where females are at least twice male size) are rather rare and taxonomically scattered, the phenomenon has intrigued early evolutionists (Darwin 1871). Nevertheless, eSSD has not been subjected to rigorous modern evolutionary research, and the phenomenon continues to receive mere occasional bursts of interest (reviewed in Fairbairn et al. 2007). We consequently lack solid comparative data that would explain common mechanisms responsible for the repeated origin, convergence, and maintenance of eSSD in several lineages.

Most studies have investigated intraspecific patterns of SSD on selected model species (Fairbairn 2005; Cox and Calsbeek 2009; Blanckenhorn et al. 2011). These works largely corroborate the differential equilibrium model of SSD evolution that invokes opposing selection pressures on the sexes (Blanckenhorn 2005). However, having ignored the phylogenetic basis of SSD, these single-species studies offer only limited explanations of processes that lead to SSD because they look at individual fitness costs and benefits in phylogenetic isolation. To gain a more complete picture of the phenomenon, comparative studies are therefore needed, but these are relatively rare (Hormiga et al. 2000; Teder and Tammaru 2005; Foellmer and Moya-Laraño 2007; Webb and Freckleton 2007; Cheng and Kuntner 2014; Kuntner and Elgar 2014; Teder 2014) or cannot be directly compared. In this review, we revisit the evolution of eSSD in spiders within the context of convergent evolution that looks for common themes in phylogenetically independent lineages.

Orb web spiders (Araneoidea) represent an ideal animal group for comparatively testing evolutionary hypotheses regarding the origin and maintenance of eSSD, as eSSD has evolved at least four, but more likely up to nine times independently in this group of spiders (Hormiga et al. 2000; Kuntner et al. 2015). Yet most araneoid lineages remain sexually size monomorphic. Originally, cases of eSSD were referred to as male dwarfism (Vollrath and Parker 1992), but authors have more recently argued that eSSD is generally better explained through female size increase rather than male size decrease (Coddington et al. 1997; Hormiga et al. 2000). However, although such interpretation, if general, would elegantly explain eSSD in spiders as female gigantism due to selection for increased fecundity (Head 1995; Higgins 2002), analyses at a finer taxonomic scale reveal that the origin and shifts

of eSSD are much more complex (Kuntner and Coddington 2009; Kuntner and Elgar 2014). These studies have linked eSSD to a combination of sexual and natural selection components, but as we show here, interspecific patterns of eSSD commonly show no directional trends (Cheng and Kuntner 2014).

Our recent work demonstrates that studying the species-level lineages provides increased resolution compared with studies at higher taxonomic levels. Our understanding of the evolutionary mechanisms of eSSD in spiders, however, is confined to only two clades of a comparable evolutionary age: the family Nephilidae (Kuntner et al. 2008, 2013; Kuntner and Coddington 2009; Higgins et al. 2011; Kuntner and Elgar 2014) and the araneid subfamily Argiopinae (Cheng and Kuntner 2014, 2015). As we show in this review, these clades show, against all predictions, two strikingly differing patterns which suggest that eSSD in these clades arose and is maintained through different mechanisms.

8.2 Studied Clades

We studied independent evolution of eSSD in two orb-weaving spider clades: Nephilidae and Argiopinae (Fig. 8.1). Both clades have over 50 species, are distributed worldwide, and have been extensively studied (e.g., reviews by Kuntner

Fig. 8.1 Representatives of the two clades (Nephilidae and Argiopinae) investigated for evolution of extreme female-biased sexual size dimorphism (eSSD). The *left image* shows a small orange *Nephila pilipes* male climbing on the body of a large female; *right image* shows a small male *Argiope australis* hanging in the web above a large female. Image copyright M. Kuntner

et al. 2008, 2013; Cheng et al. 2010; Walter and Elgar 2012; Cheng and Kuntner 2014; Schneider et al. 2015). Their SSD is pronounced and varies within the clades (Cheng and Kuntner 2014; Kuntner and Elgar 2014), with most species showing eSSD. Recent research robustly resolved species-level phylogenies for these groups (Kuntner et al. 2013; Cheng and Kuntner 2014) and used them to reconstruct sex-specific trends in evolution of size (Cheng and Kuntner 2014; Kuntner and Elgar 2014). We here briefly recap these patterns by first bringing them to the same scale to make them fully comparable, then adding analyses that were reported in one but not both prior studies.

8.3 Macroevolutionary Patterns

Figure 8.2 plots macroevolutionary patterns of sex-specific size changes and the evolution of eSSD in the two investigated clades with comparable ages, 40–50 Myr (Kuntner et al. 2013; Cheng and Kuntner 2014). The upper graphs show female and the middle ones show male body size changes reconstructed across the phylogenies of both lineages, with the *x* axis scaled to cladogenetic events (Cheng and Kuntner 2014; Kuntner and Elgar 2014). The bottom graphs show the reconstructed changes in eSSD, plotted as sexual size dimorphism index (SDI = [female body length/male body length] − 1) (Lovich and Gibbons 1992; Cheng and Kuntner 2014). In nephilids, both female (Linear regression, $P = 0.003$) and male sizes (Linear regression, $P = 0.001$) rise significantly through evolutionary time. However, the female slope rises more steeply than the male slope (Linear regression, $P = 0.015$; Kuntner and Elgar 2014), and these combined maintain eSSD as shown by the SDI slope in Fig. 8.2 that stagnates (Linear regression, $P = 0.346$). The nephilid pattern has been labeled as sexually dimorphic gigantism (Kuntner and Elgar 2014), meaning that both sexes increase in size, but the females more so than the males. On the other hand, Fig. 8.2 shows that the argiopine sex-specific sizes show no trends across cladogenetic events (Cheng and Kuntner 2014): Linear regressions, $P_{\text{female size}} = 0.511$; $P_{\text{male size}} = 0.280$. Therefore, size evolution in argiopines is nondirectional (Cheng and Kuntner 2014). However, the new pattern reported here is that SDI in this clade decreases through evolutionary time (Linear regression, $P = 0.019$).

8.4 Correlation Between Male and Female Size Evolution

Phylogenetically independent contrasts (PIC) analyses of sex-specific size data also establish strikingly contrasting patterns in the two clades. As reported in prior studies on nephilids (Kuntner and Coddington 2009; Kuntner and Elgar 2014), female and male sizes in this clade show no phylogenetic correlation ($r^2 = 0.05$, $t = 1.24$, $F_{1,27} = 1.5$, 2-tailed $P = 0.23$). This independence between male and female size evolution implies a broken genetic linkage between the sexes and suggests that the

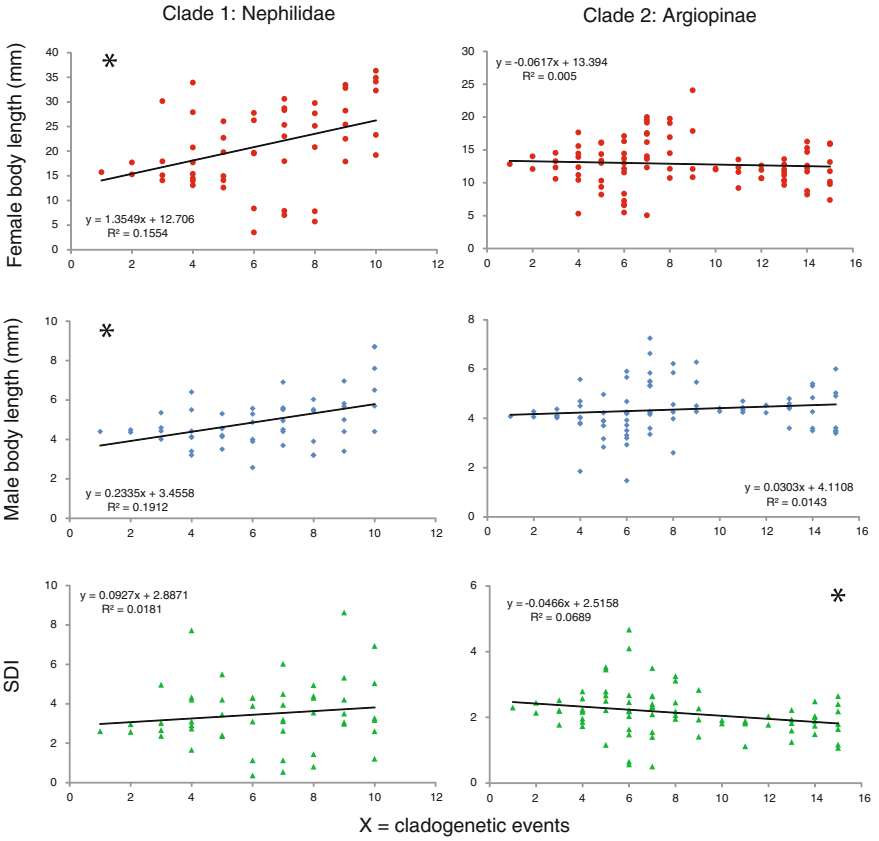


Fig. 8.2 Macroevolutionary patterns of sex-specific size changes and the evolution of eSSD in the two investigated clades. SDI = sexual dimorphism index. Time on X corresponds to cladogenetic events from the phylogenies. The original analyses (Cheng and Kuntner 2014; Kuntner and Elgar 2014) were here brought to the same scale to be comparable. *Asterisks* mark significance

sexes can independently respond to different selection regimes (see below). In contrast, argiopines show correlated phylogenetic changes in male and female sizes ($r^2 = 0.36$, $t = 4.90$, $F_{1,38} = 24.00$, 2-tailed $P = 0.008$) (Cheng and Kuntner 2014). In argiopines, male size is linked to female size and vice versa so that selection on the size of any sex results in size changes in the other (Cheng and Kuntner 2015).

8.5 Detecting Evolutionary Signal

Our prior work on argiopines performed model fitting to sex-specific size data and showed no evolutionary trends for either sex and also no evolutionary trend for SSD (Cheng and Kuntner 2014). Although we would expect models that reveal

Table 8.1 Evolutionary model fitting for nephilid and argiopine spider size and eSSD

	Nephilidae			Argiopinae		
Model	BM	Trend	OU	BM	Trend	OU
Female body length						
lnL	-90.627	-90.588	-90.594	-141.084	-139.505	-139.349
AIC	185.254	187.176	187.187	286.168	285.011	284.698
P value		0.780	0.797		0.076	0.062
Male body length						
lnL	-54.743	-52.752	-49.509	-67.333	-66.828	-67.138
AIC	113.486	111.505	105.017	138.665	139.657	140.276
P value		0.046	0.001		0.315	0.532
SDI						
lnL	-56.544	-55.364	-54.744	-53.808	-53.347	-53.540
AIC	117.088	116.727	115.489	111.617	112.694	113.080
P value		0.124	0.058		0.337	0.464

Nephilid data were analyzed after the methodology reported in the argiopine size evolution paper (Cheng and Kuntner 2014). Abbreviations: *SDI* sexual dimorphism index; *BM* Brownian motion model; *trend* Brownian motion model with a directional trend; *OU* single-optimum Ornstein–Uhlenbeck model. *P* values are from the likelihood ratio test as compared with BM. *Shaded fields* indicate the best-fit model for each comparison

selection for size (either optimum size or directional size), those analyses could not reject the Brownian motion model for size and SSD (Table 8.1). Here, we also performed model fitting to the nephilid size and SSD data following the methodology from the argiopine study (Cheng and Kuntner 2014). We tested the fit of Brownian motion (BM), Brownian motion with a directional trend (Trend), and the single-optimum Ornstein–Uhlenbeck (OU) on the nephilid tree (Kuntner et al. 2013; Kuntner and Elgar 2014) pruned for those taxa for which we had no data for one sex. We first estimated the log-likelihood value of each model, and then selected the best-fit model using the likelihood ratio test. For model fitting methodology and model assumptions, see Cheng and Kuntner (2014) and the literature cited there.

The results cannot reject the BM model for nephilid female body size changes, while the OU model best explains nephilid male size changes (Table 8.1). This suggests that nephilid male evolution is directed toward a body size optimum. Although our analyses cannot unequivocally establish this optimum, we find it logical that it should be on the trend line of the graph in Fig. 8.2, i.e., between 3.5 and 5.7 mm. While for SDI (and thus eSSD) we cannot reject BM, the OU model shows only a marginally nonsignificant result (Table 8.1).

8.6 Allometric Patterns

Allometric analyses of male (on *y* axis) and female sizes (on *x* axis) provide a different aspect of understanding SSD evolution (Fairbairn 1997). An isometric slope would imply that size changes of males and females are comparable, and that SSD is maintained. In organisms with a male-biased SSD, a positive allometric slope ($\beta > 1$) is usually detected, meaning that SSD increases as male body size increases (Rensch 1950). While this pattern, known as Rensch's rule, is common in birds and mammals, its opposite is predicted to hold in animals with a female-biased SSD, including spiders (Abouheif and Fairbairn 1997; Fairbairn 1997). The converse of Rensch's rule predicts a negative allometric slope ($\beta < 1$) meaning that SSD increases as female body size increases (Abouheif and Fairbairn 1997; Foellmer and Moya-Laraño 2007; Cheng and Kuntner 2014). Note that by convention allometric analyses of sex-specific sizes plot male size on y and female size on *x* axes (Fairbairn 1997).

In spiders, the literature therefore expects to find the converse pattern of Rensch's rule (Abouheif and Fairbairn 1997). However, against these predictions, a comparative study at higher phylogenetic levels in spiders failed to find any departure from isometry (Foellmer and Moya-Laraño 2007). Likewise, our study established that in argiopines size patterns do not depart from isometry (Cheng and Kuntner 2014). These same patterns were detected when analyzing tip data and phylogenetically controlled (PIC) data shown in Fig. 8.3. Here, we reanalyzed the

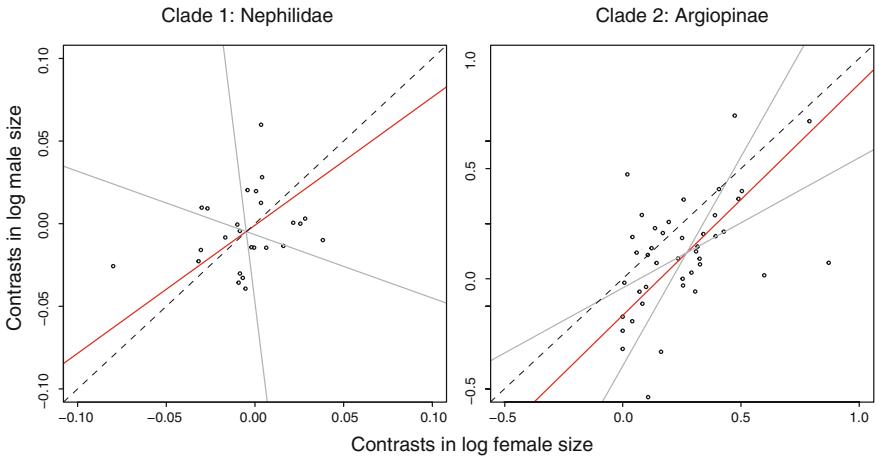


Fig. 8.3 The results from allometric analyses of size data using major axis regression on phylogenetically independent contrast data. The nephilid analysis is new and follows the methodology described in the argiopine study (Cheng and Kuntner 2014). Red lines are allometric slopes; gray lines are 95 % confidence intervals of the slope; and dashed lines indicate a slope equal to one. The nephilid allometric slope does not significantly depart from zero, implying that male and female sizes evolve independently. The allometric slope for argiopines, on the other hand, is not significantly different from one, implying isometric evolution

nephilid size data in a comparable way using major axis regression following Cheng and Kuntner (2014). When analyzing raw size data from the tips of the phylogeny, we detected a negative allometry resembling the converse pattern of Rensch's rule (MA slope = 0.365 (0.147–0.619), $P = 0.009$). However, when analyzing PIC data, we detected no relationship between male and female size (Fig. 8.3; MA slope = 0.775 (−0.384 to −8.695), $P = 0.139$). Although this result defies expected allometric patterns from the literature, it is consistent with the above reported lack of any correlation between female and male size evolution in nephilids.

Our results in the two clades, combined with those from Foellmer and Moya-Laraño (2007), provide no comparative evidence for Rensch's rule or its converse in spiders.

8.7 Evolutionary Pressures

The correlated isometric evolution in size between the sexes in argiopines in contrast to nephilids suggests that the species from each clade either respond differently to the same sex-specific selection pressures, or that these are different in each clade. In argiopines, any sex-specific tendency in size changes must also affect the changes in the opposite sex. In nephilids, on the other hand, one sex is free to directly respond to selection pressures for size changes without affecting the opposing sex. This key difference has direct implications for establishing which selection pressures drive size changes in each sex and consequently eSSD.

Female gigantism in spiders is usually attributed to selection for increased female fecundity (i.e., fecundity selection; Head 1995; Kuntner and Elgar 2014). However, our model fitting results detect random fluctuations in female size rather than a trend toward large size as would be expected under the fecundity model. Nevertheless, female size is on the rise in macroevolutionary time in nephilids and this contributes to the eSSD observed in that clade (Fig. 8.2). This is not the case in argiopines, where eSSD has been suggested to be phylogenetically undetectable, perhaps a consequence of ecological factors operating at the population level (Cheng and Kuntner 2014). Female fecundity has also been questioned as the driver of sexual *shape* dimorphism in argiopines (Cheng and Kuntner 2015).

Several hypotheses explain the selection pressures that prevent male sizes from following those of females. In size dimorphic spiders, small male sizes have been proposed to be advantageous in mate searching, either to reduce male mortality during risky search for sedentary females (Vollrath and Parker 1992; Walker and Rypstra 2003; Kasumovic et al. 2007), or because maturation at small sizes may give males an advantage in scramble competition for virgin females (Danielson-François et al. 2012; Neumann and Schneider 2015). Small male size may also be advantageous during episodes of sexual conflict, perhaps because smaller males more easily evade sexual cannibalism (Elgar 1991). Gravity selection, likewise, may operate strongly in those species that need to climb and selects

for males of small or optimal size (Moya-Laraño et al. 2002, 2009; Corcobado et al. 2010; but see Brandt and Andrade 2007). On the other hand, male–male contests favor larger male size and across the nephilid taxa, this may be the most pervasive selection trend (Kuntner and Elgar 2014).

A combination of these evolutionary pressures on males may operate in nephilid and argiopine spiders. We hypothesize that selection for male size in argiopines affects the female size as well, as the comparative results suggest that they coevolve. Thus, male–male competition that selects for larger males pushes size in both sexes into the same direction as fecundity that selects for larger females. In contrast, mate searching-related mechanisms work in the opposite direction, again affecting both sexes. In argiopines, the net result is an evolutionary decrease in eSSD. In nephilids, on the other hand, selection for small male size may not affect the females at all. Here, we hypothesize that nephilid females freely respond to selection for large size whereas males experience trade-offs that result in selection for optimal size. Such sexually decoupled size evolution is responsible for maintenance of eSSD in this clade.

8.8 Outlook

To establish pervasive patterns in spider size evolution, studies on other comparable lineages are essential. This section briefly points toward promising candidate clades that (i) show comparable levels in eSSD to nephilids and argiopines; (ii) and whose eSSD is either convergent to the one in nephilids and argiopines or alternatively shows a homologous origin to one of these clades, but with subsequent modification (Hormiga et al. 2000; Kuntner et al. 2015).

Among araneid candidates that exhibit nonhomologous origin of eSSD to nephilids and argiopines are bark spiders (genus *Caerostris*), and the groups Gasteracanthinae (*Gasteracantha, Micrathena*, and other genera) and Mastophorinae; among theridiids are widows (genus *Latrodectus*), and the genera *Tidarren* and *Echinotheridion*; a lone size dimorphic tetragnathid clade is the genus *Opadometa*. There are further cases of eSSD in crab spiders (Thomisidae) and of more moderate SSD in raft spiders (genus *Dolomedes*, Pisauridae). Studies repeatedly suggest that levels of SSD in these lineages may relate to unusual traits in their sexual biology, physiology, and web ecology (Michalik et al. 2005, 2010; Agnarsson et al. 2010; Kuntner and Agnarsson 2010; Gregorič et al. 2011a, b; Schwartz et al. 2013; Kuntner et al. 2015; Kralj-Fišer et al. 2016).

In addition, we find the following two araneid clades particularly important as they seem to share a phylogenetically deep origin of SSD with *Argiope* (Cheng and Kuntner 2014): tent spiders—genus *Cyrtophora* whose SSD varies from moderate to extreme, and scorpion-tailed spiders—genus *Arachnura* that may represent the most extreme cases of SSD in spiders.

Animal lineages with extremely female size-biased species are rather exceptional, but we argue that these phylogenetic outliers are important. However, in addition to studying the highly eSSD clades such as those proposed above, comparative research should also involve their monomorphic relatives, so phylogenetic insight is essential.

We also call for detailed genetic studies of eSSD in spiders. If the same sets of genes determine size in both sexes, then it is logical to predict more or less equal sizes in males and females. The known eSSD cases thus require a genetic explanation, but such work has not been done. It would be revealing to begin to understand the genetic differences between a lineage with a demonstrated correlated size evolution (e.g., argiopines) and another with a broken correlational patterns (e.g., nephilids).

In extremely sexually size dimorphic clades, one would expect to detect sex-specific size optima due to a combination of natural and sexual selection pressures. Fecundity selection is expected to strongly favor large female size in egg-laying animals, while pressures that relate to mate searching are predicted to maintain small male sizes (Blanckenhorn 2005). The combined fecundity and gravity hypotheses predict a trend in increased female size and an optimal (smallish) size to be maintained in males. This pattern is here found in nephilids, but not in argiopines. Experimental research should investigate the validity of the optimal nephilid male size, here hypothesized between 3.5 and 5.7 mm. Does it relate to gravity, sexual conflict, mate searching, differential mortality, other causes, or a combination of these?

8.9 Summary

Although nephilids and argiopines exhibit comparable levels of eSSD (Fig. 8.1), we demonstrate clear differences in evolutionary patterns that result in eSSD and maintain it. In nephilids, the correlation in sex-specific size changes is broken while this correlation is maintained in argiopines. In nephilids, the size in both sexes significantly increases in evolutionary time, but females grow faster and this difference maintains eSSD. In contrast, argiopines exhibit no direction of size change in either sex, and eSSD slowly declines. Model fitting analyses reveal that in nephilids, female size and eSSD do not depart from Brownian motion, but male size tends toward an optimum. In contrast, no directional trends can be detected in argiopines where Brownian motion best fits the data. Finally, phylogenetic allometric analyses reveal no relationships between male and female sizes in nephilids, but in contrast, argiopine size evolution is isometric. The sole similarity between the clades seems to be falsification of both Rensch's rule and its converse.

Acknowledgments We thank Pierre Pontarotti for inviting us to contribute to this volume, an anonymous reviewer for helpful suggestions on how to improve our presentation, as well as Ingi Agnarsson, Matjaž Gregorič, Simona Kralj-Fišer, Klemen Čandek, Shakira Quiñones-Lebrón, and Javad Malekhosseini for valuable feedback on early drafts.

References

Abouheif E, Fairbairn DJ (1997) A comparative analysis of allometry for sexual size dimorphism: assessing Rensch's rule. Am Nat 149:540–562. doi:10.1086/286004

Agnarsson I, Kuntner M, Blackledge TA (2010) Bioprospecting finds the toughest biological material: extraordinary silk from a giant riverine orb spider. PLoS ONE 5:e11234. doi: 10.1371/journal.pone.0011234

Blanckenhorn W (2005) Behavioral causes and consequences of sexual size dimorphism. Ethology 111:977–1016. doi:10.1111/j.1439-0310.2005.01147.x

Blanckenhorn WU, Stillwell RC, Young KA, Fox CW, Ashton KG (2011) When Rensch meets Bergmann: does sexual size dimorphism change systematically with latitude? Evolution 60:2004–2011. doi:10.1111/j.0014-3820.2006.tb01838.x

Brandt Y, Andrade MCB (2007) Testing the gravity hypothesis of sexual size dimorphism: are small males faster climbers? Funct Ecol 21:379–385. doi:10.1111/j.1365-2435.2007.01243.x

Cheng RC, Kuntner M (2014) Phylogeny suggests non-directional and isometric evolution of sexual size dimorphism in argiopine spiders. Evolution 68:1–31. doi:10.1111/evo.12504

Cheng RC, Kuntner M (2015) Disentangling the size and shape components of sexual dimorphism. Evol Biol 42:223–234. doi:10.1007/s11692-015-9313-z

Cheng RC, Yang EC, Lin CP, Herberstein ME, Tso IM (2010) Insect form vision as one potential shaping force of spider web decoration design. J Exp Biol 213:759–768. doi:10.1242/jeb.037291

Coddington JA, Hormiga G, Scharff N (1997) Giant female or dwarf male spiders? Nature 385:687–688

Corcobado G, Rodríguez-Gironés MA, De Mas E, Moya-Laraño J (2010) Introducing the refined gravity hypothesis of extreme sexual size dimorphism. BMC Evol Biol 10:236. doi:10.1186/1471-2148-10-236

Cox RM, Calsbeek R (2009) Sex-specific selection and intraspecific variation in sexual size dimorphism. Evolution 64:798–809. doi:10.1111/j.1558-5646.2009.00851.x

Danielson-François A, Hou C, Cole N, Tso IM (2012) Scramble competition for moulting females as a driving force for extreme male dwarfism in spiders. Anim Behav 84:937–945. doi:10.1016/j.anbehav.2012.07.018

Darwin C (1871) The descent of man, and selection in relation to sex. John Murray, London

Elgar MA (1991) Sexual cannibalism, size dimorphism, and courtship in orb-weaving spiders (Araneidae). Evolution 45:444–448

Fairbairn DJ (1997) Allometry for sexual size dimorphism: patterns and process in the coevolution of body size in males and females. Annu Rev Ecol Syst 28:659–687. doi:10.1146/annurev.ecolsys.28.1.659

Fairbairn DJ (2005) Allometry for sexual size dimorphism: testing two hypotheses for Rensch's rule in the water strider *Aquarius remigis*. Am Nat 166(Suppl):S69–S84. doi:10.1086/444600

Fairbairn DJ, Blanckenhorn WU, Székely T (2007) Sex, size, and gender roles: evolutionary studies of sexual size dimorphism. Oxford University Press, Oxford

Foellmer MW, Moya-Laraño J (2007) Sexual size dimorphism in spiders: patterns and processes. In: Fairbairn DJ, Blanckenhorn WU, Székely TS (eds) Sex, size, and gender roles: evolutionary studies of sexual size dimorphism. Oxford University Press, Oxford, pp 71–82

Gregorič M, Agnarsson I, Blackledge TA, Kuntner M (2011a) How did the spider cross the river? Behavioral adaptations for river-bridging webs in *Caerostris darwini* (Araneae: Araneidae). PLoS ONE 6:e26847. doi:10.1371/journal.pone.0026847

Gregorič M, Agnarsson I, Blackledge TA, Kuntner M (2011b) Darwin's bark spider: giant prey in giant orb webs (*Caerostris darwini*, Araneae: Araneidae)? J Arachnol 39:287–295. doi:10.1636/CB10-95.1

Head G (1995) Selection on fecundity and variation in the degree of sexual size dimorphism among spider species (Class Araneae). Evolution 49:776–781. doi:10.2307/2410330

Higgins L (2002) Female gigantism in a New Guinea population of the spider *Nephila maculata*. Oikos 99:377–385. doi:10.1034/j.1600-0706.2002.990220.x

Higgins L, Coddington J, Goodnight C, Kuntner M (2011) Testing ecological and developmental hypotheses of mean and variation in adult size in nephilid orb-weaving spiders. Evol Ecol 25:1289–1306. doi:10.1007/s10682-011-9475-9

Hormiga G, Scharff N, Coddington JA (2000) The phylogenetic basis of sexual size dimorphism in orb-weaving spiders (Araneae, Orbiculariae). Syst Biol 49:435–62. doi:10.1080/10635159950127330

Isaac JL (2005) Potential causes and life-history consequences of sexual size dimorphism in mammals. Mamm Rev 35:101–115

Kasumovic MM, Bruce MJ, Herberstein ME, Andrade MCB (2007) Risky mate search and mate preference in the golden orb-web spider (*Nephila plumipes*). Behav Ecol 18:189–195. doi:10.1093/beheco/arl072

Kralj-Fišer S, Čandek K, Lokovšek T, Čelik T, Cheng RC, Elgar MA, Kuntner M (2016) Mate choice and sexual size dimorphism, not personality, explain female aggression and sexual cannibalism in raft spiders. Anim Behav 111:49–55. doi:10.1016/j.anbehav.2015.10.013

Kuntner M, Agnarsson I (2010) Web gigantism in Darwin's bark spider, a new species from Madagascar (Araneidae: *Caerostris*). J Arachnol 38:346–356. doi:10.1636/B09-113.1

Kuntner M, Coddington JA (2009) Discovery of the largest orbweaving spider species: the evolution of gigantism in *Nephila*. PLoS ONE 4:2–6. doi:10.1371/journal.pone.0007516

Kuntner M, Elgar MA (2014) Evolution and maintenance of sexual size dimorphism: aligning phylogenetic and experimental evidence. Front Ecol Evol 2:1–8. doi:10.3389/fevo.2014.00026

Kuntner M, Coddington JA, Hormiga G (2008) Phylogeny of extant nephilid orb-weaving spiders (Araneae, Nephilidae): testing morphological and ethological homologies. Cladistics 24:147–217

Kuntner M, Zhang S, Gregorič M, Li D (2012) *Nephila* female gigantism attained through post-maturity molting. J Arachnol 40:345–347. doi:10.1636/B12-03.1

Kuntner M, Arnedo MA, Trontelj P, Lokovšek T, Agnarsson I (2013) A molecular phylogeny of nephilid spiders: Evolutionary history of a model lineage. Mol Phylogenet Evol 69:961–979. doi:10.1016/j.ympev.2013.06.008

Kuntner M, Agnarsson I, Li D (2015) The eunuch phenomenon: adaptive evolution of genital emasculation in sexually dimorphic spiders. Biol Rev 90:279–296. doi:10.1111/brv.12109

Lovich JE, Gibbons JW (1992) A review of techniques for quantifying sexual size dimorphism. Growth Dev Aging 56:269–281

Michalik P, Knoflach B, Thaler K, Alberti G (2005) The spermatozoa of the one-palped spider *Tidarren argo* (Araneae, Theridiidae). J Arachnol 33:562–568. doi:10.1636/04-65.1

Michalik P, Knoflach B, Thaler K, Alberti G (2010) Live for the moment: adaptations in the male genital system of a sexually cannibalistic spider (Theridiidae, Araneae). Tissue Cell 42:32–36. doi:10.1016/j.tice.2009.06.004

Moya-Laraño J, Halaj J, Wise DH (2002) Climbing to reach females: Romeo should be small. Evolution 56:420–425. doi:10.1111/j.0014-3820.2002.tb01351.x

Moya-Laraño J, Vinković D, Allard CM, Foellmer MW (2009) Optimal climbing speed explains the evolution of extreme sexual size dimorphism in spiders. J Evol Biol 22:954–63. doi:10.1111/j.1420-9101.2009.01707.x

Neumann R, Schneider JM (2015) Differential investment and size-related mating strategies facilitate extreme size variation in contesting male spiders. Anim Behav 101:107–115. doi:10.1016/j.anbehav.2014.12.027

Norman MD, Paul D, Finn J, Tregenza T (2002) First encounter with a live male blanket octopus: The world's most sexually size-dimorphic large animal. New Zeal J Mar Freshw Res 36:733–736. doi:10.1080/00288330.2002.9517126

Rensch B (1950) Die Abhängigkeit der relativen Sexualdifferenz von der Körpergrösse. Bonner Zool Beiträge 1:58–69

Schneider J, Uhl G, Herberstein ME (2015) Cryptic female choice within the genus *Argiope*: a comparative approach. In: Cryptic Female Choice in Arthropods. Springer, Berlin, pp 55–77

Schwartz SK, Wagner WE, Hebets EA (2013) Spontaneous male death and monogyny in the dark fishing spider. Biol Lett 9:20130113. doi:10.1098/rsbl.2013.0113

Teder T (2014) Sexual size dimorphism requires a corresponding sex difference in development time: a meta-analysis in insects. Funct Ecol 28:479–486. doi:10.1111/1365-2435.12172

Teder T, Tammaru T (2005) Sexual size dimorphism within species increases with body size in insects. Oikos 108:321–334. doi:10.1111/j.0030-1299.2005.13609.x

Trivers RL (1972) Parental investment and sexual selection. Sex Sel Descent Man, 1871–1971 136–179. doi:10004055386

Vollrath F, Parker GA (1992) Sexual dimorphism and distorted sex ratios in spiders. Nature 355:156–159. doi:10.1038/355242a0

Walker SE, Rypstra AL (2003) Sexual dimorphism and the differential mortality model: is behaviour related to survival? Biol J Linn Soc 78:97–103. doi:10.1046/j.1095-8312.2003.00134.x

Walter A, Elgar MA (2012) The evolution of novel animal signals: silk decorations as a model system. Biol Rev 87:686–700. doi:10.1111/j.1469-185X.2012.00219.x

Webb TJ, Freckleton RP (2007) Only half right: species with female-biased sexual size dimorphism consistently break Rensch's rule. PLoS ONE. doi:10.1371/journal.pone.0000897

Part II
Evolution of Complex Traits

Chapter 9
Evolution of the BCL-2-Regulated Apoptotic Pathway

Abdel Aouacheria, Emilie Le Goff, Nelly Godefroy and Stephen Baghdiguian

Abstract The mitochondrion descends from a bacterium that, about two billion years ago, became endosymbiotic. This organelle represents a Pandora's box whose opening triggers cytochrome-c release and apoptosis of cells from multicellular animals, which evolved much later, about six hundred million years ago. BCL-2 proteins, which are critical apoptosis regulators, were recruited at a certain time point in evolution to either lock or unlock this mitochondrial Pandora's box. Hence, particularly intriguing is the issue of when and how the "BCL-2 proteins–mitochondria–apoptosis" triptych emerged. This chapter explains what it takes from an evolutionary perspective to evolve a BCL-2-regulated apoptotic pathway, by focusing on the events occurring upstream of mitochondria.

9.1 Introduction

It is in the form of cells that life has continued over generations for billions of years. Most of the time, these building blocks of life are defined as self-replicating elements, overlooking the fact that cells endowed with the ability to self-destruct were described in all branches of the tree of life (Bozhkov and Lam 2011; Dwyer and Winkler 2013; Kerr et al. 1972; Madeo et al. 1997). In multicellular animals (metazoans), organismal success and complexity are built upon the silent destruction and rapid removal of cells through a genetically encoded cell death process called apoptosis. This form of active (or programmed) cell death functions to sculpt shapes, optimize functions and eliminate damaged, superfluous, or harmful cells from the body, thus playing crucial roles in animal development and homeostasis.

A. Aouacheria (✉) · E. Le Goff · N. Godefroy · S. Baghdiguian
ISEM - Institut Des Sciences de L'Evolution de Montpellier, UMR 5554, Université de Montpellier, CNRS, IRD, CIRAD, EPHE, Place Eugène Bataillon, 34095 Montpellier, France (publication number: ISEM 2016-113)
e-mail: abdelouahab.aouacheria@umontpellier.fr; aouacheria.abdel@gmail.com

P. Pontarotti (ed.), *Evolutionary Biology*, DOI 10.1007/978-3-319-41324-2_9

Adding to its importance as a physiological phenomenon, apoptosis dysregulation is involved in a wide range of diseases such as cancer (Czabotar et al. 2014; Elmore 2007). Studies on vertebrate cells have revealed the existence of two major apoptotic pathways: the extrinsic pathway initiated by the ligation of death receptors by extracellular ligands at the surface of target cells and the intrinsic (mitochondrial or BCL-2-regulated) pathway, which can be stimulated by a plethora of signals (e.g., DNA damage, endoplasmic reticulum stress, hypoxia, growth factor deprivation, and developmental cues) (Czabotar et al. 2014; Tait and Green 2013).

The BCL-2-regulated apoptotic pathway is initiated through transcriptional and/or post-transcriptional activation of so-called BH3-only proteins, which form a disparate group of proteins traditionally considered as sensors of cellular stress and damage (Doerflinger et al. 2015; Shamas-Din et al. 2011). In response to distinct upstream signaling events, some of these death effectors (i.e., BIM, PUMA, tBID) serve as ligands to activate the pro-apoptotic BCL-2 family members BAX and BAK through direct interaction, while all of them can activate BAX/BAK indirectly by binding to and inhibiting the prosurvival BCL-2 homologous proteins. Once activated through a complex multi-step process, BAX and BAK are thought to homo-oligomerize and form (or participate to) pores in the mitochondrial outer membrane (Tait and Green 2013; Volkmann et al. 2014; Westphal et al. 2014). These oligomeric pores cause the release of mitochondrial intermembrane space proteins, including cytochrome-c (cyt-c), in the cytosol (in a process termed MOMP, for mitochondrial outer membrane permeabilization). Leaked cyt-c then triggers the activation of a family of death proteases called caspases through a well-defined post-mitochondrial pathway (which will not be reviewed here). Prosurvival BCL-2 proteins can inhibit BAX-BAK activity through one or more possible mechanisms: sequestration of the "direct activator" BH3-only proteins (Llambi et al. 2011) or local inhibition of BAX (and BAK) at the mitochondrial outer membrane level (via inhibitory complex formation, oligomer disassembly, and/or retrotranslocation to the cytosol) (Billen et al. 2008a; Edlich et al. 2011; Subburaj et al. 2015). The "sensitizer" BH3-only proteins can neutralize the prosurvival BCL-2 proteins through direct binding, thus releasing the direct activators to promote BAX-BAK activation and apoptosis. An important concept that emerges from this mechanistic description is that the BCL-2-regulated pathway appears to be organized in a hierarchical manner, from cellular sentinels (BH3-only proteins) to the BCL-2/BAX apoptotic switch controlling cytosolic release of mitochondrial apoptogenic factors (Fig. 9.1). The next sections will address evolutionary perspectives on all three categories of constituents, with a particular emphasis on BCL-2 homologous proteins [aka BCL-2 family members, see https://bcl2db.ibcp.fr/BCL2DB/BCL2DBNomenclature for an explanation of nomenclature (Aouacheria 2014; Rech de Laval et al. 2014)].

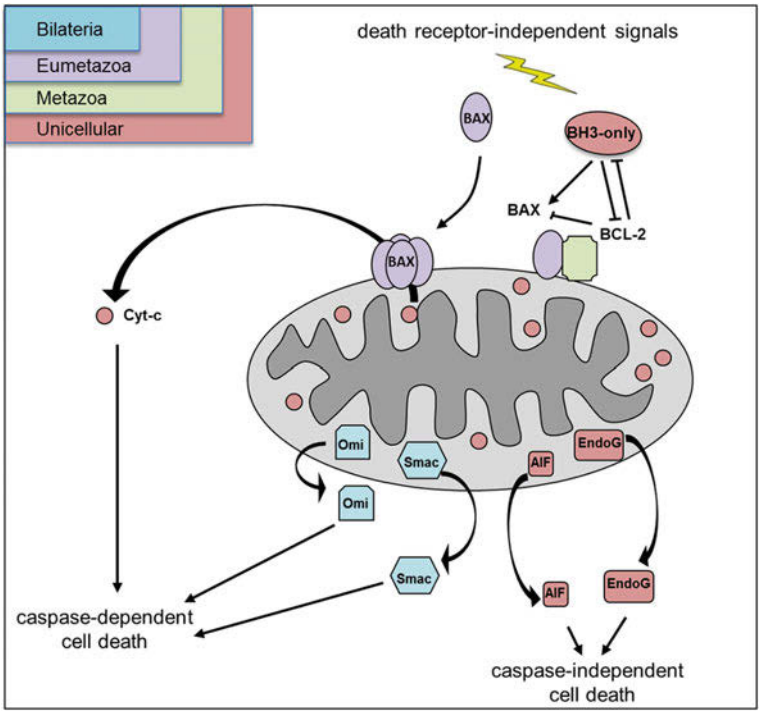


Fig. 9.1 Simplified representation of the mitochondrial apoptotic pathway. Mitochondrial outer membrane permeabilization (MOMP) constitutes the pivotal event in the mitochondrial or BCL-2-regulated intrinsic death pathway and results in the release of cytochrome-c (cyt-c) and other mitochondrial apoptogenic factors from the mitochondrial intermembrane space to the cytosol. Once in the cytoplasm, cyt-c activates a family of death proteases called caspases that leads to the cleavage of a myriad of cellular substrates, causing cell demise. This pathway is initiated by activation of BH3-only proteins which serve as ligands to activate proapoptotic BCL-2 family members (e.g., BAX) through direct interaction or by binding to and inhibiting prosurvival BCL-2 homologous proteins (like BCL-2). Once activated, BAX homo-oligomerizes and forms pores in the mitochondrial outer membrane which cause the release of apoptogenic factors. Among these apoptogenic proteins, cytochrome-c, Omi/HtrA2, and SMAC/Diablo promote caspase-dependent cell death, whereas AIF and endoG induce caspase-independent cell death. Gene products are colored by their phyletic distribution (*inset*, see text for details)

9.2 Mitochondrial Intermembrane Space Proteins

The output of MOMP corresponds to the cytosolic release of mitochondrial apoptogenic proteins normally sequestered within the intermembrane space. The following five death-promoting factors have received significant characterization: cyt-c, apoptosis-inducing factor (AIF), second mitochondrial activator of caspases (Smac)/Diablo, Omi/HtrA2 and endonuclease G (endoG). During apoptosis, cyt-c directly induces caspase activation whereas Smac/Diablo and Omi/HtrA2 neutralize

the inhibition of caspase activation (Lorenzo and Susin 2004; Saelens et al. 2004). AIF and endoG translocates to the nucleus to trigger caspase-independent DNA fragmentation (Arnoult et al. 2003; Cregan et al. 2004). All of these mitochondrial factors are encoded in the nucleus. Cyt-c and AIF have the widest phylogenetic distribution as they are found both in prokaryotes (Archaea, bacteria) and eukaryotes including protists, plants, fungi, and animals. Omi/HtrA2 and endoG also display a wide phylogenetic pattern and are present in all kingdoms of life except Archaea. Based on this, it seems reasonable to infer that these four mitochondrial apoptogenic proteins represent endosymbiotic acquisitions from the mitochondrial ancestor. In contrast, Smac/Diablo orthologues are only present in vertebrate species, suggesting a late phylogenetic origin. Interestingly, most of these mitochondrial intermembrane space proteins can act as "pencils–erasers": cyt-c, for instance, has a vital daily job in respiration (as an essential electron carrier) and becomes cytotoxic only when its gets to the cytosol (Garrido and Kroemer 2004). AIF was also suggested to exert vital normal functions (possibly pertaining to its oxidoreductase activity) (Porter and Urbano 2006; Vahsen et al. 2004; Sorrentino et al. 2015). EndoG may be involved in DNA recombination and repair in addition to proliferation (Buttner et al. 2007; Huang et al. 2006). These examples illustrate an important but often neglected aspect of many apoptotic players: their polyfunctional nature. Pleiotropy is backed by a peculiar subcellular compartmentation, i.e., sequestration of conditionally toxic proteins in a normally non-accessible subcellular compartment (the mitochondrial intermembrane space). The molecular determinants underlying the apoptotic and non-apoptotic functions have been deciphered for cyt-c and AIF (Cheung et al. 2006; Hao et al. 2005) but await characterization for the other mitochondrial factors.

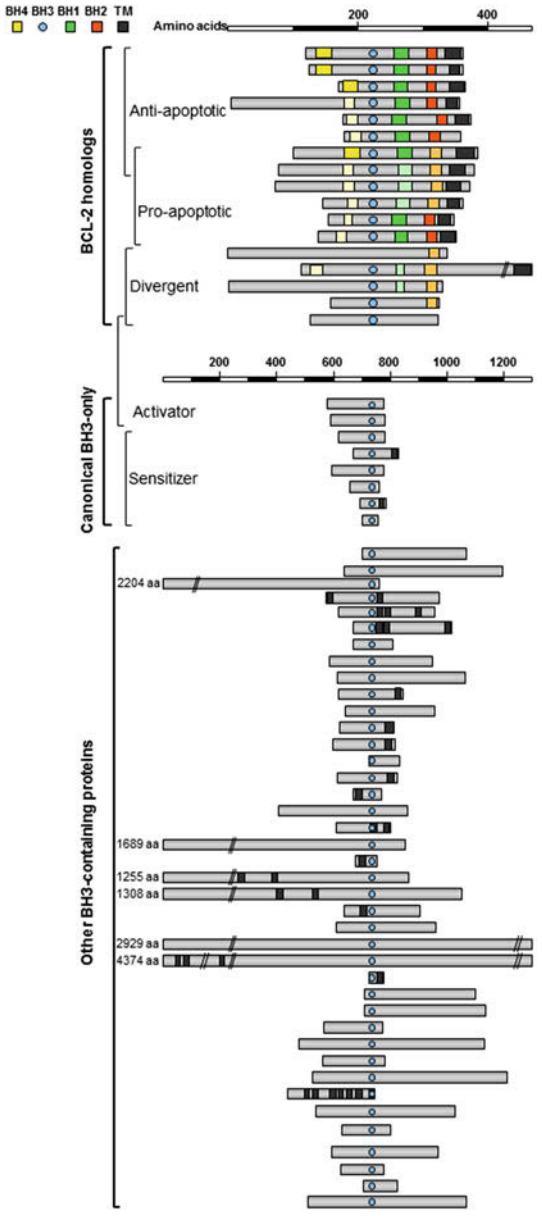
9.3 BCL-2 Homologous Proteins

BCL-2 family proteins control apoptosis upstream of the release of mitochondrial apoptogenic proteins and subsequent activation of caspases. This family of proteins comprises anti-apoptotic BCL-2 and pro-apoptotic BAX and their respective homologs. Our previous phylogenomic studies have revealed that BCL-2 homologous genes were restricted to metazoan species (and some animal viruses) and absent in fully sequenced genomes from Archaea, Eubacteria, Viridiplantae, Fungi, and other unicellular Eukaryota (Aouacheria et al. 2005), suggesting that this family arose in the metazoan stem. Logically, BCL-2 homologs are not found in the fully sequenced genomes of the choanoflagellate *Monosiga brevicollis* and the filasterean *Capsaspora owczarzaki* (King et al. 2008; Suga et al. 2013). Therefore, BCL-2 homologs most probably evolved only about 600 or 700 MYA and do not trace their origin back to the mitochondrial ancestor, their appearance being concomitant to the emergence of metazoan multicellularity.

Previous analysis indicated that gene duplication (for instance in marine invertebrates and fishes) and loss (e.g., in nematodes) played a prominent role in the

evolution of the BCL-2 family and contributed to the generation of lineage-specific diversity. Six representatives are present in the demosponge *Amphimedon queenslandica* (Srivastava et al. 2010), four in the placozoan *Trichoplax adhaerens* (Srivastava et al. 2008), nine in the cnidarian *Hydra vulgaris* (Lasi et al. 2010), ten in the zebrafish Danio rerio (Kratz et al. 2006), and fourteen in the humans (Aouacheria et al. 2005), whereas the worm *Caenorhabditis elegans* has a unique BCL-2-like gene (called CED-9) (Hengartner and Horvitz 1994) and the fruit fly (Drosophila melanogaster) only a pair of homologs (known as Buffy and Debcl) (Clavier et al. 2015). Thus, the BCL-2 gene complement of extant metazoans is not a mere function of organismal complexity but include differential gene expansion and loss across lineages. As a result, BCL-2 homologous genes are present in multiple paralogs showing substantial sequence divergence and BH (BCL-2 Homology) motif arrangements (Aouacheria et al. 2005; Aouacheria et al. 2013; Guillemin et al. 2009) (see Fig. 9.2). Phylogenetic reconstruction indicates that, in vertebrates, BCL-2 homologs segregate into three major clades: BCL-2-like, BAX-like, and BID-like members (Aouacheria et al. 2013). BCL-2-like and BAX-like members correspond to prosurvival and pro-apoptotic proteins, respectively, while BID-like members form a divergent group of proteins with diverse activities toward apoptosis. Within this last group, BPR/BCL2L12 is an anti-apoptotic protein (shown to reside in the nucleocytoplasmic compartment rather than mitochondria) (Stegh and DePinho 2011), BFK/BCL2L15 is poorly characterized but may constitute a pro-apoptotic protein (Coultas et al. 2003), and BCL-G/BCL2L14 appears to be neutral against apoptosis (Tischner and Villunger 2012). BCL-2 family members have been reported in multiple invertebrate species, including sponges, cnidarians, echinoderms, and mollusks (see Table 9.1), and many more are predicted [e.g., in *Trichoplax adhaerens* (Srivastava et al. 2008), *Ciona intestinalis* (Terajima et al. 2003), *Bombyx mori* (Zhang et al. 2010), *Apis mellifera* (Dallacqua and Bitondi 2014), and *Octopus vulgaris* (Castellanos-Martinez et al. 2014)]. Unfortunately, only a small proportion of these proteins have been fully characterized in an experimental way. Although there have been numerous published phylogenetic studies, none of them has addressed the full diversity of BCL-2 family members in invertebrates and a robust, comprehensive phylogenetic analysis is not yet available.

Since most bcl-2 family genes and proteins share commonalities in structure (e.g., an intron dividing the BH2 motif, and a similar "helical bundle" tridimensional fold—see Fig. 9.3), it appears likely that the diversity of metazoan BCL-2 genes was generated from a single precursor. The origin and functions of this ancestral BCL-2 protein are unknown. The early discovery that BCL-2 homologous proteins bear structural resemblance (analogy) to microbial toxins like colicins or the translocation domain of diphtheria toxin (Muchmore et al. 1996) has led to the speculation that they might have been acquired by horizontal gene transfer from the bacterial world. However, a set of viral proteins structurally related to BCL-2 (but functionally divergent) were recently characterized (Graham et al. 2008; Neidel et al. 2015), suggesting that the hypothesis of a viral origin for the founder gene should also be considered (Fig. 9.3). Whatever their origins, it seems reasonable to



Protein	Gene	Acc.°
BCL2	Bcl-2	P10415
BCL2L1	Bcl-x$_L$, Bcl2l1	Q07817
BCL2L2	Bcl-w, Bcl2l2	Q92843
MCL1	Mcl-1, Bcl2l3	Q07820
BCL2L10	Bcl2l10, Bcl-B, Nrh, Diva, Boo	Q9HD36
BCL2A1	Bcl2a1, Bfl1, Bcl2l5	Q16548
CED-9	CED-9 (Ce)	P41958
BUFFY	Buffy, dBorg2 (Dm)	Q8T8Y5
DEBCL	Debcl, dBorg1 (Dm)	Q9V9C8
BOK	Bok, Bcl2l9, Mtd	Q9UMX3
BAX	Bax, Bcl2l4	Q07812
BAK1	Bak, Bcl2l7	Q16611
BCL2L12	Bcl2l12, Bpr	Q9HB09
BCL2L13	Bcl2l13, Bcl-rambo, Mil1	Q9BXK5
BCL2L14	Bcl2l14, Bcl-G	Q9BZR8
BCL2L15	Bcl2l15, Bfk	Q5TBC7
BID	Bid	P55957
BCL2L11	Bim, Bod, Bcl2l11	O43521
BBC3	Puma	Q9BXH1
BAD	Bad, Bbc2, Bcl2l8	Q92934
BIK	Bik, Nbk, Blk	Q13323
BMF	Bmf	Q96LC9
EGL-1	EGL-1 (Ce)	O61667
HRK	Hrk, Dp5	O00198
PMAIP1	Noxa, Apr	Q13794
M	AF2240-M (NDV)	Q9WBL3
F0	F (NDV)	Q9DLD4
L	L (NDV)	Q6X1B6
APOL1	Apol1	O14791
APOL2	Apol2	Q9BQE5
APOL6	Apol6	Q9BWW8
ATG12	Atg12	O94817
AVEN	Aven	Q9NQS1
BECN1	Beclin-1	Q14457
BNIP1	Bnip1	Q12981
BNIP2	Bnip2	Q12982
BNIP3	Bnip3	Q12983
BNIP3L	Bnip3l, Nix	O60238
BLID	Brcc2	Q8IZY5
BNIP3	CEBNIP3 (Ce)	Q09969
CED-13	CED-13 (Ce)	Q9TY06
CLU	Clusterin	P10909
CORE	Core (HCV)	Q00269
CUL7	Cullin-7, p193 (Mm)	Q8VE73
E	E (SARS-CoV)	P59637
ERBB2	Erbb2, Her2	P04626
ERBB4	Erbb4, Her4	Q15303
ITM2B	Itm2b$_B$	Q9Y287
MOAP1	Map-1	Q96BY2
MCF	Mcf (Pl)	Q8KT65
HUWE1	Mule	Q7Z6Z7
PXT1	Pxt1 (Mm)	Q8K459
RAD9A	Rad9a, hRad9	Q99638
RAD9	Rad9sp (Sp)	P26306
HEPB2	Soul	Q9Y5Z4
SPHK2	Sphk2	Q9NRA0
CHMP5	Spike	Q9NZZ3
TGM2	Tg2, TGase	P21980
YNL305C	Ynl305cp, Ybh3p, Bxi1p (Sc)	C8ZGL9
VIRF-1	vIRF-1 (HHV8)	P88946
BLM-s	Ccdc132	Q0GET0
BOP	Bop	Q7L3V2
HBX	X (HBV)	Q65Z75
HBSP	P (HBV)	B5TZH4
Casp8p41	Casp8	Q14790

◀ **Fig. 9.2** BH motif composition in BCL-2 homologous proteins and BH3-containing proteins. Schematic representation and BH motif composition of BCL-2 homologous proteins (including BCL-2-like, BAX-like and BID-like subgroups), canonical BH3-only proteins and other reported BH3-containing proteins (with UniProtKB accession numbers). *Light shades* depicted BH motif is uncertain. Total amino acid (aa) number is indicated for proteins that were not drawn to scale. Abbreviations for non-human proteins: *Ce Caenorhabditis elegans*; *Dm Drosophila melanogaster*; *NDV* Newcastle disease virus; *Mm* Mus musculus; *HCV* hepatitis C virus; *Pl Photorhabdus luminescens*; *Sp Schizosaccharomyces pombe*; *SARS-CoV* human SARS coronavirus; *Sc Saccharomyces cerevisiae*; *HHV8* Human herpesvirus 8 (Kaposi's sarcoma-associated herpesvirus)

infer that the appearance of BCL-2 proteins might have been instrumental to the emergence of metazoan multicellularity, through the recruitment of mitochondria to a cell death program enabling tissue differentiation and homeostasis. However, the opposite assertion could also be true, namely recruitment of BCL-2 family proteins into a mitochondrial death program as a consequence of animal multicellularity. This issue is particularly interesting because in addition to their daily job in apoptosis, BCL-2 proteins have been shown to exert physiological, non-apoptotic functions such as regulation of mitochondrial dynamics, metabolism, DNA damage response, calcium homeostasis, general autophagy, and mitophagy (Pattingre et al. 2005; Alavian et al. 2011; Autret and Martin 2009; Chen et al. 2011; Chen and Pervaiz 2007; Gross 2006; Hardwick and Soane 2013; Hollville et al. 2014; Karbowski et al. 2006; Laulier and Lopez 2012; Murakawa et al. 2015; Perciavalle et al. 2012; Pinton and Rizzuto 2006; Wang et al. 2013). These findings suggesting that the function of BCL-2 proteins is pleiotropically linked to prosurvival traits in extant metazoan species raise the possibility that the ancestral function of BCL-2 proteins was unrelated to apoptosis regulation and that these proteins were exapted from an ancestor with an originally different function.

Evolutionary information is scarce about the beginnings of paralog divergence in the family and about how the repertoire of BCL-2 family genes evolved in the different metazoan lineages. Evidence of conserved colinearity (i.e., relict linkage) was gathered for the divergent BCL-2 homologs BID and BCL2L13 in vertebrate genomes (Aouacheria et al. 2005), providing information about the time when the cross talk between the intrinsic and extrinsic apoptosis pathways—which is mediated by BID—evolved. An early duplication involving these genes at the invertebrate-to-vertebrate transition, followed by two further duplications gave rise to the BID-like clade characterized by the absence of the C-terminal transmembrane segment (TM) (Aouacheria et al. 2013), illustrating the fact that many gene (sub) family expansions probably originally occur as tandem or proximal duplications (Charon et al. 2012; Fan et al. 2008; Srivastava et al. 2008). In the case of BCL2A1/BFL-1, a prosurvival BCL-2 homolog which appears to be found only in mammals, the exon encoding this TM segment has been replaced by a heterologous sequence, possibly as a result of duplication and shuffling events (Ko et al. 2007). BID, BCL2L13, and BCL2A1 correspond to phylogenetically recent innovations in metazoans, but other BCL-2 family genes are of more ancient origin such as proapoptotic BAK and prosurvival BCL2L1 (BCL-xL), for which close or

Table 9.1 Reported invertebrate BCL-2 proteins

Species	Reference	Gene/protein (Acc. °)	Motif composition (as published)	Function
Echinodermata				
Strongylocentrotus purpuratus	Robertson et al. (2006)	SPU_024469 SPU_006124 SPU_021416 SPU_001916 SPU_016028 SPU_014028 SPU_010641 SPU_010786 SPU_017154	BH4 TM TM BH3, TM BH3, TM BH3, TM	
Arthropoda				
Apis mellifera	Dallacqua and Bitondi (2014)	Ambuffy (A0A088ACI8)	BH1-4, TM	
Bombyx mori	Pan et al. (2014)	Bmbuffy (E9JEG2)	BH1-3, TM	Anti
Drosophila melanogaster	Quinn et al. (2003) Colussi et al. (2000)	Buffy (Q8T8Y5) Debcl (Q9V9C8)	BH1-3, TM BH1-3, TM	Anti Pro
Mollusca				
Ruditapes philippinarum	Lee et al. (2013)	RpBCL-2A (KC506418) RpBCL-2B (KC506419)	BH1-4, TM BH1-3, no TM	
Mytilus galloprovincialis	Estevez-Calvar et al. (2013)	Bcl2 (KC545829) Bax (KC545830)	BH1-4, TM BH1-3, TM	
Chlamys farreri	Qi et al. (2015)	CfBcl-2 (KJ611244) CfBax (KJ620057)	CfBcl-2: BH4, BH3, BH1, BH2, no TM CfBax: BH3, BH1, BH2, no TM	
Crassostrea hongkongensis	Xiang et al. (2015)	ChBax (KM262836) ChBak (KM262837)	BH3, BH1, BH2, TM	
Nematoda				
Caenorhabditis elegans	Hengartner and Horvitz (1994)	P41958	BH1-4, TM	Anti

(continued)

Table 9.1 (continued)

Species	Reference	Gene/protein (Acc. °)	Motif composition (as published)	Function
Platyhelminthes				
Schmidtea mediterranea	Bender et al. (2012)	Smed-Bak-1 (JN621808) Smed-Bak-2 (JN621809) Smed-bak-3 (JN621810) Smed-bok-2 (JN621814) Smed-bcl2-1 (FJ807655) Smed-bcl2-3 (JN621816)		Pro
Schistosoma mansoni	Lee et al. (2011)	sjA/smA sjB /smB sjBcl2/2 smBcl2/2 sjBcl2/1 smBcl2/1 sjC /smC sjD	BH1-4, TM BH1-4, TM BH1-4, TM BH1-4, no TM BH1 BH1 BH3 BH3	Anti Pro Pro
Cnidaria				
Aiptasia pallida	Dunn et al. (2006)	ABHP (DQ211980)	BH1, BH2, no TM	
Stylophora pistillata	Kvitt et al. (2011)	StyBcl-2-like (EU715319)	BH1-4, TM	
Hydra magnipapillata	Lasi et al. (2010)	HyBak-like 1 (EF104645) HyBak-like 2 (EU035760) HyBcl-2-like 1 (EF104646) HyBcl-2-like 2 (EF104647) HyBcl-2-like 3 (EU035765) HyBcl-2-like 4 (EU035764) HyBcl-2-like 5 (EU035763) HyBcl-2-like -6 (EU035762) HyBcl-2-like 7 (EU035761) HyBH3-only 1 (hma2.230679)	BH1-3, TM BH1-3, TM BH1-4, TM BH1-4, TM BH1-4, TM BH1-4, TM BH1-4, TM BH1-4, TM BH1-4, TM BH3 BH3 BH3 BH3	Pro Pro Anti Anti Anti Anti Slightly pro Anti Anti ND Pro Neutral Neutral

(continued)

Table 9.1 (continued)

Species	Reference	Gene/protein (Acc. °)	Motif composition (as published)	Function
		HyBH3-only 2 (hma2.221399) HyBH3-only 3 (hma2.218794) HyBH3-only 4 (hma2.224514)		
Porifera				
Geodia cydonium	Wiens et al. (2001)	BHP2-GC (AJ293508)	BH1, BH2, TM	Anti
Geodia cydonium *Suberites domuncula*	Wiens et al. (2000)	BHP1_GC (CAB97129) BHP1_SD (CAB97205)	BH1, BH2, TM	
Lubomirskia baicalensis	Wiens et al. (2006)	BAK-2_LUBAI (CAJ12144) BCL-2a_LUBAI (CAJ12145)	BH3, BH2, TM BH1-4, TM	

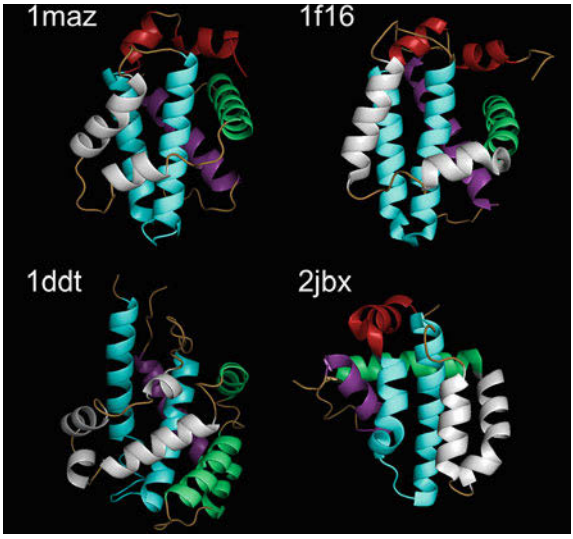


Fig. 9.3 Structural similarity between BCL-2 homologs and microbial proteins. Ribbon diagrams of anti-apoptotic protein BCL-xL (PDB code: 1maz), proapoptotic protein BAX (1f16), diphtheria toxin translocation domain (1ddt), and myxoma virus antiapoptotic protein M11L (2jbx). These proteins form a compact α-helical bundle with a pair of central helices (*in cyan*) surrounded by other (mainly amphipathic) helices. The figure was made with PyMol

divergent homologs, respectively, can be found in early-branching metazoans (Srivastava et al. 2010; Wiens et al. 2001; Wiens et al. 2000).

Particularly, intriguing is the issue of how opposite activities evolved in proteins that share a similar 3D structure, as do BCL-2-type and BAX-type proteins. The precise molecular determinants that underpin the extreme functional divergence

between structural homologs of the BCL-2 family are not completely understood. Distinct regions of the BCL-2 domain were shown to be involved in the functional dichotomy between pro- and anti-apoptotic members: the BH4 region (which is often located in the first α-helix) (Borner et al. 1994; Lee et al. 1996), the BH3 motif (Lee et al. 2014), and the α5-α6 helical hairpin motif (often referred to as a "pore-forming" domain) (Bleicken et al. 2013; Guillemin et al. 2010). However, subtle differences are scattered along the entire protein domain of pro- and anti-apoptotic BCL-2 proteins and it is expected that various sites may contribute to their antagonist actions on cell survival. A prime difference between BAX-type and BCL-2-type proteins might be related to the ability of proapoptotic homologs to self-assemble by forming dimers and higher order oligomers (Subburaj et al. 2015; Westphal et al. 2014), and of the prosurvival ones to inhibit these association processes in the context of the mitochondrial membrane by behaving like chain terminators, i.e., dominant negative forms (Reed 2006; Westphal et al. 2014). If this scenario is correct, then what distinguishes both types of proteins should be looked for in contact interfaces (with partners and/or lipids) in addition to isolated regions, bearing in mind that those interactions can be "cloudy" and involve distant residues. An alternative view may be that the separation between BCL-2-like and BAX-like family members has been overstated and that some kind of dualistic thinking is at work that somehow hides the partly artificial nature of the pro- versus anti-apoptotic dichotomy. This alternative scenario is not without support from a variety of experimental data, including the demonstration that (i) most prosurvival BCL-2 family proteins can be converted into death factors following proteolytic cleavage (Cheng et al. 1997; Clem et al. 1998; Kucharczak et al. 2005; Michels et al. 2004; Xue and Horvitz 1997); (ii) proapoptotic isoforms can be produced by alternative splicing of prosurvival bcl-2-like genes (Bae et al. 2000; Boise et al. 1993); (iii) pro-apoptotic BAK or BAK proteins can behave as prosurvival factors in specific cellular contexts or cell types (Kiefer et al. 1995; Lewis et al. 1999); and (iv) a number of BCL-2 family members were characterized both as pro- and anti-apoptotic factors (e.g., BCL2L10, BOK, and Bcl-rambo/BCL2L13) (Lee et al. 2001; Song et al. 1999; Aouacheria et al. 2001; Inohara et al. 1998; Ke et al. 2001; Zhang et al. 2001; Jensen et al. 2014). Hence, the scission between prosurvival and proapoptotic BCL-2 family members might be less definitive and clear than currently assumed.

9.4 BH3-Only Proteins

BCL2 homologous proteins act as receptors for BH3-only proteins, which are structurally unrelated proapoptotic molecules. In response to death signals, BH3-only proteins either inhibit the BCL-2-like apoptosis inhibitors or activate the BAX-like death activators. These proteins therefore form a signal-processing layer that connects onto the BCL-2/BAX core machinery the various inputs telling the cell either to survive or to commit suicide. Synthetic peptides encompassing the

BH3 motif of various BH3-only proteins were shown to bind with high affinities to a hydrophobic groove at the surface of prosurvival BCL-2 homologous proteins (Petros et al. 2000). Following this finding, BH3-mimetic drugs were developed that rapidly entered clinical trials as anticancer agents (Davids and Letai 2012). Given their key roles, the discovery of novel BH3-only proteins has represented and continues to represent a critical endeavor in the cell death field. Historically, this protein group contained nine non-homologous proteins discovered in the "1990s and early 2000s" (BIM, BMF, PUMA, NOXA, BAD, HRK, BIK, EGL1, and BID, herein termed "canonical" BH3-only proteins), which were sometimes erroneously appended to the protein family of BCL-2 homologs. In fact, only BID qualifies both as a BH3-only protein, as it contains a single BH3 motif, and as a BCL-2 homologous protein, because it shares a similar 3D structure with BCL-2 and BAX (Billen et al. 2008b; Chou et al. 1999; McDonnell et al. 1999). Current models of apoptosis regulation and a majority of review articles exclusively focus on the proapoptotic activity of these nine BH3-only proteins, ignoring the fact that the number of claimed BH3-only proteins has dramatically increased to reach a total of $\sim$40 (Aouacheria et al. 2015; Aouacheria et al. 2013) (Fig. 9.2). Contrary to the other BH motifs that were only detected in BCL-2 homologous proteins, BH3 motifs are now found in a gamut of folded (e.g., BCL-2) and unstructured protein domains [note that, except BID, all BH3-only proteins are intrinsically disordered proteins (Barrera-Vilarmau et al. 2011; Craxton et al. 2012; Hinds et al. 2007; Rogers et al. 2013; Yan et al. 2004)], bringing the grand total number of reported BH3 sequences to more than 60 unique instances. As a result, the evolutionary histories of BH3 motifs are singular, inherently coupled to the evolution of the proteins that harbor them, and therefore difficult to disentangle collectively. Depending on the case, evolution of BH3 motifs can be attributed to homologous processes (e.g., duplication divergence of BCL-2 family genes) or homoplastic mechanisms (e.g., random coincidence or convergence, as in the case of the E3 ubiquitin ligase MULE and the insecticidal toxin Mcf1, among many other putative instances). Interestingly, inspection of gene structures suggests that transfer events (e.g., exon shuffling) could also be involved, as illustrated by the relatively high similarity of the BCL-2 homolog BAK and the BH3-only gene BIK in their BH3 regions (Aouacheria et al. 2015).

The reason that explains this complicated situation has its root in the very nature of the BH3 motif, whose sequence signatures are diverse and of low complexity (i.e., very predictable) (Aouacheria et al. 2013). Following on this observation, we recently advanced the argument that the BH3 motif meets the criteria for classification as a short linear motif (SLiM) or a molecular recognition element/feature (MoRE/MoRF) involved in protein–protein interactions between structured domains (e.g., globular domains of the BCL-2 type) and between structured domains and intrinsically disordered proteins (as exemplified by the interaction between canonical BH3-only proteins and BCL-2-like or BAX-like proteins). Rather than considering the BH3 as an apoptotic motif per se, this novel conceptual framework poses this motif as a versatile and evolutionary plastic module associated with binding events in various branches of the tree of life, within metazoans

but also probably outside the animal kingdom as well. Future experiments will have to (i) assess the prevalence of BH3 motifs in proteins from non-metazoan species, (ii) unravel the identity of their putative receptors, and (iii) determine their possible roles in the biology of the cognate organisms.

9.5 Conclusion

To sum up, the BCL-2-regulated apoptotic pathway (a metazoan synapomorphy) emerged as the result of the interplay between an eukaryotic organelle (the mitochondrion) sequestrating proteins which have both vital and proapoptotic roles, a membranotropic structural domain (the BCL-2 globular fold) able to convey opposite activities toward cell survival and cell death, and a short and evolutionary plastic module (BH3) mediating protein–protein interactions. It is likely that acquisition of a proto-bcl-2 gene occurred only once during the evolution of the first multi-celled animals, followed by vertical evolutionary descent, lineage-specific diversification, and gene losses, contributing to the numerous morphological and lifestyle features of animals. Although sequences are "documents of evolutionary history" [in reference to Zuckerkandl and Pauling (1965)], it is hard to figure out in any real way whether the repertoire of molecules involved in the control of active cell death was "simple" or "complex" in the last common ancestor of modern-day animal species. Yet, as they are descendants of lineages that diverged early in the history of multicellular animals, the study of basal metazoan species can offer useful clues, e.g., about the presence of a BH3-dependent mitochondrial apoptotic pathway in their ancestors, or about the possible non-apoptotic function(s) of the BCL-2 ancestral protein. Whether metazoan BCL-2 homologous proteins emerged as stress-signaling molecules, or as switches connecting and controlling the execution of the various pathways involved in cell survival and death (including apoptosis, autophagy and programmed necrosis), or as key players serving biochemical functions distinct from cell death regulation remains an open question.

Acknowledgements The authors thank P. Pontarotti for the invitation to write this chapter. We are grateful to Dr. Valentine Rech De Laval for help with illustrations.

References

Alavian KN, Li H, Collis L, Bonanni L, Zeng L, Sacchetti S, Lazrove E, Nabili P, Flaherty B, Graham M, Chen Y, Messerli SM, Mariggio MA, Rahner C, McNay E, Shore GC, Smith PJ, Hardwick JM, Jonas EA (2011) Bcl-xL regulates metabolic efficiency of neurons through interaction with the mitochondrial F1FO ATP synthase. Nat Cell Biol 13(10):1224–1233. doi:10.1038/ncb2330

Aouacheria A (2014) The BCL-2 database, Act 2: moving beyond dualism to diversity and pleiotropy. Cell Death Dis 5:e981. doi:10.1038/cddis.2013.511

Aouacheria A, Arnaud E, Venet S, Lalle P, Gouy M, Rigal D, Gillet G (2001) Nrh, a human homologue of Nr-13 associates with Bcl-Xs and is an inhibitor of apoptosis. Oncogene 20 (41):5846–5855. doi:10.1038/sj.onc.1204740

Aouacheria A, Brunet F, Gouy M (2005) Phylogenomics of life-or-death switches in multicellular animals: Bcl-2, BH3-Only, and BNip families of apoptotic regulators. Mol Biol Evol 22 (12):2395–2416. doi:10.1093/molbev/msi234

Aouacheria A, Combet C, Tompa P, Hardwick JM (2015) Redefining the BH3 death domain as a short linear motif. Trends Biochem Sci 40(12):736–748. doi:10.1016/j.tibs.2015.09.007

Aouacheria A, Rech de Laval V, Combet C, Hardwick JM (2013) Evolution of Bcl-2 homology motifs: homology versus homoplasy. Trends Cell Biol 23(3):103–111. doi:10.1016/j.tcb.2012.10.010

Arnoult D, Gaume B, Karbowski M, Sharpe JC, Cecconi F, Youle RJ (2003) Mitochondrial release of AIF and EndoG requires caspase activation downstream of Bax/Bak-mediated permeabilization. The EMBO J 22(17):4385–4399. doi:10.1093/emboj/cdg423

Autret A, Martin SJ (2009) Emerging role for members of the Bcl-2 family in mitochondrial morphogenesis. Mol Cell 36(3):355–363. doi:10.1016/j.molcel.2009.10.011

Bae J, Leo CP, Hsu SY, Hsueh AJ (2000) MCL-1S, a splicing variant of the antiapoptotic BCL-2 family member MCL-1, encodes a proapoptotic protein possessing only the BH3 domain. J Biol Chem 275(33):25255–25261. doi:10.1074/jbc.M909826199

Barrera-Vilarmau S, Obregon P, de Alba E (2011) Intrinsic order and disorder in the bcl-2 member harakiri: insights into its proapoptotic activity. PLoS ONE 6(6):e21413. doi:10.1371/journal.pone.0021413

Bender CE, Fitzgerald P, Tait SW, Llambi F, McStay GP, Tupper DO, Pellettieri J, Sanchez Alvarado A, Salvesen GS, and Green DR (2012) Mitochondrial pathway of apoptosis is ancestral in metazoans. Proc Natl Acad Sci USA 109:4904–4909

Billen LP, Kokoski CL, Lovell JF, Leber B, Andrews DW (2008a) Bcl-XL inhibits membrane permeabilization by competing with Bax. PLoS Biol 6(6):e147. doi:10.1371/journal.pbio.0060147

Billen LP, Shamas-Din A, Andrews DW (2008b) Bid: a Bax-like BH3 protein. Oncogene 27 (Suppl 1):S93–104. doi:10.1038/onc.2009.47

Bleicken S, Wagner C, Garcia-Saez AJ (2013) Mechanistic differences in the membrane activity of Bax and Bcl-xL correlate with their opposing roles in apoptosis. Biophys J 104(2):421–431. doi:10.1016/j.bpj.2012.12.010

Boise LH, Gonzalez-Garcia M, Postema CE, Ding L, Lindsten T, Turka LA, Mao X, Nunez G, Thompson CB (1993) bcl-x, a bcl-2-related gene that functions as a dominant regulator of apoptotic cell death. Cell 74(4):597–608

Borner C, Martinou I, Mattmann C, Irmler M, Schaerer E, Martinou JC, Tschopp J (1994) The protein bcl-2 alpha does not require membrane attachment, but two conserved domains to suppress apoptosis. J cell biol 126(4):1059–1068

Bozhkov PV, Lam E (2011) Green death: revealing programmed cell death in plants. Cell Death Differ 18(8):1239–1240. doi:10.1038/cdd.2011.86

Buttner S, Eisenberg T, Carmona-Gutierrez D, Ruli D, Knauer H, Ruckenstuhl C, Sigrist C, Wissing S, Kollroser M, Frohlich KU, Sigrist S, Madeo F (2007) Endonuclease G regulates budding yeast life and death. Mol Cell 25(2):233–246. doi:10.1016/j.molcel.2006.12.021

Castellanos-Martinez S, Arteta D, Catarino S, Gestal C (2014) De novo transcriptome sequencing of the Octopus vulgaris hemocytes using Illumina RNA-Seq technology: response to the infection by the gastrointestinal parasite Aggregata octopiana. PLoS ONE 9(10):e107873. doi:10.1371/journal.pone.0107873

Charon C, Bruggeman Q, Thareau V, Henry Y (2012) Gene duplication within the Green Lineage: the case of TEL genes. J Exp Bot 63(14):5061–5077. doi:10.1093/jxb/ers181

Chen YB, Aon MA, Hsu YT, Soane L, Teng X, McCaffery JM, Cheng WC, Qi B, Li H, Alavian KN, Dayhoff-Brannigan M, Zou S, Pineda FJ, O'Rourke B, Ko YH, Pedersen PL, Kaczmarek LK, Jonas EA, Hardwick JM (2011) Bcl-xL regulates mitochondrial energetics by

stabilizing the inner membrane potential. J cell biol 195(2):263–276. doi:10.1083/jcb.201108059

Chen ZX, Pervaiz S (2007) Bcl-2 induces pro-oxidant state by engaging mitochondrial respiration in tumor cells. Cell Death Differ 14(9):1617–1627. doi:10.1038/sj.cdd.4402165

Cheng EH, Kirsch DG, Clem RJ, Ravi R, Kastan MB, Bedi A, Ueno K, Hardwick JM (1997) Conversion of Bcl-2 to a Bax-like death effector by caspases. Science 278(5345):1966–1968

Cheung EC, Joza N, Steenaart NA, McClellan KA, Neuspiel M, McNamara S, MacLaurin JG, Rippstein P, Park DS, Shore GC, McBride HM, Penninger JM, Slack RS (2006) Dissociating the dual roles of apoptosis-inducing factor in maintaining mitochondrial structure and apoptosis. EMBO J 25(17):4061–4073. doi:10.1038/sj.emboj.7601276

Chou JJ, Li H, Salvesen GS, Yuan J, Wagner G (1999) Solution structure of BID, an intracellular amplifier of apoptotic signaling. Cell 96(5):615–624

Clavier A, Rincheval-Arnold A, Colin J, Mignotte B, Guenal I (2015) Apoptosis in Drosophila: which role for mitochondria? Apoptosis Int J Program Cell Death. doi:10.1007/s10495-015-1209-y

Clem RJ, Cheng EH, Karp CL, Kirsch DG, Ueno K, Takahashi A, Kastan MB, Griffin DE, Earnshaw WC, Veliuona MA, Hardwick JM (1998) Modulation of cell death by Bcl-XL through caspase interaction. Proc Natl Acad Sci USA 95(2):554–559

Colussi PA, Quinn LM, Huang DC, Coombe M, Read SH, Richardson H, Kumar S (2000) Debcl, a proapoptotic Bcl-2 homologue, is a component of the Drosophila melanogaster cell death machinery. J Cell Biol 148:703–714

Coultas L, Pellegrini M, Visvader JE, Lindeman GJ, Chen L, Adams JM, Huang DC, Strasser A (2003) Bfk: a novel weakly proapoptotic member of the Bcl-2 protein family with a BH3 and a BH2 region. Cell Death Differ 10(2):185–192. doi:10.1038/sj.cdd.4401204

Craxton A, Butterworth M, Harper N, Fairall L, Schwabe J, Ciechanover A, Cohen GM (2012) NOXA, a sensor of proteasome integrity, is degraded by 26S proteasomes by an ubiquitin-independent pathway that is blocked by MCL-1. Cell Death Differ 19(9):1424–1434. doi:10.1038/cdd.2012.16

Cregan SP, Dawson VL, Slack RS (2004) Role of AIF in caspase-dependent and caspase-independent cell death. Oncogene 23(16):2785–2796. doi:10.1038/sj.onc.1207517

Czabotar PE, Lessene G, Strasser A, Adams JM (2014) Control of apoptosis by the BCL-2 protein family: implications for physiology and therapy. Nat Rev Mol Cell Biol 15(1):49–63. doi:10.1038/nrm3722

Dallacqua RP, Bitondi MM (2014) Dimorphic ovary differentiation in honeybee (Apis mellifera) larvae involves caste-specific expression of homologs of ark and buffy cell death genes. PLoS ONE 9(5):e98088. doi:10.1371/journal.pone.0098088

Davids MS, Letai A (2012) Targeting the B-cell lymphoma/leukemia 2 family in cancer. J Clin Oncol: Official J Ame Soc Clin Oncol 30(25):3127–3135. doi:10.1200/JCO.2011.37.0981

Doerflinger M, Glab JA, Puthalakath H (2015) BH3-only proteins: a 20-year stock-take. FEBS J 282(6):1006–1016. doi:10.1111/febs.13190

Dunn SR, Phillips WS, Spatafora JW, Green DR, Weis VM (2006) Highly conserved caspase and Bcl-2 homologues from the sea anemone Aiptasia pallida: lower metazoans as models for the study of apoptosis evolution. J Mol Evol 63:95–107

Dwyer DJ, Winkler JA (2013) Identification and characterization of programmed cell death markers in bacterial models. Methods Mol Biol 1004:145–159. doi:10.1007/978-1-62703-383-1_11

Edlich F, Banerjee S, Suzuki M, Cleland MM, Arnoult D, Wang C, Neutzner A, Tjandra N, Youle RJ (2011) Bcl-x(L) retrotranslocates bax from the mitochondria into the cytosol. Cell 145(1):104–116. doi:10.1016/j.cell.2011.02.034

Estevez-Calvar N, Romero A, Figueras A, Novoa B (2013) Genes of the mitochondrial apoptotic pathway in Mytilus galloprovincialis. PloS one. 8:e61502

Elmore S (2007) Apoptosis: a review of programmed cell death. Toxicol Pathol 35(4):495–516. doi:10.1080/01926230701320337

Fan C, Chen Y, Long M (2008) Recurrent tandem gene duplication gave rise to functionally divergent genes in drosophila. Mol Biol Evol 25(7):1451–1458. doi:10.1093/molbev/msn089

Garrido C, Kroemer G (2004) Life's smile, death's grin: vital functions of apoptosis-executing proteins. Curr Opin Cell Biol 16(6):639–646. doi:10.1016/j.ceb.2004.09.008

Graham SC, Bahar MW, Cooray S, Chen RA, Whalen DM, Abrescia NG, Alderton D, Owens RJ, Stuart DI, Smith GL, Grimes JM (2008) Vaccinia virus proteins A52 and B14 Share a Bcl-2-like fold but have evolved to inhibit NF-kappaB rather than apoptosis. PLoS Pathog 4 (8):e1000128. doi:10.1371/journal.ppat.1000128

Gross A (2006) BID as a double agent in cell life and death. Cell Cycle 5(6):582–584

Guillemin Y, Lalle P, Gillet G, Guerin JF, Hamamah S, Aouacheria A (2009) Oocytes and early embryos selectively express the survival factor BCL2L10. J Mol Med (Berl) 87(9):923–940. doi:10.1007/s00109-009-0495-7

Guillemin Y, Lopez J, Gimenez D, Fuertes G, Valero JG, Blum L, Gonzalo P, Salgado J, Girard-Egrot A, Aouacheria A (2010) Active fragments from pro- and antiapoptotic BCL-2 proteins have distinct membrane behavior reflecting their functional divergence. PLoS ONE 5 (2):e9066. doi:10.1371/journal.pone.0009066

Hao Z, Duncan GS, Chang CC, Elia A, Fang M, Wakeham A, Okada H, Calzascia T, Jang Y, You-Ten A, Yeh WC, Ohashi P, Wang X, Mak TW (2005) Specific ablation of the apoptotic functions of cytochrome C reveals a differential requirement for cytochrome C and Apaf-1 in apoptosis. Cell 121(4):579–591. doi:10.1016/j.cell.2005.03.016

Hardwick JM, Soane L (2013) Multiple functions of BCL-2 family proteins. Cold Spring Harb Perspect Biol 5(2). doi:10.1101/cshperspect.a008722

Hengartner MO, Horvitz HR (1994) C. elegans cell survival gene ced-9 encodes a functional homolog of the mammalian proto-oncogene bcl-2. Cell 76(4):665–676

Hinds MG, Smits C, Fredericks-Short R, Risk JM, Bailey M, Huang DC, Day CL (2007) Bim, Bad and Bmf: intrinsically unstructured BH3-only proteins that undergo a localized conformational change upon binding to prosurvival Bcl-2 targets. Cell Death Differ 14 (1):128–136. doi:10.1038/sj.cdd.4401934

Hollville E, Carroll RG, Cullen SP, Martin SJ (2014) Bcl-2 family proteins participate in mitochondrial quality control by regulating Parkin/PINK1-dependent mitophagy. Mol Cell 55 (3):451–466. doi:10.1016/j.molcel.2014.06.001

Huang KJ, Ku CC, Lehman IR (2006) Endonuclease G: a role for the enzyme in recombination and cellular proliferation. Proc Natl Acad Sci USA 103(24):8995–9000. doi:10.1073/pnas.0603445103

Inohara N, Gourley TS, Carrio R, Muniz M, Merino J, Garcia I, Koseki T, Hu Y, Chen S, Nunez G (1998) Diva, a Bcl-2 homologue that binds directly to Apaf-1 and induces BH3-independent cell death. J Biol Chem 273(49):32479–32486

Jensen SA, Calvert AE, Volpert G, Kouri FM, Hurley LA, Luciano JP, Wu Y, Chalastanis A, Futerman AH, Stegh AH (2014) Bcl2L13 is a ceramide synthase inhibitor in glioblastoma. Proc Natl Acad Sci USA 111(15):5682–5687. doi:10.1073/pnas.1316700111

Karbowski M, Norris KL, Cleland MM, Jeong SY, Youle RJ (2006) Role of Bax and Bak in mitochondrial morphogenesis. Nature 443(7112):658–662. doi:10.1038/nature05111

Ke N, Godzik A, Reed JC (2001) Bcl-B, a novel Bcl-2 family member that differentially binds and regulates Bax and Bak. J Biol Chem 276(16):12481–12484. doi:10.1074/jbc.C000871200

Kerr JF, Wyllie AH, Currie AR (1972) Apoptosis: a basic biological phenomenon with wide-ranging implications in tissue kinetics. Br J Cancer 26(4):239–257

Kiefer MC, Brauer MJ, Powers VC, Wu JJ, Umansky SR, Tomei LD, Barr PJ (1995) Modulation of apoptosis by the widely distributed Bcl-2 homologue Bak. Nature 374(6524):736–739. doi:10.1038/374736a0

King N, Westbrook MJ, Young SL, Kuo A, Abedin M, Chapman J, Fairclough S, Hellsten U, Isogai Y, Letunic I, Marr M, Pincus D, Putnam N, Rokas A, Wright KJ, Zuzow R, Dirks W, Good M, Goodstein D, Lemons D, Li W, Lyons JB, Morris A, Nichols S, Richter DJ, Salamov A, Sequencing JG, Bork P, Lim WA, Manning G, Miller WT, McGinnis W, Shapiro H, Tjian R, Grigoriev IV, Rokhsar D (2008) The genome of the choanoflagellate

Monosiga brevicollis and the origin of metazoans. Nature 451(7180):783–788. doi:10.1038/nature06617

Ko JK, Choi KH, Pan Z, Lin P, Weisleder N, Kim CW, Ma J (2007) The tail-anchoring domain of Bfl1 and HCCS1 targets mitochondrial membrane permeability to induce apoptosis. J Cell Sci 120(Pt 16):2912–2923. doi:10.1242/jcs.006197

Kratz E, Eimon PM, Mukhyala K, Stern H, Zha J, Strasser A, Hart R, Ashkenazi A (2006) Functional characterization of the Bcl-2 gene family in the zebrafish. Cell Death Differ 13(10):1631–1640. doi:10.1038/sj.cdd.4402016

Kucharczak JF, Simmons MJ, Duckett CS, Gelinas C (2005) Constitutive proteasome-mediated turnover of Bfl-1/A1 and its processing in response to TNF receptor activation in FL5.12 pro-B cells convert it into a prodeath factor. Cell Death Differ 12(9):1225–1239. doi:10.1038/sj.cdd.4401684

Kvitt H, Rosenfeld H, Zandbank K, Tchernov C (2011) Regulation of apoptotic pathways by Stylophora pistillata (Anthozoa, Pocilloporidae) to survive thermal stress and bleaching. PloS one. 6:e28665.

Lasi M, Pauly B, Schmidt N, Cikala M, Stiening B, Kasbauer T, Zenner G, Popp T, Wagner A, Knapp RT, Huber AH, Grunert M, Soding J, David CN, Bottger A (2010) The molecular cell death machinery in the simple cnidarian hydra includes an expanded caspase family and pro- and anti-apoptotic Bcl-2 proteins. Cell Res 20(7):812–825. doi:10.1038/cr.2010.66

Laulier C, Lopez BS (2012) The secret life of Bcl-2: apoptosis-independent inhibition of DNA repair by Bcl-2 family members. Mutat Res 751(2):247–257. doi:10.1016/j.mrrev.2012.05.002

Lee EF, Clarke OB, Evangelista M, Feng Z, Speed TP, Tchoubrieva EB, Strasser A, Kalinna BH, Colman PM, Fairlie WD (2011) Discovery and molecular characterization of a Bcl-2-regulated cell death pathway in schistosomes. Proc Natl Acad Sci USA 108:6999–7003

Lee EF, Dewson G, Evangelista M, Pettikiriarachchi A, Gold GJ, Zhu H, Colman PM, Fairlie WD (2014) The functional differences between pro-survival and pro-apoptotic B cell lymphoma 2 (Bcl-2) proteins depend on structural differences in their Bcl-2 homology 3 (BH3) domains. J Biol Chem 289(52):36001–36017. doi:10.1074/jbc.M114.610758

Lee LC, Hunter JJ, Mujeeb A, Turck C, Parslow TG (1996) Evidence for alpha-helical conformation of an essential N-terminal region in the human Bcl2 protein. J Biol Chem 271 (38):23284–23288

Lee R, Chen J, Matthews CP, McDougall JK, Neiman PE (2001) Characterization of NR13-related human cell death regulator, Boo/Diva, in normal and cancer tissues. Biochim Biophys Acta 1520(3):187–194

Lee Y, Whang I, Lee S, Menike U, Oh C, Kang DH, Heo GJ, Lee J, De Zoysa M (2013) Two molluscan BCL-2 family members from Manila clam, Ruditapes philippinarum: molecular characterization and immune responses. Fish Shellfish Immu 34:1628–1634

Lewis J, Oyler GA, Ueno K, Fannjiang YR, Chau BN, Vornov J, Korsmeyer SJ, Zou S, Hardwick JM (1999) Inhibition of virus-induced neuronal apoptosis by Bax. Nat Med 5(7):832–835. doi:10.1038/10556

Llambi F, Moldoveanu T, Tait SW, Bouchier-Hayes L, Temirov J, McCormick LL, Dillon CP, Green DR (2011) A unified model of mammalian BCL-2 protein family interactions at the mitochondria. Mol Cell 44(4):517–531. doi:10.1016/j.molcel.2011.10.001

Lorenzo HK, Susin SA (2004) Mitochondrial effectors in caspase-independent cell death. FEBS Lett 557(1–3):14–20

Madeo F, Frohlich E, Frohlich KU (1997) A yeast mutant showing diagnostic markers of early and late apoptosis. J cell biol 139(3):729–734

McDonnell JM, Fushman D, Milliman CL, Korsmeyer SJ, Cowburn D (1999) Solution structure of the proapoptotic molecule BID: a structural basis for apoptotic agonists and antagonists. Cell 96(5):625–634

Michels J, O'Neill JW, Dallman CL, Mouzakiti A, Habens F, Brimmell M, Zhang KY, Craig RW, Marcusson EG, Johnson PW, Packham G (2004) Mcl-1 is required for Akata6 B-lymphoma cell survival and is converted to a cell death molecule by efficient caspase-mediated cleavage. Oncogene 23(28):4818–4827. doi:10.1038/sj.onc.1207648

Muchmore SW, Sattler M, Liang H, Meadows RP, Harlan JE, Yoon HS, Nettesheim D, Chang BS, Thompson CB, Wong SL, Ng SL, Fesik SW (1996) X-ray and NMR structure of human Bcl-xL, an inhibitor of programmed cell death. Nature 381(6580):335–341. doi:10.1038/381335a0

Murakawa T, Yamaguchi O, Hashimoto A, Hikoso S, Takeda T, Oka T, Yasui H, Ueda H, Akazawa Y, Nakayama H, Taneike M, Misaka T, Omiya S, Shah AM, Yamamoto A, Nishida K, Ohsumi Y, Okamoto K, Sakata Y, Otsu K (2015) Bcl-2-like protein 13 is a mammalian Atg32 homologue that mediates mitophagy and mitochondrial fragmentation. Nat Commun 6:7527. doi:10.1038/ncomms8527

Neidel S, Maluquer de Motes C, Mansur DS, Strnadova P, Smith GL, Graham SC (2015) Vaccinia virus protein A49 is an unexpected member of the B-cell Lymphoma (Bcl)-2 protein family. J Biol Chem 290(10):5991–6002. doi:10.1074/jbc.M114.624650

Pan C, Hu YF, Yi HS, Song J, Wang L, Pan MH, Lu C (2014) Role of Bmbuffy in hydroxycamptothecine-induced apoptosis in BmN-SWU1 cells of the silkworm, Bombyx mori. Biochem Bioph Res Co 447:237–243

Pattingre S, Tassa A, Qu X, Garuti R, Liang XH, Mizushima N, Packer M, Schneider MD, Levine B (2005) Bcl-2 antiapoptotic proteins inhibit Beclin 1-dependent autophagy. Cell 122(6):927–939. doi:10.1016/j.cell.2005.07.002

Perciavalle RM, Stewart DP, Koss B, Lynch J, Milasta S, Bathina M, Temirov J, Cleland MM, Pelletier S, Schuetz JD, Youle RJ, Green DR, Opferman JT (2012) Anti-apoptotic MCL-1 localizes to the mitochondrial matrix and couples mitochondrial fusion to respiration. Nat Cell Biol 14(6):575–583. doi:10.1038/ncb2488

Petros AM, Nettesheim DG, Wang Y, Olejniczak ET, Meadows RP, Mack J, Swift K, Matayoshi ED, Zhang H, Thompson CB, Fesik SW (2000) Rationale for Bcl-xL/Bad peptide complex formation from structure, mutagenesis, and biophysical studies. Protein Sci: Publ Protein Soc 9(12):2528–2534. doi:10.1110/ps.9.12.2528

Pinton P, Rizzuto R (2006) Bcl-2 and Ca2 + homeostasis in the endoplasmic reticulum. Cell Death Differ 13(8):1409–1418. doi:10.1038/sj.cdd.4401960

Porter AG, Urbano AG (2006) Does apoptosis-inducing factor (AIF) have both life and death functions in cells? BioEssays: News Rev Mol Cell Dev Biol 28(8):834–843. doi:10.1002/bies.20444

Qi H, Miao G, Li L, Que H, Zhang G (2015) Identification and functional characterization of two Bcl-2 family protein genes in Zhikong scallop Chlamys farreri. Fish Shellfish Immun 44:147–155

Quinn L, Coombe M, Mills K, Daish T, Colussi P, Kumar S, Richardson H (2003) Buffy, a Drosophila Bcl-2 protein, has anti-apoptotic and cell cycle inhibitory functions. EMBO J 22:3568–3579

Rech de Laval V, Deleage G, Aouacheria A, Combet C (2014) BCL2DB: database of BCL-2 family members and BH3-only proteins. Database: J Biol Databases Curation 2014: bau013. doi:10.1093/database/bau013

Reed JC (2006) Proapoptotic multidomain Bcl-2/Bax-family proteins: mechanisms, physiological roles, and therapeutic opportunities. Cell Death Differ 13(8):1378–1386. doi:10.1038/sj.cdd.4401975

Robertson AJ, Croce J, Carbonneau S, Voronina E, Miranda E, McClay DR, Coffman JA. 2006. The genomic underpinnings of apoptosis in Strongylocentrotus purpuratus. Dev Biol 300:321–334

Rogers JM, Steward A, Clarke J (2013) Folding and binding of an intrinsically disordered protein: fast, but not diffusion-limited. J Am Chem Soc 135(4):1415–1422. doi:10.1021/ja309527h

Saelens X, Festjens N, Vande Walle L, van Gurp M, van Loo G, Vandenabeele P (2004) Toxic proteins released from mitochondria in cell death. Oncogene 23(16):2861–2874. doi:10.1038/sj.onc.1207523

Shamas-Din A, Brahmbhatt H, Leber B, Andrews DW (2011) H3-only proteins: orchestrators of apoptosis. Biochim Biophys Acta 4:508–520. doi:10.1016/j.bbamcr.2010.11.024

Song Q, Kuang Y, Dixit VM, Vincenz C (1999) Boo, a novel negative regulator of cell death, interacts with Apaf-1. EMBO J 18(1):167–178. doi:10.1093/emboj/18.1.167

Sorrentino L, Calogero AM, Pandini V, Vanoni MA, Sevrioukova IF, Aliverti A (2015) Key role of the adenylate moiety and integrity of the adenylate-binding site for the NAD(+)/H binding to mitochondrial apoptosis-inducing factor. Biochemistry 54(47):6996–7009. doi:10.1021/acs.biochem.5b00898

Srivastava M, Begovic E, Chapman J, Putnam NH, Hellsten U, Kawashima T, Kuo A, Mitros T, Salamov A, Carpenter ML, Signorovitch AY, Moreno MA, Kamm K, Grimwood J, Schmutz J, Shapiro H, Grigoriev IV, Buss LW, Schierwater B, Dellaporta SL, Rokhsar DS (2008) The Trichoplax genome and the nature of placozoans. Nature 454(7207):955–960. doi:10.1038/nature07191

Srivastava M, Simakov O, Chapman J, Fahey B, Gauthier ME, Mitros T, Richards GS, Conaco C, Dacre M, Hellsten U, Larroux C, Putnam NH, Stanke M, Adamska M, Darling A, Degnan SM, Oakley TH, Plachetzki DC, Zhai Y, Adamski M, Calcino A, Cummins SF, Goodstein DM, Harris C, Jackson DJ, Leys SP, Shu S, Woodcroft BJ, Vervoort M, Kosik KS, Manning G, Degnan BM, Rokhsar DS (2010) The Amphimedon queenslandica genome and the evolution of animal complexity. Nature 466(7307):720–726. doi:10.1038/nature09201

Stegh AH, DePinho RA (2011) Beyond effector caspase inhibition: Bcl2L12 neutralizes p53 signaling in glioblastoma. Cell Cycle 10(1):33–38

Subburaj Y, Cosentino K, Axmann M, Pedrueza-Villalmanzo E, Hermann E, Bleicken S, Spatz J, Garcia-Saez AJ (2015) Bax monomers form dimer units in the membrane that further self-assemble into multiple oligomeric species. Nature Commun 6:8042. doi:10.1038/ncomms9042

Suga H, Chen Z, de Mendoza A, Sebe-Pedros A, Brown MW, Kramer E, Carr M, Kerner P, Vervoort M, Sanchez-Pons N, Torruella G, Derelle R, Manning G, Lang BF, Russ C, Haas BJ, Roger AJ, Nusbaum C, Ruiz-Trillo I (2013) The Capsaspora genome reveals a complex unicellular prehistory of animals. Nat Commun 4:2325. doi:10.1038/ncomms3325

Tait SW, Green DR (2013) Mitochondrial regulation of cell death. Cold Spring Harb perspect biol 5(9). doi:10.1101/cshperspect.a008706

Terajima D, Shida K, Takada N, Kasuya A, Rokhsar D, Satoh N, Satake M, Wang HG (2003) Identification of candidate genes encoding the core components of the cell death machinery in the Ciona intestinalis genome. Cell Death Differ 10(6):749–753. doi:10.1038/sj.cdd.4401223

Tischner D, Villunger A (2012) Bcl-G acquitted of murder! Cell Death Dis 3:e405. doi:10.1038/cddis.2012.147

Vahsen N, Cande C, Briere JJ, Benit P, Joza N, Larochette N, Mastroberardino PG, Pequignot MO, Casares N, Lazar V, Feraud O, Debili N, Wissing S, Engelhardt S, Madeo F, Piacentini M, Penninger JM, Schagger H, Rustin P, Kroemer G (2004) AIF deficiency compromises oxidative phosphorylation. EMBO J 23(23):4679–4689. doi:10.1038/sj.emboj.7600461

Volkmann N, Marassi FM, Newmeyer DD, Hanein D (2014) The rheostat in the membrane: BCL-2 family proteins and apoptosis. Cell Death Differ 21(2):206–215. doi:10.1038/cdd.2013.153

Wang X, Bathina M, Lynch J, Koss B, Calabrese C, Frase S, Schuetz JD, Rehg JE, Opferman JT (2013) Deletion of MCL-1 causes lethal cardiac failure and mitochondrial dysfunction. Genes Dev 27(12):1351–1364. doi:10.1101/gad.215855.113

Westphal D, Kluck RM, Dewson G (2014) Building blocks of the apoptotic pore: how Bax and Bak are activated and oligomerize during apoptosis. Cell Death Differ 21(2):196–205. doi:10.1038/cdd.2013.139

Wiens M, Belikov SI, Kaluzhnaya OV, Schroder HC, Hamer B, Perovic-Ottstadt S, Borejko A, Luthringer B, Muller IM, Muller WE (2006) Axial (apical-basal) expression of pro-apoptotic and pro-survival genes in the lake baikal demosponge Lubomirskia baicalensis. DNA Cell Biol 25:152–164

Wiens M, Diehl-Seifert B, Muller WE (2001) Sponge Bcl-2 homologous protein (BHP2-GC) confers distinct stress resistance to human HEK-293 cells. Cell Death Differ 8(9):887–898. doi:10.1038/sj.cdd.4400906

Wiens M, Krasko A, Muller CI, Muller WE (2000) Molecular evolution of apoptotic pathways: cloning of key domains from sponges (Bcl-2 homology domains and death domains) and their phylogenetic relationships. J Mol Evol 50(6):520–531

Xiang Z, Qu F, Wang F, Xiao S, Jun L, Zhang Y, Yu Z (2015) ChBax/Bak as key regulators of the mitochondrial apoptotic pathway: cloned and characterized in Crassostrea hongkongensis. Fish Shellfish Immun. 42:225–232

Xue D, Horvitz HR (1997) Caenorhabditis elegans CED-9 protein is a bifunctional cell-death inhibitor. Nature 390(6657):305–308. doi:10.1038/36889

Yan N, Gu L, Kokel D, Chai J, Li W, Han A, Chen L, Xue D, Shi Y (2004) Structural, biochemical, and functional analyses of CED-9 recognition by the proapoptotic proteins EGL-1 and CED-4. Mol Cell 15(6):999–1006. doi:10.1016/j.molcel.2004.08.022

Zhang H, Holzgreve W, De Geyter C (2001) Bcl2-L-10, a novel anti-apoptotic member of the Bcl-2 family, blocks apoptosis in the mitochondria death pathway but not in the death receptor pathway. Hum Mol Genet 10(21):2329–2339

Zhang JY, Pan MH, Sun ZY, Huang SJ, Yu ZS, Liu D, Zhao DH, Lu C (2010) The genomic underpinnings of apoptosis in the silkworm. Bombyx mori BMC genomics 11:611. doi:10.1186/1471-2164-11-611

Zuckerkandl E, Pauling L (1965) Molecules as documents of evolutionary history. J Theor Biol 8(2):357–366

Chapter 10
The Axial Level of the Heart in Snakes

J.W. Faber, M.K. Richardson, E.M. Dondorp and R.E. Poelmann

Abstract Snakes are remarkable for their extremely elongated body plan, loss of limbs and increase in numbers of vertebrae. These adaptations have a number of functional advantages including the ability to enter the underground burrows of prey, to move around the tree canopies with relative ease, to strike at prey and to swim in water. Compared to four-legged squamate ancestors, snakes show a remarkable displacement and elongation of their viscera along their extended body axes. The heart in snakes varies widely in position along the body axis, and this variation in axial level may be correlated with phylogeny and/or with habitat. Here, we review the existing literature and examine the variation in the axial level of the heart in a sample of 15 mature snake specimens from arboreal, terrestrial, subterrestrial, semi-aquatic and aquatic habitats. Alcohol-preserved animals were dissected and examined radiographically. The results show that there is a trend towards a more rostral positioning of the heart along the body axis in arboreal snakes. However, this axial level falls within the range of that observed in terrestrial species. Phylogeny does not fully account for the observed axial level of the heart in the species examined. We conclude that a combination of habitat (both arboreal and aquatic) and phylogeny may explain the axial level of the heart in snakes.

J.W. Faber · M.K. Richardson
Sylvius Laboratory, Institute of Biology, Leiden University, Sylviusweg 72,
2333BE Leiden, The Netherlands
e-mail: m.k.richardson@biology.leidenuniv.nl

E.M. Dondorp
Naturalis Biodiversity Center, Darwinweg 2, C.04.32, 2333CR Leiden,
The Netherlands
e-mail: esther.dondorp@naturalis.nl

R.E. Poelmann (✉)
Department of Cardiology, Leiden University Medical Center, Leiden University,
Albinusdreef 2, Leiden, The Netherlands
e-mail: r.e.poelmann@lumc.nl

P. Pontarotti (ed.), *Evolutionary Biology*, DOI 10.1007/978-3-319-41324-2_10

10.1 Introduction

Snakes (Squamata: Serpentes) are characterised by a number of specialised morphological features including elongation of the body and viscera, high vertebral count and loss of limbs. There are at least 3378 snake species known from 23 different families (Pincheira-Donoso et al. 2013). Snakes have been very successful and have adapted to huge differences in climate and can therefore be found in almost every habitat except for the polar regions.

We can distinguish at least five types of snakes according to their habitats: they are arboreal, terrestrial, subterrestrial, semi-aquatic, and aquatic species (Seymour and Lillywhite 1976; Murphy 2012). Arboreal snakes spend most of their life in trees, terrestrial snakes live on the ground, subterrestrial snakes live in burrows and spend their life mostly underground, semi-aquatic snakes spend a lot of time in and around water but lay eggs on land, while aquatic snakes spend their entire life in water and are ovoviviparous (Neill et al. 1964).

The heart is usually located between the first and second quarter of the total body length in snakes, although its precise position varies between species (Lillywhite 1987, 1988). According to some researchers, this difference in the axial level of the heart is related to the habitat (Seymour 1987; Lillywhite 1996). Those authors argue that, in terrestrial environments, gravity has a profound influence on the cardiovascular system; in aquatic environments, this effect is possibly nullified by the buoyancy of water (Seymour 1987; Lillywhite 1996). Additionally, they argue that when a snake is positioned vertically in air, the influence of gravity will reduce the blood flow to the brain, making it important for the cardiovascular system to adapt (Seymour 1987; Lillywhite 1996). Other researchers have suggested that any differences in axial positioning of the heart evolved independently of habitat are therefore more related to phylogeny (Gartner et al. 2008).

Here, we give an overview of the literature and describe our own new data relating to the axial level of the heart in snakes.

10.2 Literature Overview

Badeer assumed the cardiovascular system to function as a siphon, such that the weight of the venous return column was opposite to the gravitational load of the blood pumped towards the head (Badeer 1985). These two loads were assumed to cancel each other out, rendering the position of the heart irrelevant to cardiovascular dynamics. However, this assumption may have overlooked the effects of vasoconstriction in caudal arteries and veins (Seymour and Johansen 1987).

Further circulatory adaptations to gravity in snakes of different habitats have been explored by Lillywhite and colleagues (Lillywhite and Pough 1983; Lillywhite and Gallagher 1985; Lillywhite 1987, 1988, 1993, 1996). These workers found that the heart in terrestrial and arboreal snakes is positioned more rostrally along the

body axis level compared to the heart in semi-aquatic and aquatic snakes; they also found that terrestrial and arboreal snakes are much better adapted to being tilted upright (Lillywhite 1987).

The baroreceptor reflex is one of the physiological mechanisms that might be important in this context. This reflex ensures that the cardiac output increases when the animal experiences low blood pressure in the carotid arteries (the main arteries supplying the brain). This reflex is most strongly developed in arboreal snakes, with the result that a higher blood pressure can be maintained at all times (Lillywhite and Gallagher 1985; Lillywhite 1996). In aquatic species the baroreceptor reflex is present, but is not capable of maintaining an adequate cerebral blood flow when the animal is fully upright. Slopes of 30 % have been described as being the maximum tilt in, for example, *Aipysurus laevis* (Lillywhite and Pough 1983). It is thought that the baroreceptor reflex in aquatic species became less effective because it became redundant in an underwater environment (Lillywhite and Pough 1983). Being upright under water does not create a lower cerebral blood pressure because the gravitational force is countered by the outside water pressure (Lillywhite 1996).

In snakes, no venous valves (such as those that prevent backflow of blood in mammals) have been observed. However, according to Lillywhite (1987), the following venous adaptations may prevent the pooling of blood in the lower body segments: (1) hairpin loops, present in many veins, reduce retrograde blood flow; (2) the corkscrew morphology of the portal vein in terrestrial snakes can trap blood depending on posture. This helical morphology of the portal vein is absent in aquatic species, suggesting that it plays a role in regulating blood pressure during postural changes; and (3) localised vasoconstriction, induced by catecholamines, create a bottleneck in the vessel preventing backflow.

Another strategy to prevent pooling of blood is the tighter integument which tree climbing snakes have developed. Their tails in particular, show tighter skin and less subcutaneous compliance compared to aquatic species (Lillywhite 1993, 1996).

Because an upright posture has hemodynamic effects on the lung vasculature, the pulmonary circulation has to adapt also to the increase in pulmonary blood pressure during tilting (Lillywhite 1988). An elongated lung suffers more from blood pooling in the capillaries than a short one. With an increase in blood pooling, more fluid will transfer out of the vessels into the interstitial space; this in turn creates oedema, impaired oxygen transfer into the blood and ultimately hypoxia (Lillywhite and Pough 1983; Lillywhite 1987). To prevent oedema from forming, the blood pressure in the lung circulation is usually lower than the systemic blood pressure. However, further adaptations may be necessary to keep the pulmonary blood pressure from becoming too high. It has been observed that terrestrial and arboreal snakes have shorter lungs than aquatic snakes; the latter tend to have a long vascular lung extending along almost the whole body length (Lillywhite 1987). After being tilted upright outside water, aquatic species may exhibit pulmonary blood clotting, swelling and disruption of pulmonary tissue, and death (Lillywhite and Pough 1983).

The elongated lung in aquatic species might have developed as an adaptation to underwater life (Lillywhite 1987). The long, unpaired lung is used as an air sac when diving, and allows for better buoyancy (Graham et al. 1987). However, it is

known that diving snakes spend a longer time under water than would be expected based on the oxygen content in their lung (Heatwole and Seymour 1975; Rubinoff et al. 1986). For many aquatic species, it has consequently been shown that they can take up to 33 % of their required oxygen from the water through their skin, the remainder of oxygen being absorbed from the lung reserve. These species also manage to release most of their carbon dioxide through the skin, thereby preventing respiratory acidosis (Graham 1974). This permits them to stay for up to 213 min under water (Rubinoff et al. 1986).

Locomotor patterns can influence blood pooling and cardiac output. Terrestrial snakes start moving their tails when their arterial pressure drops below 20 mm Hg, the lowest pressure at which snakes can function normally (Seymour and Lillywhite 1976; Lillywhite 1987). However, cardiovascular adaptation to tilting has mainly been tested with animals in a tight Perspex tube without space for movement; thus, the effect of movement may not have been fully determined.

Also, by forcing the animals to adopt a stretched vertical position, the natural curves of climbing animals are bypassed. When looking into climbing behaviour of snakes, it has been described that they tend to make use of two different types of movement (Astley and Jayne 2007, 2009). The first is a more energy-demanding movement called arboreal concertina movement, in which the body has always at least one static gripping point on the branch while the rest of the body follows one path. The second type of movement is preferred and is called lateral undulation, in which the body moves in a constant flow. However, for lateral undulation more grip is needed than for arboreal concertina movement. In both of these movements the body is never fully upright but always in a sinuous conformation.

As branches become thicker, the lower body moves to a more perpendicular plane to create additional grip with ventral flexion (Jayne and Herrmann 2011). Also, on smoother surfaces snakes switch to ventral flexion to create additional grip (Astley and Jayne 2007). In all the cases, the body is twisted and not straight, influencing the continuity of the blood column.

Furthermore, though arboreal snakes spend some time climbing, a lot of time is spent resting horizontally on a branch. The question that remains is how far these snakes will have had to adapt to the short periods of upright posture.

Gartner has reported evidence that for land-dwelling species, arboreal or terrestrial habitat had significantly less influence on axial heart level than their genetic background (Gartner et al. 2008). The trend he thereby also observed was that arboreal snakes tended to have a more caudally located heart compared to terrestrial specimens. This, according to him, was opposite to what Lillywhite and Seymour (2011) had previously reported.

Lillywhite and Seymour responded by saying they had never stated that arboreal snakes had a higher axial heart level compared to terrestrial snakes (Lillywhite and Seymour 2011). Indeed, in one of their first articles mentioning axial heart level, they described one arboreal snake specimen, *Boiga dendrophila,* and two terrestrial specimens, *Austrelaps superbus* and *Notechis scutatus*, having their hearts at 17, 17 and 16 % of their total body lengths, respectively (Seymour and Lillywhite 1976). In this article, they also looked at three semi-aquatic and three aquatic species and

found them to have the most distant hearts. Therefore, they only divided snakes into the land-dwelling and the water-dwelling group with regard to axial heart level (Seymour 1987).

However, in other articles by the same authors they mentioned that a high blood pressure seemed to be correlated with a higher axial heart level and that the arterial blood pressure is highest in arboreal snakes (Seymour 1987; Lillywhite 1996). Though not explicitly mentioned, the conclusion that arboreal species would therefore have a higher axial heart level was implied. This was further supported by their schematic overviews of snakes in upright positions in which all arboreal snakes are portrayed with their hearts at a higher axial level than the terrestrial animals (Lillywhite 1987, 1988, 1996). These facts might explain the conclusion drawn by Gartner.

Confirming Gartner's theory about the importance of phylogeny is the finding that Viperidae have the most caudally located hearts compared to other families, even though they do not show blood pooling during tilting experiments as aquatic species do (Lillywhite 1987; Gartner et al. 2008). Also it has been seen that within the Viperidae the arboreal species have their heart positioned more cranially along the primary body axis compared to their terrestrial counterparts (Seymour 1987).

In summary, there is evidence that arboreal snakes are better adapted to being in an upright position than aquatic snakes or even terrestrial species. However, whether the axial level of the heart is one of these adaptations is unclear. Currently, it seems possible that a mix of phylogenetic and behavioural factors cause the differences of axial heart level between snake species (Gartner et al. 2008; Lillywhite et al. 2012).

We have examined the axial level of the heart in adult snakes from different habitats to see whether we could confirm the trend observed by Gartner et al. (2008) or the link between arboreal species and a cranial axial heart level as suggested by Lillywhite (1987).

10.3 Materials and Methods

We recorded the axial level of the heart in 15 snakes from 7 different families (Table 10.1). Since the curled-up positions in which many of the animals were fixed did not allow for them to be stretched without causing damage, outside measurements were not reliable (Fig. 10.1). Therefore, we chose to take X-rays to allow us to count the vertebrae; we then used vertebral count to provide a proxy for distance along the body axis (Fig. 10.2).

To establish the axial level at which the heart was located in the body cavity without disturbing the organ connections, the specimens were opened along the ventral midline and radiopaque pins were inserted to mark the cranial extent of the heart, the atrioventricular junction and the apex of the heart. The cranial limit of the

Table 10.1 Specimens used to compare heart position in snakes

#	Species	Habitat	Family	Skull-heart	Heart-cloaca	Cloaca-tailtip	Heart-tailtip
12	*Acrochordus granulatus*	Aquatic	Acrochordidae	91	145	35	180
84	*Natrix tessellata*	Terrestrial + Aquatic	Colubridae	36	142	75	217
119	*Xenochrophis piscator*	Terrestrial + Aquatic	Colubridae	30	110	124	234
127	*Acanthophis rugosus*	Terrestrial	Elapidae	36	136	47	183
269	*Dendroaspis viridis*	Arboreal	Elapidae	61	156	‡90	‡246
272	*Dendroaspis viridis*	Arboreal	Elapidae	65	154	‡80	‡234
188	*Morelia viridis*	Arboreal	Pythonidae	56	‡192	‡77	‡269
274	*Morelia viridis*	Arboreal	Pythonidae	55	‡171	‡84	‡255
230	*Typhlops sp.*	underground	Typhlopidae	57	151	7	158
236	*Bothrops alternatus*	Terrestrial	Viperidae	58	169	53	222
242	*Calloselasma rhodostoma*	Terrestrial	Viperidae	50	102	47	149
246	*Crotalus artox*	Terrestrial	Viperidae	77	103	36	139
247	*Crotalus atrox*	Terrestrial	Viperidae	74	109	30	139
257	*Vipera Schweizeri*	Terrestrial	Viperidae	51	126	18	144
262	*Xenopeltis sp.*	Terrestrial	Xenopeltidae	49	133	31	164

Skull-heart: amount of vertebrae between cranium and top of the heart; Heart-Cloaca: amount of vertebrae between top of the heart and the cloaca; Cloaca-Tailtip: amount of vertebrae between the cloaca and the end of the tail; Heart-Tailtip: amount of vertebrae between top of the heart and the end of the tail. ‡ Estimated values because of insufficient resolution of the X-rays. Numbers refer to the alcohol collection catalogue of Leiden University

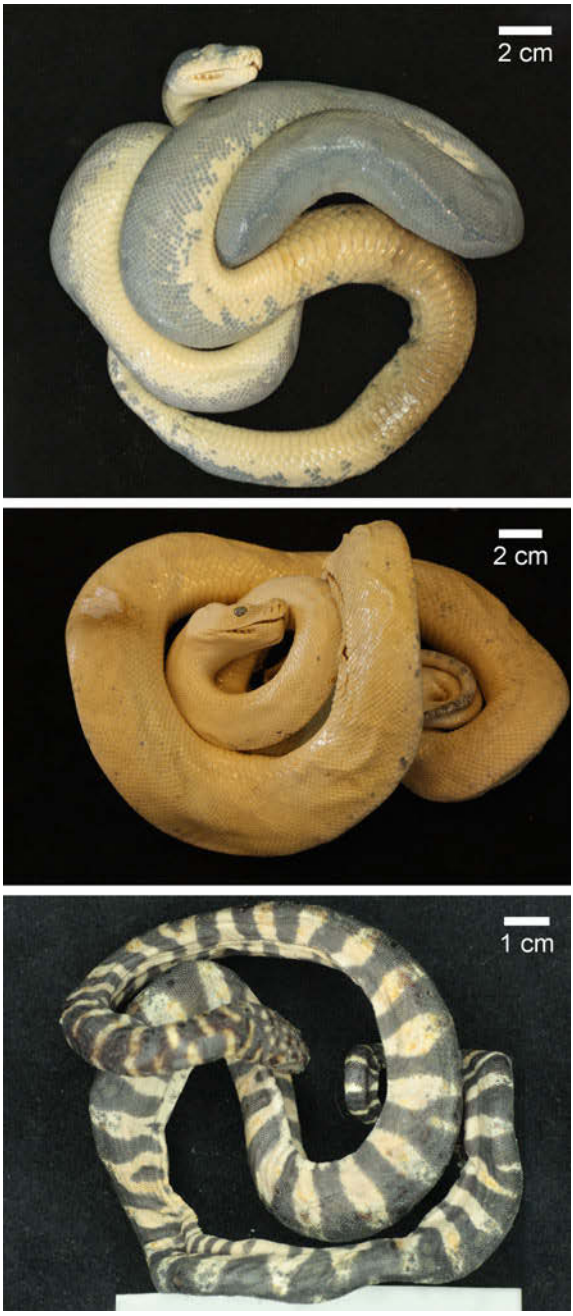


Fig. 10.1 Representative specimens. From *top* to *bottom Morelia viridis* #274, *Morelia viridis* #188, *Acrochordus granulatus* #12. The animals were too stiffened by fixation and storage to be extended without damage

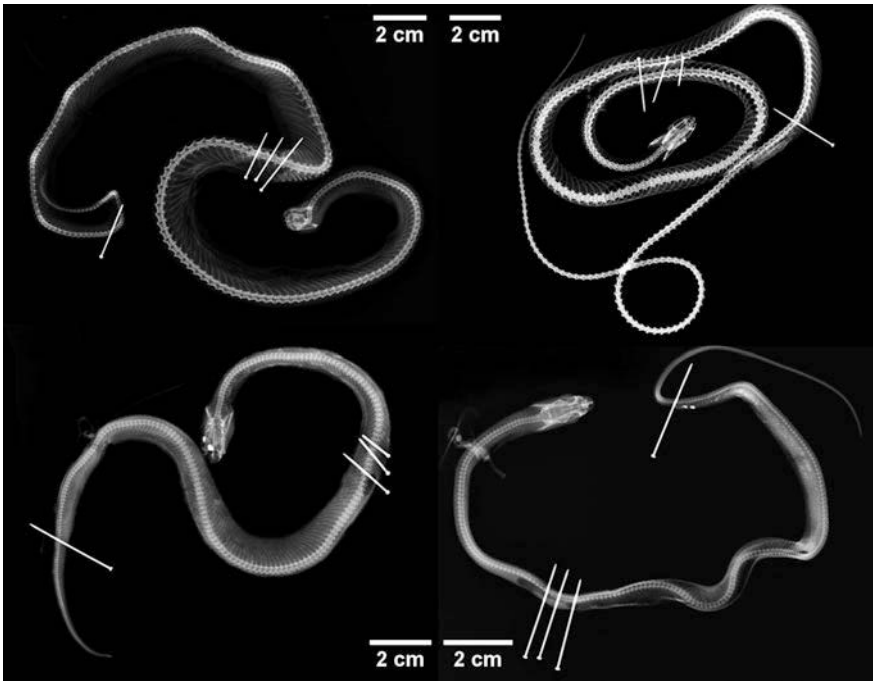


Fig. 10.2 X-rays of snakes. Clockwise from *top left*: *Acrochordus granulatus* #12 a fully aquatic species; *Xenochrophis piscator* #119 a semi-aquatic species, pins are located in the most cranial wreath; The arboreal *Dendroaspis viridis* #272, part of the label still visible; *Bothrops alternatus* #236 a terrestrial species. The three cranial pins indicate top of the heart, atrioventricular transition and apex of the heart respectively. The most caudal pin indicates tail base

heart was defined as the notch between the left and right aortae. The cloaca was also marked with a pin. X-rays were taken at the Naturalis Biodiversity Center in Leiden. All specimens were from the Institute of Biology, Leiden University, and had been stored in ethanol. No animals were sacrificed for the purpose of this study. To place our findings in a phylogenetic context, we used the phylogeny of Pincheira-Donoso et al. (2013), reproduced here as Fig. 10.3.

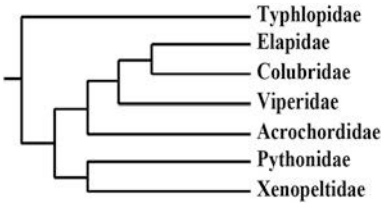


Fig. 10.3 Phylogeny of the snake taxa studied. Adapted from Pincheira-Donoso et al. (2013)

10.4 Results and Discussion

We found that vertebrae get smaller towards the tail. However, since we observed the same trend in all species in a similar manner we decided not to correct for differences in vertebral size. Since Gartner et al. (2008) measured axial level of the heart along the 'total' length from snout to cloaca, omitting the tail, we made a graphical representation of our results both with and without including tail length in the total (Figs. 10.4, 10.5). When considering the total body length including tail, snakes with a semi-aquatic lifestyle showed the shortest relative distance between skull and heart. This distance was 12–14 % of the total body length. In the arboreal snakes examined, the heart was located between in the range of 18–22 % body length, while the equivalent range in terrestrial snakes was greater, with measurements ranging from 17 to 37 %. The axial level of the heart in aquatic and subterrestrial species fell completely within the terrestrial scatter (Fig. 10.4). When we excluded the tail from the total body length, we still found all arboreal snakes fitting within the terrestrial range (Fig. 10.5). When analysing the results by phylogeny, we saw a wide scatter in the Viperidae. The Colubridae and Pythonidae showed the most cranial axial level of the heart (Fig. 10.6).

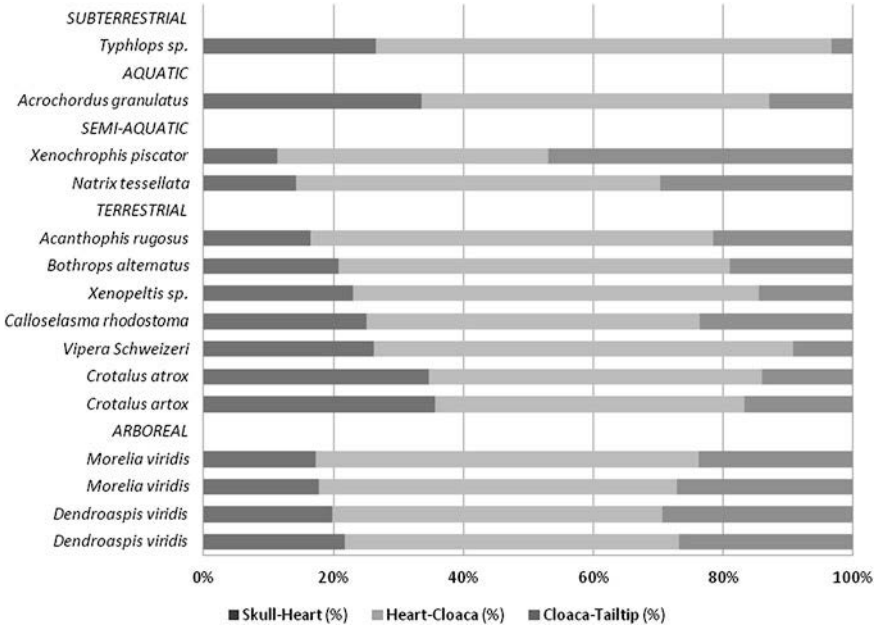


Fig. 10.4 Heart position in snakes with tail length included in the analysis. Key: Skull-heart: percentage of number of vertebrae between the skull base and the top level of the heart; Heart-cloaca: percentage of number of vertebrae between the top of the heart and the cloaca; Cloaca-tail: percentage of number of vertebrae between the cloaca and the tail tip. Tail lengths in the arboreal species are estimated values

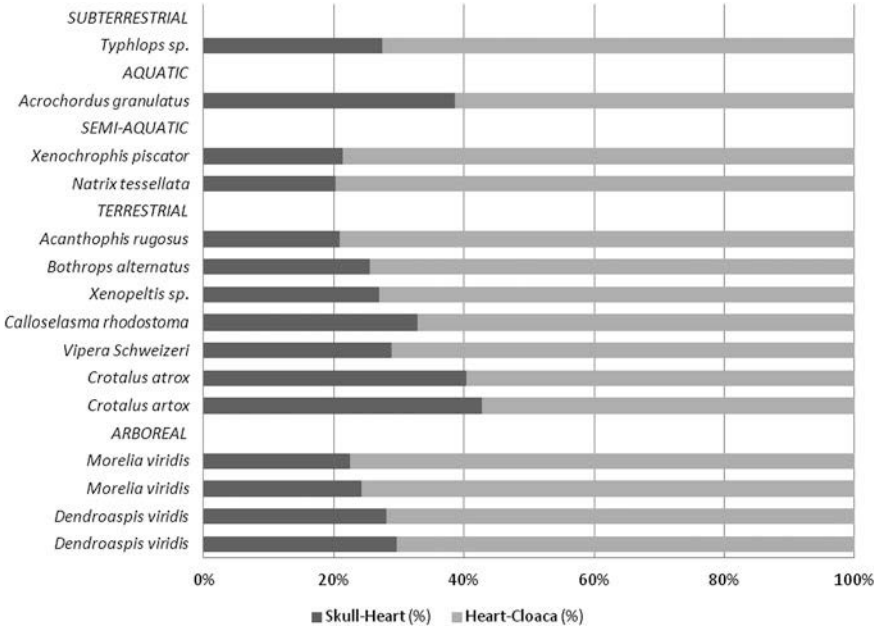


Fig. 10.5 Heart position with tail length excluded from the analysis. All animals are subdivided by habitat. Key: Skull-heart: percentage of number of vertebrae between the skull base and the top level of the heart; Heart-cloaca: percentage of number of vertebrae between the top of the heart and the cloaca

We could not confirm the observation that arboreal snakes have their hearts closest to the head (Lillywhite 1987). From our results, it seems that semi-aquatic species such as *Natrix tesselata* and *Xenochrophis piscator*, both members of the Colubridae, have their hearts closest to their cranium and that the arboreal species we examined fall within the range of the terrestrial species. This is also in contrast to the findings of Gartner who stated that arboreal snakes in general tend to have a more caudally located heart when excluding the tail from measurements of total body length (Gartner et al. 2008). However, regardless of whether the tail was included, we found no difference between the axial level of the heart in terrestrial versus arboreal snakes.

When we considered the species with regard to their phylogeny, the Acrochordidae and Viperidae had species with the most caudal positioning of the heart, in agreement with the findings of Gartner et al. (2008) (Fig. 10.6). However, it should be mentioned that *Bothrops alternatus*, also a viper, had a heart positioned very cranially compared to, for instance, *Crotalus artrox*.

We are therefore unable to confirm the conclusion that phylogeny plays a major role in establishing the axial level of the heart (Gartner et al. 2008). Furthermore, something that we were mostly unable to do, except for *Morelia viridis* and

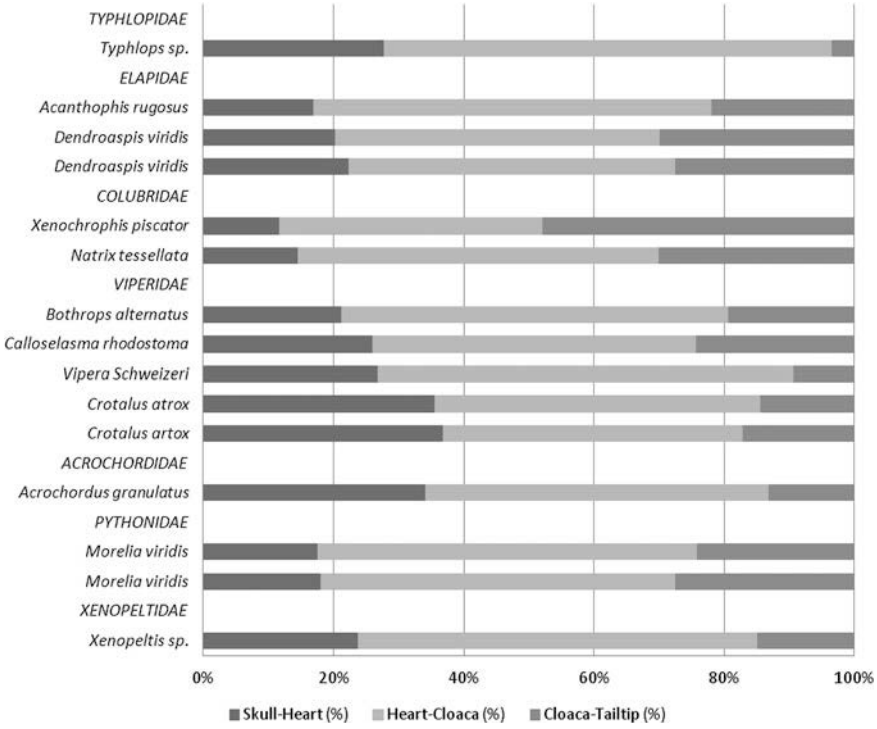


Fig. 10.6 Heart position according to phylogeny. Compare with Fig. 10.3 to see evolutionary relation. Key: Skull-heart: percentage of number of vertebrae between the skull base and the top level of the heart; Heart-cloaca: percentage of number of vertebrae between top of the heart and the cloaca; Cloaca-tail: percentage of number of vertebrae between the cloaca and the tail tip. Tail lengths in the arboreal species are estimated values

Dendroaspis viridis, was to have replicates for each species. It is possible that, when a snake grows older, shifts might occur in the axial position of the heart because of differential growth (Van Soldt et al. 2015). Because of the unknown age of our specimens, and because the two *Morelia viridis* and *Dendroaspis viridis* were of roughly the same size, we cannot exclude such a phenomenon.

The aquatic *Acrochordus granulatus* had a heart positioned relatively caudally, consistent with previous studies (Kashyap and Sohoni 1973; Seymour and Lillywhite 1976). It could be that this morphology too results from its aquatic lifestyle. The weight of the heart muscle itself, by being close to the body centre, might improve weight distribution during swimming or diving. Also, the close relation between pulmonary and circulatory organs might explain why this aquatic snake shows such a caudal heart positioning, since the aquatic snake lung is so elongated (Lillywhite 1987). More research into aquatic species should be done to determine the effect of the aquatic lifestyle on the axial position of the heart.

10.5 Conclusions

We conclude that there are still several possible scenarios to explain the differences found in the axial level of the heart in snakes. Among these are not only the arboreal habitat of some snakes and the gravitational demands of that habit, but also an aquatic lifestyle, as well as individual changes during ontogeny, phylogeny and the close relationship of the cardiovascular system with other organ systems.

Author Contributions and Acknowledgments Dissections were performed by JWF. X-rays were made by EMD in collaboration with Naturalis Biodiversity Center. JWF, REP and MKR devised the project and wrote the manuscript.

References

Astley HC, Jayne BC (2007) Effects of perch diameter and incline on the kinematics, performance and modes of arboreal locomotion of corn snakes (Elaphe guttata). J Exp Biol 210:3862–3872

Astley H, Jayne BC (2009) Arboreal habitat structure affects the performance and modes of locomotion of corn snakes (Elaphe guttata). J Exp Zool Part A Ecol Genet Physiol 311:207–216

Badeer H (1985) Elementary hemodynamic principles based on modified Bernoulli's equation. Physiologist 28:41–46

Gartner GE, Hicks JW, Manzani PR, Andrade DV, Abe AS, Wang T, Secor SM, Garland T (2008) Phylogeny, ecology, and heart position in snakes. Physiol Biochem Zool 83:43–54

Graham JB (1974) Aquatic respiration in the sea snake Pelamis platurus. Respir Physiol 21:1–7

Graham JB, Gee JH, Motta J, Rubinoff I (1987) Subsurface buoyancy regulation by the sea snake Pelamis platurus. 60:251–261

Heatwole H, Seymour R (1975) Pulmonary and cutaneous oxygen uptake in sea snakes and a file snake. Comp Biochem Physiol 51:399–405

Jayne BC, Herrmann MP (2011) Perch size and structure have species-dependent effects on the arboreal locomotion of rat snakes and boa constrictors. J Exp Biol 214:2189–2201

Kashyap H, Sohoni P (1973) The heart and arterial system of Acrochordus granulatus (Schneider). J univ Bombay 42:34–52

Lillywhite HB (1987) Circulatory adaptations of snakes to gravity. Am Zool 27:81–95

Lillywhite HB (1988) Snakes, blood circulation and gravity. Sci Am 259:92–98

Lillywhite HB (1993) Subcutaneous compliance and gravitational adaptation in snakes. J Exp Zool 267:557–562

Lillywhite HB (1996) Gravity, blood circulation, and the adaptation of form and function in lower vertebrates. J Exp Zool 275:217–225

Lillywhite HB, Gallagher KP (1985) Hemodynamic adjustments to head-up posture in the partly arboreal snake, Elaphe-Obsoleta. J Exp Zool 235:325–334

Lillywhite HB, Pough FH (1983) Control of arterial pressure in aquatic sea snakes. Am J Physiol 244:66–73

Lillywhite HB, Seymour RS (2011) Heart position in snakes: response to "phylogeny, ecology, and heart position in snakes". Physiol Biochem Zool 84:99–101

Lillywhite HB, Albert JS, Sheehy CM, Seymour RS (2012) Gravity and the evolution of cardiopulmonary morphology in snakes. Comp Biochem Physiol A: Mol Integr Physiol 161:230–242

Murphy JC (2012) Marine invasions by non-sea snakes, with thoughts on terrestrial-aquatic-marine transitions. Integr Comp Biol 52:217–226

Neill WT (1964) Viviparity in snakes: some ecological and zoogeographical considerations. 48:35–55

Pincheira-Donoso D, Bauer AM, Meiri S, Uetz P (2013) Global taxonomic diversity of living reptiles. PLoS ONE 8:e59741

Rubinoff I, Graham JB, Motta J (1986) Diving of the sea snake Pelamis platurus in the Gulf of Panamá. Mar Biol 91:181–191

Seymour RS (1987) Scaling of cardiovascular physiology in snakes. Integr Comp Biol 27:97–109

Seymour R, Johansen K (1987) Blood flow uphill and downhill: does a siphon facilitate circulation above the heart? Comp Biochem Physiol 88:167–170

Seymour R, Lillywhite H (1976) Blood pressure in snakes from different habitats. Nature 264:644–646

Van Soldt BJ, Metscher BD, Poelmann RE, Vervust B, Vonk FJ, Müller GB, Richardson MK (2015) Heterochrony and early left-right asymmetry in the development of the cardiorespiratory system of snakes. PLoS ONE 10:e116416

Chapter 11
On the Neo-Sex Chromosomes of Lepidoptera

Petr Nguyen and Leonela Carabajal Paladino

Abstract Chromosome rearrangements can play an important role in adaptive evolution and speciation with gene flow. Here, we briefly review state of the art in chromosomal speciation, along with the classic model of sex chromosome evolution. The main focus lies on sex chromosome–autosome fusions, i.e., neo-sex chromosomes. We describe the presence of neo-sex chromosomes in moth and butterflies (Lepidoptera), the largest group with female heterogamety. Despite the relative stability of lepidopteran karyotypes, fusions which result either in multiple sex chromosomes (W_1W_2Z or WZ_1Z_2) or large sex chromosome pairs occurred at a surprisingly high frequency throughout their evolution. We discuss the role of meiotic drive, genetic drift, and selection in the establishing of these derived sex chromosome systems. It is hypothesized that the association between sex-linked reproductive isolation or female preference and larval performance may contribute to ecological specialization and species formation in Lepidoptera.

11.1 Chromosomal Speciation

Observations that even closely related species usually differ in their karyotypes, i.e., chromosome number and morphology, have been feeding long-standing controversy over the role of chromosome rearrangements in species formation (White 1973; Faria and Navarro 2010). A correlation between rates of chromosome rearrangements and speciation reported in some taxa (Bush et al. 1977; Olmo 2005) added even more fuel to the flames.

P. Nguyen (✉) · L. Carabajal Paladino
Biology Centre CAS, Institute of Entomology, Branišovská 31,
37005 České Budějovice, Czech Republic
e-mail: petr.nguyen@prf.jcu.cz

P. Nguyen
Faculty of Science, University of South Bohemia in České Budějovice,
Branišovská 1760, 37005 České Budějovice, Czech Republic

P. Pontarotti (ed.), *Evolutionary Biology*, DOI 10.1007/978-3-319-41324-2_11

First models of chromosomal speciation assumed that karyotype change can constitute a reproductive barrier via its underdominance, i.e., deleterious effect in hybrids heterozygous for the rearrangement (White 1973; Baker and Bickham 1986). A main flaw of these models lies in the fact that underdominant rearrangements strong enough to act as a barrier are unlikely to spread and get fixed in a population. The problem was alleviated by the hybrid–sterility monobrachial fusion model, which proposes that a strong sterility barrier can be formed upon secondary contact if independent centric fusions with little or no fitness costs get fixed in isolated populations (Baker and Bickham 1986). This concept, however, applies only to the so-called Robertsonian fusions of acrocentric chromosomes. Furthermore, it is clear that at least some organisms can cope with the segregation of complex multivalents observed in hybrids of different karyomorphs (Sharp and Rowell 2007; Potter et al. 2015).

Speciation has long been viewed as a result of random processes requiring geographic isolation (Bird et al. 2012). However, this perspective has recently changed and the evolution of reproductive barriers without spatial isolation is now considered to be a plausible scenario (Smadja and Butlin 2011). In its initial phase, speciation with gene flow is supposedly characterized by strong disruptive selection on locally adapted alleles resulting in genomic islands of divergence, which tend to spread by means of genetic hitchhiking. Selection is opposed by gene flow, which homogenizes variation in the remaining genomic regions. New theoretical models on chromosomal speciation postulate that recombination modifiers would be favored by selection under speciation with gene flow (Ortiz-Barrientos et al. 2016). These models predict that restriction of recombination promotes adaptive divergence by sheltering associations of adapted alleles (Ayala et al. 2014; Lohse et al. 2015; Roesti et al. 2015). Compared to genic modifiers, structural modifiers of recombination, i.e., chromosome rearrangements, affect the recombination rate by altering linkage relationships (Yeaman 2013; Ortiz-Barrientos et al. 2016) and are more likely to spread in a population as they are linked to selected loci. Furthermore, they avoid accumulation of deleterious mutations due to Muller's ratchet by suppressing recombination only in heterozygotes (Felsenstein 1974).

These theoretical models have recently been extended from inversions to other rearrangements such as fusions and transpositions, which can tighten linkage between adaptive loci (Ortiz-Barrientos et al. 2016) and couple previously unlinked genes (Yeaman 2013; Guerrero and Kirkpatrick 2014).

11.2 Sex Chromosome–Autosome Fusions

Several authors have recently noted that there are similarities between evolution of sex and formation of species with gene flow (Yeaman 2013; Ortiz-Barrientos et al. 2016). Under the canonical model, sex chromosomes evolved from a pair of autosomes in which one homologue acquired a sex-determining factor (Ohno 1967; Charlesworth and Charlesworth 1978; Charlesworth 2013). Once the separate sexes

are established, selection favors accumulation of sexually antagonistic alleles, i.e., those beneficial to one sex but deleterious to the other, in the vicinity of the sex-determining region. To maintain their linkage disequilibrium, recombination between the newly arisen sex chromosomes is suppressed in the heterogametic sex (Rice 1987; but cf. Ironside 2010) either by means of inversions (Charlesworth et al. 2005) or heterochiasmy, i.e., sex-specific differences in recombination (Perrin 2009). Lack of recombination leads to molecular degeneration of sex-limited chromosomes Y (in male heterogametic organisms) or W (in female heterogametic taxa) resulting in a loss of coding sequences and accumulation of repetitive sequences, which are then silenced epigenetically by changes in chromatin structure (Charlesworth et al. 2005; Bachtrog 2013).

There is a plethora of ancient sex chromosome systems which evidence this final step of sex chromosome differentiation (e.g., Skaletsky et al. 2003; Vicoso and Bachtrog 2013), yet there is much to learn about early stages of sex chromosome evolution. To that end, evolutionary young sex chromosomes formed either de novo or by sex chromosome–autosome rearrangements, i.e., neo-sex chromosomes, have been investigated (e.g., Bachtrog 2013; Yoshida et al. 2014). For example, independent fusions gave rise to neo-sex chromosomes of different age and level of differentiation in *Drosophila* fruit flies. Hence, detailed genomic analyses of neo-sex chromosomes in *D. albomicans*, *D. miranda*, and *D. pseudoobscura*, which are about 0.1 My, 1 My, and 15 My old, respectively, provide an exceptional insight into molecular evolution of sex-limited loci over evolutionary time (reviewed in Bachtrog 2013). An even younger neo-sex chromosome has been reported in *D. americana*, in which the rearrangement is polymorphic and its frequency changes along a latitudinal transect. The *D. americana* neo-sex chromosome is favored by selection and its frequency correlates with winter severity and presumably contributes to local adaptation to different climatic conditions (McAllister et al. 2008).

Recent studies also showed that sex chromosome–autosome fusions could play a role in formation of new species. In Japanese threespine stickleback fish *Gasterosteus aculeatus*, the fusion brought together loci responsible for behavioral isolation and hybrid sterility (Kitano et al. 2009). These observations prompted interest in forces driving neo-sex chromosome evolution. It has been argued that some chromosomes are predisposed to a specialized role in sex determination (Ross et al. 2009; O'Meally et al. 2012). Accordingly, it was suggested that gene content of autosomes involved in fusions with sex chromosomes can be relevant for maintenance of resulting neo-sex chromosomes in fish and birds (Kitano et al. 2009; Ross et al. 2009; Pala et al. 2012; Kitano and Peichel 2012; Yoshida et al. 2014).

In theory, neo-sex chromosomes can become fixed in a population by random genetic drift (Lande 1985) or selection, which can favor either heterozygosity in inbred populations (Charlesworth and Wall 1999) or linkage of sexually antagonistic loci to sex chromosomes (Charlesworth and Charlesworth 1980). The latter has recently been topped up by a model describing how a newly formed neo-sex chromosome could further expand into the species range. It was shown that under

realistic conditions selection against sexually antagonistic effects of the neo-Y chromosome indirectly promotes neo-X in the hybrid zone, which in turn shields the neo-Y from selection and allows it to spread (Veltsos et al. 2008).

Moreover, comparative analyses of available data on the presence of multiple sex chromosomes in vertebrates revealed that sex chromosome–autosome fusions are much more frequent in male heterogametic taxa (XX/XY) than in female heterogametic taxa (WZ/ZZ) (Pokorná et al. 2014; Pennell et al. 2015). Female meiotic drive supposedly propels chromosome evolution via nonrandom segregation of trivalents formed by fused chromosomes in heterozygotes due to asymmetry of female meiosis and polarity of the meiotic spindle (de Villena and Sapienza 2001). It has been suggested that female meiotic drive also plays an important role in the evolution of neo-sex chromosomes. In mammals, prevalence of XY_1Y_2 and X_1X_2Y sex chromosome systems coincide with the preponderance in the karyotype of biarmed and acrocentric chromosomes, respectively, as expected under female meiotic drive (Yoshida and Kitano 2012). Female meiotic drive can also explain the paucity of neo-sex chromosomes in female heterogametic taxa. Pokorná et al. (2014) hypothesized that nonrandom segregation of multivalent, comprising Z and W chromosomes in female meiosis, would result in a distorted sex ratio and could be selected against. Other factors such as underdominace of fusions or biased reproductive sex ratios can further contribute to the observed pattern (Pennell et al. 2015).

11.3 Neo-Sex Chromosomes in Lepidoptera

Moths and butterflies (Lepidoptera) represent the largest group of organisms with a female heterogametic sex determination system (Traut et al. 2007; Marec et al. 2010). Although Lepidoptera comprises about 160,000 species (van Nieukerken et al. 2011), data on their sex chromosomes remains scarce despite considerable effort. The reasons are clear, lepidopteran mitotic complements consist of a high number of small and morphologically uniform elements (Robinson 1971). That, along with the absence of any applicable banding technique (De Prins and Saitoh 2003) makes identification of sex chromosomes in lepidopteran metaphase complements nearly impossible (cf. Fontana 1976) without the use of molecular cytogenetic techniques (Fuková et al. 2005; Yoshido et al. 2005; Vítková et al. 2007). Hence, Traut et al. (2007) included merely 40 species in their list of known karyotypes with identified sex chromosomes. The list is far from being complete (see, e.g., Suomalainen 1969), yet it contains 12 species with multiple sex chromosomes. These correspond to seven independent origins (cf. Pokorná et al. 2014) of neo-sex chromosomes in Lepidoptera. Many more neo-sex chromosome systems have recently been reported (Nguyen et al. 2013; Šíchová et al. 2013, 2015, 2016; Walters and Mongue 2016; Carabajal Paladino et al. (2016) and others are waiting for detailed analyses (e.g., Suomalainen 1969; Maeki 1981). Therefore, available data in Lepidoptera points to a relatively high incidence of neo-sex chromosomes,

which is comparable to rates observed in male heterogametic taxa such as mammals and reptiles (Pennell et al. 2015).

Our knowledge on the evolution of lepidopteran sex chromosomes largely stems from the female-specific sex chromatin, i.e., heterochromatic body observed in female interphase nuclei, which corresponds to the W chromosome and can thus be used as indirect evidence of its presence (Ennis 1976; Traut and Marec 1996). Systematic screening for the presence of sex chromatin revealed that basal lepidopteran lineages and caddisflies share the sex chromosome constitution Z0/ZZ. The W chromosome originated later in the evolution of Lepidoptera, in the common ancestor of the Tischerioidea–Ditrysia clade encompassing 95 % of all extant species (Traut and Marec 1996; Lukhtanov 2000; Marec et al. 2010).

The neo-sex chromosomes observed in Lepidoptera are of various ages. Remarkably, one of the hypotheses on the evolutionary origin of the W chromosome postulates that it actually derived from an autosome, which became female-limited upon a fusion of its homologue with the ancestral Z chromosome (Traut and Marec 1996; Lukhtanov 2000; Marec et al. 2010). If confirmed, this pair of neo-sex chromosomes would predate the major diversification of moths and butterflies, and it seems it would not be the only case.

In his paper published in 1969, Suomalainen reported one of the first multiple sex chromosome systems observed in female heterogametic taxa. Among others, he described a sex chromosome trivalent W_1W_2Z in two moths of the family Tortricidae, *Bactra furfurana* and *B. lacteana*. Other members of this family are characterized by the presence of two conspicuously large elements that correspond to a sex chromosome pair which most likely arose from a sex chromosome–autosome fusion (Suomalainen 1971; Ennis 1976). This large chromosome pair has been observed in all tortricids karyotyped so far (for a comprehensive list of karyotyped species, see Šíchová et al. 2013) and was subjected to detailed molecular analyses (Fuková et al. 2005; Šíchová et al. 2013). Nguyen et al. (2013) performed comparative physical mapping of the Z chromosome in the codling moth, *Cydia pomonella*, to test a hypothesis on the fusion of the original sex chromosomes with an autosomal pair corresponding to chromosome 15 from *Bombyx mori* (Bombycidae) reference genome (Heckel et al. 1998). The fusion was confirmed not only in the codling moth but also in the other tortricids examined, which suggests that it occurred in a common ancestor of the Tortricinae and Olethreutinae subfamilies, which comprise nearly 10,000 species representing 97 % of extant tortricids (Nguyen et al. 2013).

In addition, a similar noticeably large chromosome pair has repeatedly been reported in several families within the mega-diverse superfamily Gelechioidea (Ennis 1976; Kawazoé 1987; Lukhtanov and Puplesiene 1999). Recently, this large chromosome pair has been proven to be the sex chromosomes in the tomato leafminer *Tuta absoluta* (Gelechiidae) (Carabajal Paladino et al. 2016). A detailed analysis of neo-sex chromosomes in Gelechioidea is under way (Carabajal Paladino LZ, Nguyen P, unpublished) and preliminary results confirm that neo-sex chromosomes are present in all three gelechioid clades (sensu Sohn et al. 2015) encompassing almost 18,000 described species (van Nieukerken et al. 2011).

Neo-sex chromosomes have recently been described also in monarch butterflies, *Danaus plexippus* (Danaidae), as well as in several of its closely related congeners. Using genomic data, Walters and Mongue (2016) found that the Z chromosome fused with an autosome corresponding to the *B. mori* chromosome 16 at least in a common ancestor of the genus *Danaus* as suggested by a shared karyotype of 2n = 60 (Brown et al. 2004). As noted by the authors, given the fact that *Danaus* shares the chromosome number also with representatives of two basal danaini groups, namely *Lycorea* spp. (Itunina) (Brown et al. 2004), *Idea* sp. (Maeki and Ae 1969), and *Euploea* spp. (Euploeina) (Saitoh and Abe 1970), it is tempting to speculate that the neo-sex chromosomes originated earlier in the evolution of the group. A comparative study of synteny of sex-linked genes within Danainae is needed to test this hypothesis.

Yet another interesting case of neo-sex chromosomes is represented by ermine moths of the genus *Yponomeuta* (Yponomeutidae). The genus with 76 species (Turner et al. 2010) belongs to the superfamily Yponomeutoidea which represents one of the earliest ditrysian radiations (Sohn et al. 2013). As such, the ermine moths have been a model system for investigation into the evolution of insect–plant interactions (Menken et al. 1992). Nilsson et al. (1988) examined karyotypes of six representatives of a recently diversified European clade and found a derived sex chromosome system Z_1Z_2neo-W, which probably evolved by fusion of the ancestral W chromosome with an autosome. The evolutionary origin of the multiple sex chromosomes and their potential contribution to ecological specialization of the ermine moths is currently being investigated (Nguyen P, unpublished).

It is remarkable that in the above-mentioned systems, sex chromosome–autosome fusions were the very first rearrangements which differentiated karyotypes of given groups from the ancestral chromosome print 2n = 62 (cf. Nilsson et al. 1988; Lukhtanov and Puplesiene 1999; Šíchová et al. 2013; Walters and Mongue 2016) preserved in Lepidoptera for more than 140 My (Ahola et al. 2014). However, not all systems are so stable. Chromosome fusions have been studied in tussock moths of the genus *Orgyia* (Lymantriidae), where different species vary widely in their chromosome number from the ancestral karyotype (Traut and Marec 1997). Neo-sex chromosomes of *O. antiqua* (2n = 28) and *O. thyellina* (2n = 22–23) were examined in detail by molecular cytogenetic techniques. These revealed that in *O. antiqua*, the ancestral sex chromosome pair fused with one autosome, while in *O. thyellina* it fused with two autosomal pairs. The analysis also showed that the neo-sex chromosomes differ between populations of *O. thyellina* by putative fission being neo-Wneo-Z in one and W_1W_2neo-Z in the other population (Yoshido et al. 2005). Even more detailed studies have been performed in the species complex of the wild silkmoths of the genus *Samia* (Saturniidae). The type species of this genus, *Samia cynthia*, can be divided into geographic subspecies showing a unique karyotype polymorphism with variable sex chromosome systems derived from ancestral-like karyotype 2n = 28 WZ/ZZ. Physical mapping combined with W chromosome painting in the *S. cynthia* subspecies allowed to reconstruct step by

step the evolution of their sex chromosomes, that involved several sex chromosome–autosome fusions and independent losses of the W chromosome (Yoshido et al. 2005, 2011, 2013).

Considering the high overall stability of lepidopteran genomes, wood white butterflies of the genus *Leptidea* (Pieridae) stand out for their dynamic karyotype evolution. Besides considerable inter- and intraspecific variability in chromosome numbers mainly due to fissions, which resulted in up to 2n = 208 chromosomes in *L. duponcheli* (Šíchová et al. 2015, 2016), the wood whites feature remarkably complex neo-sex chromosome systems. Multiple sex chromosome systems consist of $W_{1\text{-}4}Z_{1\text{-}4}$ in *L. reali*, $W_{1\text{-}3}Z_{1\text{-}4}$ in *L. juvernica*, $W_{1\text{-}3}Z_{1\text{-}3}$ in *L. sinapis*, and $W_{1\text{-}3}Z_{1\text{-}6}$ in *L. amurensis* and probably arose by a series of sex chromosome–autosome rearrangements (Šíchová et al. 2015, 2016). The origin of these complex chromosome associations remains unclear. It is worth mentioning that the *Leptidea* spp. karyotypes are highly variable regarding chromosome number and distribution of chromosome markers even within the progeny of individual females. Yet, interestingly, sex chromosome systems are stable and multivalents segregate evenly in meiosis (Šíchová et al. 2015, 2016). This within-species stability can point to a role for multiple sex chromosomes in the formation of reproductive barriers in *Leptidea* wood whites or some other evolutionary constraints. Indeed, it was hypothesized that the largest chromosomes observed in species with high chromosome number, such as blue butterflies, represent sex chromosomes, which suggests that sex chromosomes could resist to fissions (Robinson 1971; Ennis 1976).

The above overview of neo-sex chromosomes in Lepidoptera is not meant to be exhaustive. More cases of putative neo-sex chromosomes have been reported in moths and butterflies (see e.g., Suomalainen 1969; Maeki 1981). These examples are, however, rather anecdotal and await detailed investigation, which is deeply encouraged.

11.4 Drivers of Neo-Sex Chromosome Evolution in Lepidoptera

What are the driving forces behind the high incidence of neo-sex chromosomes in Lepidoptera, which involve both W and Z chromosomes at a roughly similar frequency (cf. Traut et al. 2007)? Pennell et al. (2015) suggested that repetitive content of sex chromosomes could make them more prone to ectopic recombination responsible for rearrangements. It has been shown that in birds the accumulation of repeats is not confined to the sex-limited W chromosome, but occurs also in the Z chromosomes, probably due to reduced recombination (Bellott et al. 2010). Indeed, genome comparisons in Lepidoptera revealed that short chromosomes, which have a higher percentage of repetitive elements have higher inter- and intrachromosomal

rearrangement rates, and were independently involved in fusions in distant species (Ahola et al. 2014). Comparative mapping in the peppered moth *Biston betularia* (Geometridae) suggests that, compared to autosomes, the Z chromosome experienced a higher rate of intrachromosomal rearrangements, which indicates higher content of repeats (Van't Hof et al. 2013). So it is reasonable to assume that sex chromosome–autosome fusions can be facilitated by repetitive sequences accumulated on both sex chromosomes in Lepidoptera.

But how do the neo-sex chromosomes spread and get fixed? It is difficult to reconcile the prevalence of derived sex chromosome systems observed in moths and butterflies with the role that female meiotic drive supposedly plays in the evolution of neo-sex chromosomes. Nonrandom segregation under female meiotic drive relies on selfish centromeres that compete for preferential transmission to the egg. The difference in size between centromeres and their corresponding kinetochore governs the number of captured microtubules and ultimately leads to preferential segregation to the egg pole (de Villena and Sapienza 2001). However, lepidopteran chromosomes are holokinetic, i.e., they lack localized centromere, and microtubules attach along the length of the chromosomes (Murakami and Imai 1974). It was noted that holokinetic chromosome organization could influence the strength of female meiotic drive (Pokorná et al. 2014), which has been recently extended to holokinetic chromosomes giving rise to a so-called holokinetic drive hypothesis. Holokinetic drive should propel changes in chromosome number via fusions and fissions as well (Bureš and Zedek 2014). However, since the role of centromeric repeats is substituted by the entire chromosome surface in holokinetics, a decrease in chromosome number should be accompanied by an expansion of mobile elements resulting in an increase in genome size and vice versa (Bureš and Zedek 2014). Whether holokinetic drive affects segregation of chromosomes in moths and butterflies has yet to be properly tested. Nevertheless, comparison of interphase nuclei sizes, which can be used as a proxy for genome size, and chromosome numbers in the *Leptidea* wood white butterflies does not corroborate the holokinetic drive hypothesis (Šíchová et al. 2015). Further doubts are cast by an often overlooked nuance in lepidopteran kinetic organization. Ultrastructural studies of kinetochores in moths showed that microtubules attach to a kinetochore plate which covers only 30–70 % of the chromosome surface (Gassner and Klementson 1974; Wolf 1994). Since sister chromatids separate by parallel disjunction in anaphase (Murakami and Imai 1974), chromosomes of Lepidoptera should be regarded as functionally, but not fully, holokinetic. The fact that sequences responsible for kinetochore formation have not expanded along the entire chromosome in the course of evolution challenges the ubiquity of female meiotic drive, and its role as a driver of chromosomal evolution in Lepidoptera.

As noted above, sex chromosome–autosome fusions are often the first rearrangements which distinguish karyotypes of given taxa from the otherwise highly conserved ancestral genome architecture. Sex chromosomes differ from autosomes in their effective population size. Given an equal number of breeding males and females, the effective population size of the Z chromosome (N_eZ) is only three quarters compare to that of autosomes (N_eA) and further drops because of sexual

selection acting on males, decreasing the number of males involved in reproduction (Mank et al. 2010). The smaller effective population size enhances the effect of genetic drift, which seems to be responsible for a higher rate of nonsynonymous substitutions in Z-linked loci in female heterogametic taxa such as birds and snakes (Mank et al. 2010; Vicoso et al. 2013). As already mentioned, a chromosomal rearrangement can get fixed in a population by genetic drift. However, unlike in birds and snakes, accelerated rates of molecular evolution of Z-linked genes were suggested to be driven by positive selection, and not genetic drift, in the silkmoth *Bombyx huttoni* (Bombycidae) (Sackton et al. 2014). The relative effective population size of Z chromosome to autosome (N_eZ: N_eA) observed in silkmoths was higher than that in birds and snakes, but still in a range where genetic drift should predominate. However, it was argued that drift is mitigated in silkworms by the overall effective population size (N_e), which can differ from birds by orders of magnitude (Sackton et al. 2014). Inference of demographic history is needed to test whether there were any bottleneck events associated with the described sex chromosome–autosome fusions in Lepidoptera.

If both drive and drift are dismissed, the only evolutionary force left is selection. The Z chromosome of moths and butterflies plays a special role in speciation as it is involved in both postzygotic and prezygotic reproductive isolation. It is associated with preferential sterility and inviability of female hybrids (reviewed in Presgraves 2002) and show significantly higher divergence compared with autosomes (Martin et al. 2013). Sex-linked genes also disproportionately contribute to behavioral and ecological differences between closely related lepidopteran species (reviewed in Sperling 1994; Prowell 1998), such as female host preference (Thompson 1988; Nygren et al. 2006). It was proposed that linkage between isolation or preference genes and performance genes affecting growth and survival on host plants can facilitate ecological specialization and speciation in phytophagous insects (Matsubayashi et al. 2010). Thus, sex chromosome–autosome fusions identified in Lepidoptera might have resulted in beneficial association of selected traits and promoted host shifts and speciation. That could be particularly true for tortricid moths where the Z chromosome fused with an autosome corresponding to chromosome 15 of *B. mori* (Nguyen et al. 2013). This autosome bears clusters of detoxification genes, namely carboxylesterases and ATP-binding cassette transporters (e.g., Yu et al. 2009; Xie et al. 2012), involved in metabolism and regulated absorption of plant secondary metabolites (Li et al. 2007; Zangerl et al. 2012; Dermauw et al. 2013).

Furthermore, Nguyen et al. (2013) applied the classic model of sex chromosome evolution to performance genes brought under sex linkage by fusion. The authors hypothesized that upon the fusion neo-W-linked, i.e., maternally transmitted, alleles of performance genes degenerated due to female achiasmatic meiosis, which induced environmental stress. Under such conditions amplification of neo-Z-linked performance genes is beneficial and new copies get fixed by positive selection (reviewed in Innan and Kondrashov 2010). Hence, amplification of detoxification genes followed by their functional diversification would increase larval detoxification capacity and allow the moths to enter a new adaptive zone by broadening

their host plant spectra (cf. Dermauw et al. 2013) or allowing them to shift to new hosts (cf. Li et al. 2003; Bass et al. 2013; Edger et al. 2015). Given the mechanism proposed, it can be expected that the neo-sex chromosome should comprise synteny blocks enriched in detoxification genes, which would evidence lineage-specific amplification following sex chromosome–autosome fusion. A rise of neo-sex chromosomes should also correlate with changes in host plant use. An analysis of evolutionary patterns of host plant use revealed that one of the main innovations resulting in lepidopteran adaptive radiation was a broadening of the host plant spectrum in lower Ditrysia (Menken et al. 2010), where tortricid moths belong. There is no doubt that interactions of lepidopterans with their host plants generate their present diversity as it was shown that host switching correlates with butterfly diversification (Fordyce 2010; Hardy and Otto 2014). Further research will surely show whether involvement of neo-sex chromosomes in diversification could be a more general phenomenon or whether other drivers predominate in sex chromosome–autosome fusions.

To summarize, even though the architecture of lepidopteran genomes is traditionally considered stable, the presence of neo-sex chromosomes in different taxa has been reported and the list is continuously updated. Sex chromosomes are involved in postzygotic reproductive isolation, and their linkage to autosomes bearing genes involved for instance in detoxification processes may constitute a key innovation that could contribute to the radiation of Lepidoptera clades. Considering the frequency of these rearrangements and the different evolutionary forces that may act on them, the role of neo-sex chromosomes in speciation and adaptation should not be overlooked, and the results obtained until now could be considered a stepping stone to unravel the actual role of these rearrangements In the evolution of Lepidoptera.

Acknowledgements We thank Alexander Barton, Anna Voleníková, Atsuo Yoshido, Irena Hladová, Jindra Šíchová, Magda Zrzavá, Martina Dalíková, and Sander Visser for their critical reading of the manuscript. The present chapter was supported by Grants 14-35819P and 16-10298S of the Czech Science Foundation.

References

Ahola V, Lehtonen R, Somervuo P, Salmela L, Koskinen P et al (2014) The Glanville fritillary genome retains an ancient karyotype and reveals selective chromosomal fusions in Lepidoptera. Nat Commun. doi:10.1038/ncomms5737

Ayala D, Ullastres A, González J (2014) Adaptation through chromosomal inversions in *Anopheles*. Front Genet. doi:10.3389/fgene.2014.00129

Bachtrog D (2013) Y-chromosome evolution: emerging insights into processes of Y-chromosome degeneration. Nat Rev Genet 14(2):113–124

Baker RJ, Bickham JW (1986) Speciation by monobrachial centric fusions. P Natl Acad Sci USA 83(21):8245–8248

Bass C, Zimmer CT, Riveron JM, Wilding CS, Wondji CS et al (2013) Gene amplification and microsatellite polymorphism underlie a recent insect host shift. P Natl Acad Sci USA 110 (48):19460–19465
Bellott DW, Skaletsky H, Pyntikova T, Mardis ER, Graves T et al (2010) Convergent evolution of chicken Z and human X chromosomes by expansion and gene acquisition. Nature 466 (7306):612–616
Bird CE, Fernandez-Silva I, Skillings DJ, Toonen RJ (2012) Sympatric speciation in the post "modern synthesis" era of evolutionary biology. Evol Biol 39:158–180
Brown J, Keith S, Von Schoultz B, Suomalainen E (2004) Chromosome evolution in neotropical Danainae and Ithomiinae (Lepidoptera). Hereditas 141(3):216–236
Bureš P, Zedek F (2014) Holokinetic drive: centromere drive in chromosomes without centromeres. Evolution 68(8):2412–2420
Bush GL, Case SM, Wilson AC, Patton JL (1977) Rapid speciation and chromosomal evolution in mammals. P Natl Acad Sci USA 74(9):3942–3946
Carabajal Paladino LZ, Ferrari ME, Lauría JP, Cagnotti CL, Šíchová J, López SN (2016) The effect of X-rays on cytological traits of Tuta absoluta (Lepidoptera: Gelechiidae). Fla Entomol
Charlesworth D (2013) Plant sex chromosome evolution. J Exp Bot 64(2):405–420
Charlesworth B, Charlesworth D (1978) A model for the evolution of dioecy and gynodioecy. Am Nat 112:975–997
Charlesworth D, Charlesworth B (1980) Sex differences in fitness and selection for centric fusions between sex-chromosomes and autosomes. Genet Res 35:205–214
Charlesworth B, Wall JD (1999) Inbreeding, heterozygote advantage and the evolution of neo–X and neo–Y sex chromosomes. P Roy Soc Lond B Bio 266(1414):51–56
Charlesworth D, Charlesworth B, Marais G (2005) Steps in the evolution of heteromorphic sex chromosomes. Heredity 95(2):118–128
De Prins J, Saitoh K (2003) Karyology and sex determination. Handbuch Der Zoologie/Handb Zool 4(36):449–468
de Villena FPM, Sapienza C (2001) Female meiosis drives karyotypic evolution in mammals. Genetics 159(3):1179–1189
Dermauw W, Osborne EJ, Clark RM, Grbić M, Tirry L et al (2013) A burst of ABC genes in the genome of the polyphagous spider mite *Tetranychus urticae*. BMC Genom. doi:10.1186/1471-2164-14-317
Edger PP, Heidel-Fischer HM, Bekaert M, Rota J, Glöckner G et al (2015) The butterfly plant arms-race escalated by gene and genome duplications. P Natl Acad Sci USA 112(27):8362–8366
Ennis TJ (1976) Sex chromatin and chromosome numbers in Lepidoptera. Can J Genet Cytol 18 (1):119–130
Faria R, Navarro A (2010) Chromosomal speciation revisited: rearranging theory with pieces of evidence. Trends Ecol Evol 25(11):660–669
Felsenstein J (1974) The evolutionary advantage of recombination. Genetics 78(2):737–756
Fontana PG (1976) Improved resolution of the meiotic chromosomes in both sexes of *Euxoa* species and their hybrids (Lepidoptera: Noctuidae). Can J Genet Cytol 18(3):537–544
Fordyce JA (2010) Host shifts and evolutionary radiations of butterflies. P Roy Soc Lond B Bio 277(1701):3735–3743
Fuková I, Nguyen P, Marec F (2005) Codling moth cytogenetics: karyotype, chromosomal location of rDNA, and molecular differentiation of sex chromosomes. Genome 48:1083–1092
Gassner G, Klemetson DJ (1974) A transmission electron microscope examination of hemipteran and lepidopteran gonial centromeres. Can J Genet Cytol 16(2):457–464
Guerrero RF, Kirkpatrick M (2014) Local adaptation and the evolution of chromosome fusions. Evolution 68(10):2747–2756
Hardy NB, Otto SP (2014) Specialization and generalization in the diversification of phytophagous insects: tests of the musical chairs and oscillation hypotheses. P Roy Soc Lond B Bio. doi:10.1098/rspb.2013.2960

Heckel DG, Bryson PK, Brown TM (1998) Linkage analysis of insecticide-resistant acetylcholinesterase in *Heliothis virescens*. J Hered 89(1):71–78
Innan H, Kondrashov F (2010) The evolution of gene duplications: classifying and distinguishing between models. Nat Rev Genet 11(2):97–108
Ironside JE (2010) No amicable divorce? Challenging the notion that sexual antagonism drives sex chromosome evolution. BioEssays 32(8):718–726
Kawazoé A (1987) The chromosome in the primitive or microlepidopterous moth-groups. I. Proc Japan Acad Ser B 63(6):25–28
Kitano J, Peichel CL (2012) Turnover of sex chromosomes and speciation in fishes. Environ Biol Fish 94(3):549–558
Kitano J, Ross JA, Mori S, Kume M, Jones FC et al (2009) A role for a neo-sex chromosome in stickleback speciation. Nature 461(7267):1079–1083
Lande R (1985) The fixation of chromosomal rearrangements in a subdivided population with local extinction and colonization. Heredity 54(3):323–332
Li W, Schuler MA, Berenbaum MR (2003) Diversification of furanocoumarin-metabolizing cytochrome P450 monooxygenases in two papilionids: specificity and substrate encounter rate. P Natl Acad Sci USA 100(2):14593–14598
Li X, Schuler MA, Berenbaum MR (2007) Molecular mechanisms of metabolic resistance to synthetic and natural xenobiotics. Annu Rev Entomol 52:231–253
Lohse K, Clarke M, Ritchie MG, Etges WJ (2015) Genome-wide tests for introgression between cactophilic *Drosophila* implicate a role of inversions during speciation. Evolution 69(5):1178–1190
Lukhtanov VA (2000) Sex chromatin and sex chromosome systems in nonditrysian Lepidoptera (Insecta). J Zool Syst Evol Res 38(2):73–79
Lukhtanov V, Puplesiene J (1999) Polyploidy in bisexual Lepidoptera species (Insecta: Lepidoptera): old hypotheses and new data. Bonn Zool Beitr 48:313–328
Maeki K (1981) Notes on the W-chromosome of the butterfly, *Graphium sarpedon* (Papilionidae, Lepidoptera). P Jpn Acad B 57(10):371–373
Maeki K, Ae SA (1969) Studies of the chromosomes of Formosan Rhopalocera. 4. Danaidae and Satyridae. Kontyu 37(1):99–109
Mank JE, Vicoso B, Berlin S, Charlesworth B (2010) Effective population size and the Faster-X effect: empirical results and their interpretation. Evolution 64(3):663–674
Marec F, Sahara K, Traut W (2010) Rise and fall of the W chromosome in Lepidoptera. In: Goldsmith MR, Marec F (eds) Molecular biology and genetics of the Lepidoptera. CRC Press, Boca Raton, pp 49–63
Martin SH, Dasmahapatra KK, Nadeau NJ, Salazar C, Walters JR et al (2013) Genome-wide evidence for speciation with gene flow in *Heliconius* butterflies. Genome Res 23(11):1817–1828
Matsubayashi KW, Ohshima I, Nosil P (2010) Ecological speciation in phytophagous insects. Entomol Exp Appl 134(1):1–27
McAllister BF, Sheeley SL, Mena PA, Evans AL, Schlötterer C (2008) Clinal distribution of a chromosomal rearrangement: a precursor to chromosomal speciation? Evolution 62(8):1852–1865
Menken SB, Herrebout WM, Wiebes JT (1992) Small ermine moths (*Yponomeuta*): their host relations and evolution. Annu Rev Entomol 37:41–66
Menken SB, Boomsma JJ, Van Nieukerken EJ (2010) Large-scale evolutionary patterns of host plant associations in the Lepidoptera. Evolution 64(4):1098–1119
Murakami A, Imai HT (1974) Cytological evidence for holocentric chromosomes of the silkworms, *Bombyx mori* and *B. mandarina* (Bombycidae, Lepidoptera). Chromosoma 47 (2):167–178
Nguyen P, Sýkorová M, Šíchová J, Kůta V, Dalíková M et al (2013) Neo-sex chromosomes and adaptive potential in tortricid pests. P Natl Acad Sci USA 110(17):6931–6936
Nilsson NO, Löfstedt C, Dävring L (1988) Unusual sex chromosome inheritance in six species of small ermine moths (*Yponomeuta*, Yponomeutidae, Lepidoptera). Hereditas 108(2):259–265

Nygren GH, Nylin S, Stefanescu C (2006) Genetics of host plant use and life history in the comma butterfly across Europe: varying modes of inheritance as a potential reproductive barrier. J Evol Biol 19(6):1882–1893

O'Meally D, Ezaz T, Georges A, Sarre SD, Graves JAM (2012) Are some chromosomes particularly good at sex? Insights from amniotes. Chromosome Res 20(1):7–19

Ohno S (1967) Sex chromosomes and sex-linked genes. Springer, Berlin

Olmo E (2005) Rate of chromosome changes and speciation in reptiles. Genetica 125(2–3):185–203

Ortiz-Barrientos D, Engelstädter J, Rieseberg LH (2016) Recombination rate evolution and the origin of species. Trends Ecol Evol 31(3):226–236

Pala I, Naurin S, Stervander M, Hasselquist D, Bensch S et al (2012) Evidence of a neo-sex chromosome in birds. Heredity 108(3):264–272

Pennell MW, Kirkpatrick M, Otto SP, Vamosi JC, Peichel CL et al (2015) Y fuse? Sex chromosome fusions in fishes and reptiles. PLoS Genet. doi:10.1371/journal.pgen.1005237

Perrin N (2009) Sex reversal: a fountain of youth for sex chromosomes? Evolution 63(12):3043–3049

Pokorná M, Rens W, Rovatsos M, Kratochvíl L (2014) A ZZ/ZW sex chromosome system in the thick-tailed gecko (*Underwoodisaurus milii*; Squamata: Gekkota: Carphodactylidae), a member of the ancient gecko lineage. Cytogenet Genome Res 142(3):190–196

Potter S, Moritz C, Eldridge MDB (2015) Gene flow despite complex Robertsonian fusions among rock-wallaby(Petrogale) species. Biol Lett. doi:10.1098/rsbl.2015.0731

Presgraves DC (2002) Patterns of postzygotic isolation in Lepidoptera. Evolution 56(6):1168–1183

Prowell DP (1998) Sex linkage and speciation in Lepidoptera. In: Howards DJ, Berlocher SW (eds) Endless forms: species and speciation. Oxford University Press, Oxford, pp 309–319

Rice WR (1987) The accumulation of sexually antagonistic genes as a selective agent promoting the evolution of reduced recombination between primitive sex chromosomes. Evolution 41 (4):911–914

Robinson R (1971) Lepidoptera genetics. Pergamon Press, Oxford

Roesti M, Kueng B, Moser D, Berner D (2015) The genomics of ecological vicariance in three spine stickleback fish. Nat Commun. doi:10.1038/ncomms9767

Ross JA, Urton JR, Boland J, Shapiro MD, Peichel CL (2009) Turnover of sex chromosomes in the stickleback fishes (Gasterosteidae). PLoS Genet. doi:10.1371/journal.pgen.1000391

Sackton TB, Corbett-Detig RB, Nagaraju J, Vaishna L, Arunkumar KP et al (2014) Positive selection drives faster-Z evolution in silkmoths. Evolution 68(8):2331–2342

Saitoh K, Abe A (1970) On the chromosomes of five species of Danaidae from Nepal Hymalaya (Lepidoptera; Rhopalocera). Spec Bull Lep Soc Jap 4:153–158

Sharp HE, Rowell DM (2007) Unprecedented chromosomal diversity and behaviour modify linkage patterns and speciation potential: structural heterozygosity in an Australian spider. J Evol Biol 20(6):2427–2439

Šíchová J, Nguyen P, Dalíková M, Marec F (2013) Chromosomal evolution in tortricid moths: conserved karyotypes with diverged features. PLoS ONE. doi:10.1371/journal.pone.0064520

Šíchová J, Voleníková A, Dincă V, Nguyen P, Vila R et al (2015) Dynamic karyotype evolution and unique sex determination systems in *Leptidea* wood white butterflies. BMC Evol Biol. doi:10.1186/s12862-015-0375-4

Šíchová J, Ohno M, Dincă V, Watanabe M, Sahara K et al (2016) Fissions, fusions, and translocations shaped the karyotype and multiple sex chromosome constitution of the northeast Asian wood white butterfly, Leptidea amurensis. Biol J Linn Soc 118(3):457–47

Skaletsky H, Kuroda-Kawaguchi T, Minx PJ, Cordum HS, Hillier L et al (2003) The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. Nature 423 (6942):825–837

Smadja CM, Butlin RK (2011) A framework for comparing processes of speciation in the presence of gene flow. Mol Ecol 20:5123–5140

Sohn JC, Regier JC, Mitter C, Davis D, Landry JF et al (2013) A molecular phylogeny for Yponomeutoidea (Insecta, Lepidoptera, Ditrysia) and its implications for classification, biogeography and the evolution of host plant use. PLoS ONE. doi:10.1371/journal.pone.0055066

Sohn JC, Rregier JC, Mitter C, Adamski D, Landry JF et al (2015) Phylogeny and feeding trait evolution of the mega-diverse Gelechioidea (Lepidoptera: Obtectomera): new insight from 19 nuclear genes. Syst Entomol. doi:10.1111/syen.12143

Sperling FA (1994) Sex-linked genes and species differences in Lepidoptera. Can Entomol 126 (03):807–818

Suomalainen E (1969) On the sex chromosome trivalent in some Lepidoptera females. Chromosoma 28(3):298–308

Suomalainen E (1971) Unequal sex chromosomes in a moth, *Lozotaenia forsterana* F. (Lepidoptera: Tortricidae). Hereditas 68(2):313–315

Thompson JN (1988) Evolutionary ecology of the relationship between oviposition preference and performance of offspring in phytophagous insects. Entomol Exp Appl 47(1):3–14

Traut W, Marec F (1996) Sex chromatin in Lepidoptera. Q Rev Biol 71(2):239–256

Traut W, Marec F (1997) Sex chromosome differentiation in some species of Lepidoptera (Insecta). Chromosome Res 5(5):283–291

Traut W, Sahara K, Marec F (2007) Sex chromosomes and sex determination in Lepidoptera. Sex Dev 1(6):332–346

Turner H, Lieshout N, Van Ginkel WE, Menken SB (2010) Molecular phylogeny of the small ermine moth genus *Yponomeuta* (Lepidoptera, Yponomeutidae) in the Palaearctic. PloS ONE. doi:10.1371/journal.pone.0009933

van Nieukerken EJ, Kaila L, Kitching IJ, Kristensen NP, Lees DJ et al (2011) Order Lepidoptera Linnaeus, 1758. Zootaxa 3148:212–221

Van't Hof AE, Nguyen P, Dalíková M, Edmonds N, Marec F et al (2013) Linkage map of the peppered moth, *Biston betularia* (Lepidoptera, Geometridae): a model of industrial melanism. Heredity 110(3):283–295

Veltsos P, Keller I, Nichols RA (2008) The inexorable spread of a newly arisen neo-Y chromosome. PLoS Genet. doi:10.1371/journal.pgen.1000082

Vicoso B, Bachtrog D (2013) Reversal of an ancient sex chromosome to an autosome in *Drosophila*. Nature 499(7458):332–335

Vicoso B, Emmerson JJ, Zektser Y, Mahajan S, Bachtrog D (2013) Comparative sex chromosome genomics in snakes: differentiation, evolutionary strata, and lack of global dosage compensation. PLoS Biol. doi:10.1371/journal.pbio.1001643

Vítková M, Fuková I, Kubíčková S, Marec F (2007) Molecular divergence of the W chromosomes in pyralid moths (Lepidoptera). Chromosome Res 15(7):917–930

Walters J, Mongue AJ (2016) A neo-sex chromosome in the Monarch butterfly, *Danaus plexippus*. bioRxiv. doi:10.1101/036483

White MJD (1973) Animal cytology and evolution. Cambridge University Press, Cambridge

Wolf KW (1994) The unique structure of lepidopteran spindles. Int Rev Cytol 152:1–48

Xie X, Cheng T, Wang G, Duan J, Niu W et al (2012) Genome-wide analysis of the ATP-binding cassette (ABC) transporter gene family in the silkworm, *Bombyx mori*. Mol Biol Rep 39 (7):7281–7291

Yeaman S (2013) Genomic rearrangements and the evolution of clusters of locally adaptive loci. Proc Natl Acad Sci USA 110(19):E1743–E1751

Yoshida K, Kitano J (2012) The contribution of female meiotic drive to the evolution of neo-sex chromosomes. Evolution 66(10):3198–3208

Yoshida K, Makino T, Yamaguchi K, Shigenobu S, Hasebe M et al (2014) Sex chromosome turnover contributes to genomic divergence between incipient stickleback species. PLoS Genet. doi:10.1371/journal.pgen.1004223

Yoshido A, Marec F, Sahara K (2005) Resolution of sex chromosome constitution by genomic in situ hybridization and fluorescence in situ hybridization with (TTAGG)n telomeric probe in some species of Lepidoptera. Chromosoma 114:193–202

Yoshido A, Sahara K, Marec F, Matsuda Y (2011) Step-by-step evolution of neo-sex chromosomes in geographical populations of wild silkmoths, *Samia cynthia* ssp. Heredity 106(4):614–624

Yoshido A, Šíchová J, Kubíčková S, Marec F, Sahara K (2013) Rapid turnover of the W chromosome in geographical populations of wild silkmoths, *Samia cynthia* ssp. Chromosome Res 21(2):149–164

Yu QY, Lu C, Li WL, Xiang ZH, Zhang Z (2009) Annotation and expression of carboxylesterases in the silkworm, *Bombyx mori*. BMC Genomics. doi:10.1186/1471-2164-10-553

Zangerl AR, Liao LH, Jogesh T, Berenbaum MR (2012) Aliphatic esters as targets of esterase activity in the parsnip webworm (*Depressaria pastinacella*). J Chem Ecol 38(2):188–194

Chapter 12
Recent Developments on Bacterial Evolution into Eukaryotic Cells

Mauro Degli Esposti, Otto Geiger and Esperanza Martinez-Romero

Abstract This article stems from the presentation of Mauro Degli Esposti at the 19th Meeting of Evolutionary Biology held in Marseille and summarizes recent developments that have emerged in regard to the evolution of the eukaryotic cell. After updating the application of recently introduced approaches for evaluating the likely bacterial relatives to the symbionts that originated the mitochondrial organelles, the article examines the possibility that such symbionts might have engaged in a syntrophic association with ammonia-oxidizing archaea. The bacterial ancestors of mitochondria could have provided the ammonia required for chemoautotrophic growth of the archaean host, the nitrite by-product of which could have been subsequently recycled back to ammonia by the pathway of nitrate assimilation that is shared by a few proteobacteria and several eukaryotes. The similarity of this syntrophic association with the symbiosis between nitrogen-fixing bacteria and plants is discussed in detail. The article then explores the possible presence of relics of archaean lipids in current eukaryotes, a fundamental problem that weakens current views on the evolution of the eukaryotic cell.

12.1 Introduction

This article discusses recent developments on the origin of mitochondria and related organelles of eukaryotic cells from bacterial ancestors following the presentation by Mauro Degli Esposti at the Marseille conference of September 2015. According to the current consensus, these bacterial ancestors became symbionts of an archaean host that is considered to be related to a group of archaea (Spang et al. 2015; López-García and Moreira 2015; Koonin 2015) including organisms capable of nitrification, i.e. oxidation of ammonia to nitrite (Brochier-Armanet et al. 2012;

M. Degli Esposti (✉)
Italian Institute of Technology, Genoa, Italy
e-mail: Mauro.DegliEsposti@iit.it

O. Geiger · E. Martinez-Romero
Centre for Genomic Sciences, UNAM, Cuernavaca, Morelos, Mexico

P. Pontarotti (ed.), *Evolutionary Biology*, DOI 10.1007/978-3-319-41324-2_12

Hatzenpichler 2012). One peculiar metabolic trait that is present in some fungi entails the opposite reaction, namely the reduction of nitrite to ammonia for re-oxidizing NADH under anaerobic conditions (Takasaki et al. 2004). This reaction of the nitrogen cycle is part of the pathway of bacterial and fungal nitrate assimilation (Moreno-Vivián and Ferguson 1998), the crucial enzyme of which is a large Mo-containing enzyme, NADH-dependent nitrate reductase coded by the *NiaD* gene in fungi (Johnstone et al. 1990) and by the *NapA* gene in enterobacteria (Moreno-Vivián and Ferguson 1998). The multidomain enzyme and its operon constitute crucial clues to identify the closest relatives to proto-mitochondria (Degli Esposti et al. 2014), the bacterial ancestors of contemporary organelles.

The article is structured in three parts. Initially, it surveys recent phylogenomics data that expand and refine previously proposed clues and traits that can be used to identify bacterial relatives of proto-mitochondria, including assimilatory nitrate reduction (Degli Esposti et al. 2014). Subsequently, the article elaborates on the recurrent presence of such traits in bacteria that form mutualistic symbiosis with plants, often producing syntrophic associations based upon the exchange of nitrogen compounds (Martinez-Romero 2012). Finally, it presents a brief overview on the membrane lipids that are present in eukaryotes and may derive from either the bacterial guest or the archaean host that contributed to eukaryogenesis (cf. Koonin 2015).

12.1.1 Part 1. Screening Extant Proteobacteria to Identify Relatives to Proto-Mitochondria

The origin of eukaryotic cells constitutes one of the major transitions in evolution and derives from a unique event of symbiosis between different prokaryotes (Martin et al. 2015). Recent developments have solidified the view that the host in this symbiotic event was an archaean organism related to the TACK superphylum (Koonin 2015), while the guest was a proteobacterial organism that then evolved into current mitochondria. The general consensus is that such a guest belonged to the alpha class of proteobacteria (Müller et al. 2012; Martin et al. 2015; Ku et al. 2016; Gray 2015), even if the exact lineage remains unclear and controversial—for a review, see: Müller et al. (2012), Degli Esposti (2014), Gray (2015). The controversy primarily derives from the intrinsic limits of phylogenetic and phylogenomic analyses applied to such an ancient event as eukaryogenesis, which probably occurred between 1.5 and 2 billion years ago (Müller et al. 2012). Recently, three complementary approaches have been introduced to identify current relatives of proto-mitochondria (Degli Esposti et al. 2014). These approaches stem from a detailed analysis of bacterial genomes available at the beginning of 2014, which is updated here in the light of the additional genomic information accumulated in the last two years. Figure 12.1 illustrates the conceptual basis for the approaches: Mitochondria share genomic signatures and metabolic traits not just with alpha proteobacteria, but also with facultatively anaerobic beta and gamma proteobacteria

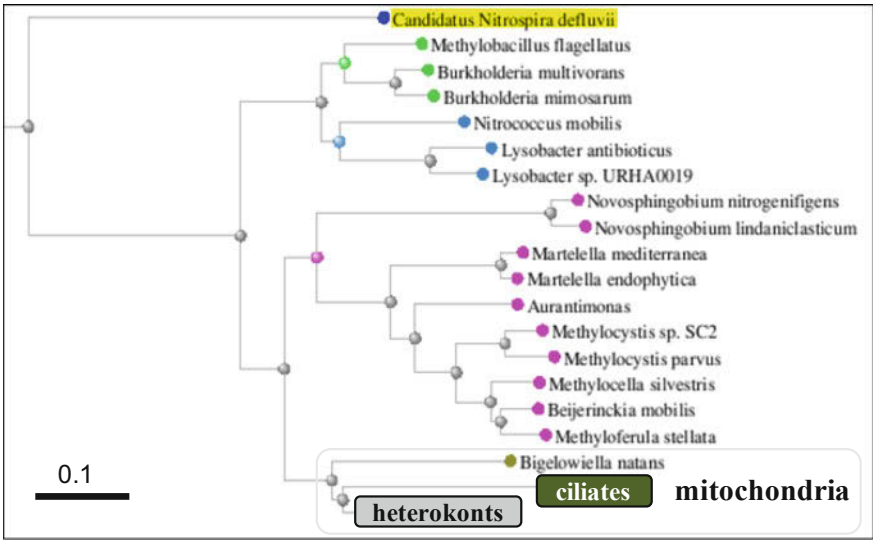


Fig. 12.1 Typical phylogenetic tree of a very conserved subunit of respiratory complex I, NuoD. The neighbor-joining tree of the ubiquinone-binding subunit NuoD (also called Nad7 or 49-kDa subunit in mitochondria) was obtained using the tools of the DELTABLAST facility as described previously (Degli Esposti et al. 2014, 2015). The NuoD protein from the Nitrospirae, *Ca. Nitrospira defluvii* (accession: CBK40015), was used as a query—hence it is highlighted in yellow. This protein appears as the outgroup of the proteobacterial and mitochondrial sequences in the tree view, which was rooted on the homologous large subunit of CO-induced hydrogenase from the rhizobial *Pleomorphomonas* (cut-off from view). Colour codes for the bacterial taxa are as follows: *purple*, alpha proteobacteria; *green*, beta proteobacteria; *pale blue*, gamma proteobacteria. The bottom branch contains the mitochondrial homologues from the following taxa: the Rhizarian *Bigelowiella natans* (Cercozoans), 5 ciliates (*dark green box*) and 4 heterokonts including the brown alga *Ectocarpus silicosus* plus the Ascomycete *Aspergillus flavus* (*grey box*) which have *NiaD* nitrate reductases. Note how the sequences of the mitochondrial and alpha proteobacterial proteins form part of the same clade that is sister to the clade containing the sequences of beta and gamma proteobacteria, as previously reported for COX1 and COX3 (Degli Esposti 2014)

(Degli Esposti 2014), consistent with the mosaic nature of bacterial inheritance in eukaryotes (Gray 2015).

The tree shown in Fig. 12.1 underlines a fundamental aspect of the phylogenesis of mitochondrial proteins, which is often overlooked. Such proteins cluster together in a clade that contains all their close homologues of alpha proteobacteria and is sister to the clade formed by their homologues in beta and gamma proteobacteria. Hence, mitochondrial proteins are effectively related to their homologues from all alpha proteobacterial taxa and not just a specific branch of such bacteria as often claimed, a concept that has been overlooked even in recent studies of phylogenomics (Pittis and Gabaldón 2016). The closeness may derive from the proteins or genes chosen for the phylogenetic analysis and is heavily influenced by phenomena of long-branch attraction for obligate endocellular parasites such as Rickettsiales (Searcy 2003; Brinkmann and Philippe 2007; Rodríguez-Ezpeleta and Embley 2012; Gray 2015).

Table 12.1 Bacteria possessing *NiaD*-like nitrate reductase of fungi and their relatives

alpha with NiaD-like nitrate reductase	COX operons of type a	relatives with COX operon type a-II	nitrogenase	metabolic traits
Methylocystis parvus	a & a-I	Methylocystis sp. SC2	yes	methanotroph, close to *Spagnum* endosymbionts
Beijerinckia mobilis	?	Methylocella silvestris, Methyloferula stellata	yes	methylotroph, sporadic plant endophyte
Martelella endophytica	a-I ?	Aurantimonas coralicida, Aurantimonas manganoxydans	yes	plant endophyte
Martelella mediterranea	a-I ?			
Martelella sp. AD-3	a			
Novosphingonbium nitrogenifigens	COX operon type a-II	Sphingonbium banderi, Sphingomonas	yes	diazotroph
Novosphingobium sp. P6W	COX operon type a-II & a			aromatic degrading, rarely plant endophyte
Novosphingobium lindaniclasticum	COX operon type a-II & a			
beta with NiaD-like nitrate reductase				
Methylobacillus flagellatus	COX operon type a-II			methylotroph, related to plant endophytes
Burkholderia mimosarum	COX operon type a-II		yes	plant N2 fixing symbiont
Burkholderia sp. CCGE1002	COX operon type a-II		yes	plant N2 fixing symbiont
Burkholderia xenovorans	COX operon type a-II		yes	polychlorinatred biphenyl degrading
Burkholderia sprentiae	COX operon type a-II		yes	
Burkholderia sp. JPY251	COX operon type a-II		yes	
Burkholderia sp. WSM4176	COX operon type a-II		yes	
Burkholderia sp. CCGE1003	COX operon type a-II			
Burkholderia sp. 9120	COX operon type a-II			
Burkholderia sp. WSM2230	COX operon type a-II			
Burkholderia multivorans	COX operon type a-II & a			
gamma with NiaD-like nitrate reductase				
Methylomicrobium album BG8	COX operon type a-II	Methylobacter luteus, Methylosarcina fibrata, Methylobacter tundripaludum	in relatives	methanotroph, close relative of endosymbionts of clams in hydrothermal vents
Methylomicrobium agile	COX operon type a-II			

The presence of a precursor form of fungal *NiaD* nitrate reductase, defined as *NiaD*-like, is probably the rarest trait that extant alpha proteobacteria share with eukaryotes (Degli Esposti 2014). Currently, about 30 alpha proteobacterial organisms contain this protein and the associated *NirB* nitrite reductase in their genomes. Besides various members of the Acetobacteraceae family, they include three *Novosphingobium* and *Martelella* organisms, two *Beijerinckia* species and the methanotroph *Methylocystis parvus* (Table 12.1 and data not shown). Conversely, the same proteins are present in many enterobacteria such as *Klebsiella* (Lin et al. 1993), two methanogenic gamma proteobacteria and a variety of beta proteobacteria, predominantly of clade I of the *Burkholderia* group (Sawana et al. 2013) (Table 12.1 and data not shown). Among such organisms, Acetobacteraceae and enterobacteria can be excluded from mitochondrial ancestry because they do not possess active aa3-type cytochrome oxidases (COX). These oxidases are assembled in eight different operon types in proteobacteria: one in some delta, one shared by both beta and gamma, another specific to alpha only (COX operon type b) and the rest present in alpha, beta and gamma proteobacteria (Degli Esposti 2014). Although the mitochondrial subunits of eukaryotic cytochrome oxidase generally resemble those of alpha proteobacterial COX operon type b, a few fungi have mitochondrial COX1 proteins containing an additional transmembrane helix at the C-terminus. The protein sequence around this additional helix shows signatures that are typical of the bacterial COX4 subunit, which is characteristically fused to COX1 in COX operon type a-II (Degli Esposti et al. 2014). Currently, five fungal species belonging to four different orders of Ascomycetes have this usual COX1 protein: the plant pathogen *Zymoseptoria tritici* (Testa et al. 2015), *Sclerotinia borealis* (Mardanov et al. 2014), *Peltigera membranacea* (Xavier et al. 2012), *Paracoccidioides brasiliensis*

(Desjardins et al. 2011) and *Trichoderma reesei* (Martinez et al. 2008), as well as the Heterolobosea *Acrasis kona*.

The presence of mitochondrial COX1 proteins resembling those of COX operon type a-II has previously suggested that the ancestors of mitochondria must have possessed this operon in combination with COX operon type b (Degli Esposti et al. 2014). Therefore, the bacteria that possess in their genome both such COX operons could be considered relatives of proto-mitochondria. If we combine this trait with the presence of *NiaD*-like nitrate reductase, only the organisms listed in Table 12.1 remain, defining a restricted group of taxa that may have the largest probability of being close to proto-mitochondria (Degli Esposti et al. 2014). Intriguingly, many of such organisms can fix N_2, since they possess a complete nitrogenase system (Table 12.1). They often are endophytic or symbionts, or are close relatives of organisms that engage in symbiotic interactions (Table 12.1). Among the taxa exhibiting these traits, the beta and gamma organisms listed in Table 12.1 clearly cannot be close to proto-mitochondria, which are known to be related to alpha proteobacteria (Gray 2015). However, they indicate that proto-mitochondria might have been close to the common ancestors of the alpha, beta and gamma proteobacteria that combined *NiaD*-like-driven assimilatory denitrification with terminal oxidases including COX operons of type a.

The major novelty of the abovementioned analysis regards three strains of *Novosphingobium* (Table 12.1), a metabolically versatile alpha proteobacterium belonging to the order of Sphingomonadales, which has been previously considered among the possible ancestors of mitochondria (Finnegan et al. 2003). However, *Novosphingobium* sequences of the Rieske iron sulphur protein of the cytochrome bc_1 complex, a protein with strong phylogenetic signal hereafter abbreviated as ISP, present a characteristic insert of five amino acids at the C-terminus (Cimit5, Fig. 12.2), which is absent in mitochondrial homologues (Degli Esposti et al. 2014).

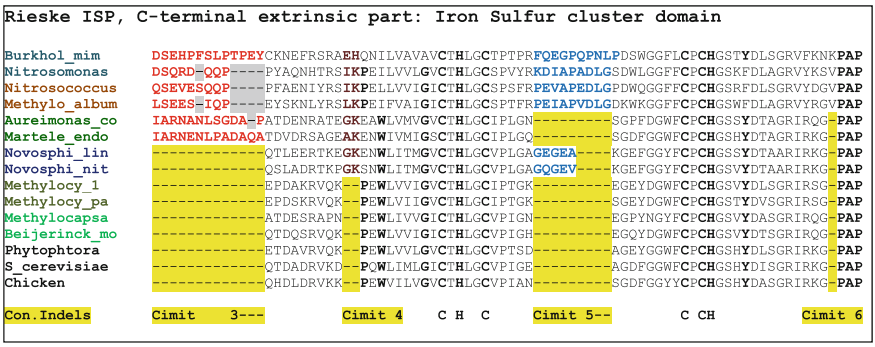


Fig. 12.2 Alignment of the cluster domain of the Rieske ISP of bacteria and mitochondria. The alignment was manually refined using the latest coordinates of the 3D structure of various mitochondrial and bacterial ISP proteins (Berry et al. 2004) as previously reported (Degli Esposti et al. 2014), using the same nomenclature for the conserved insertions and deletions (called Cimit). The residues of the Cimit4 deletion are in *dark purple*, while the residues of the Cimit3 and Cimit5 *insert* are shown in *red* and *blue*, respectively. The ligand residues of the iron sulphur cluster are in bold character and indicated at the bottom of the alignment

Conversely, the ISP sequence of *Martelella* and related organisms presents a larger central insert (Cimit3, Fig. 12.2), which is also absent in mitochondrial homologues as well as in the ISP sequence of *Beijerinckia* and *Methylocystis* (Fig. 12.2 cf. Degli Esposti et al. 2014). Hence, the application of the third approach introduced by Degli Esposti et al. (2014), namely the inspection of conserved INDELs in the ISP sequence (Fig. 12.2), restricts the possible relatives of proto-mitochondria to a few methylotrophs of the genera *Beijerinckia* and *Methylocystis* among all the organisms now found to possess both *NiaD*-like and COX operon type a-II proteins listed in Table 12.1. These findings sustain earlier conclusions that pointed at such organisms as the most likely relatives to the ancestors of mitochondria (Degli Esposti 2014).

It is of interest to consider the potential evolutionary relevance of the preservation of the assimilatory nitrate reductase system in eukaryotes from ancestral proto-mitochondria (Degli Esposti et al. 2014), enabling the pathway of ammonia fermentation present in some fungi (Takasaki et al. 2004). In the framework of syntrophic interactions, alpha proteobacterial symbionts producing ammonia would provide a thriving niche for ammonia-oxidizing archaea, usually abbreviated as AOA and classified among the Thaumarchaeota (Brochier-Armanet et al. 2007, 2012; Hatzenpichler 2012; Raymann et al. 2015). Thaumarchaeota are now included in the TACK superphylum (Guy and Ettema 2011), to which the ancestor of eukaryotic cells belong (Spang et al. 2015; López-García and Moreira 2015; Koonin 2015). Examples of microbial communities of Thaumarchaeota AOA associated with facultatively anaerobic proteobacteria (predominantly gamma, but also alpha) have been documented in diverse environments characterized by gradients of oxic–anoxic conditions (Muller et al. 2012; Hatzenpichler 2012; Tetu et al. 2013; Hawley et al. 2014; Gagliano et al. 2016). In particular, the following genera of alpha proteobacteria have been found in significant abundance within such communities: the methylotrophs *Beijerinckia* and *Methylocystis* in volcanic soils (Gagliano et al. 2016), *Reugeria* and unidentified Rhizobiales in marine low oxygen strata (Hawley et al. 2014). In the latter case, metabolic associations with nitrite-oxidizing organisms of the *Nitrospira* group have been proposed (Hawley et al. 2014). The key role of the abundant *Nitrospira* organisms in the nitrogen cycle has been recently emphasized by the discovery of *Nitrospira* strains capable of carrying out the complete process of nitrification; i.e., they oxidize ammonia to nitrite besides oxidizing nitrite to nitrate (Daims et al. 2015). In theory, the association of AOA with such organisms would be unnecessary if primordial archaean lineages would carry the same capacity of complete nitrification, which might have been subsequently lost along evolution. However, organisms of this kind have yet to be discovered.

Ammonia is produced geochemically in specific hotspots (e.g. see Gagliano et al. 2016); however, in most biological environments, it derives from atmospheric N_2 fixed by bacteria such as rhizobia (Ormeño-Orrillo et al. 2013), while ammonia produced via assimilatory pathways is usually consumed within bacterial cells. Hence, bacterial guests capable of N_2 fixation under low oxygen concentrations, as well as nitrate assimilation under anaerobic conditions, would be ideal syntrophic partners for nitrifying archaean hosts. Ancient organisms close to the root

separating Nitrospirae, which are the dominant nitrite oxidizers in most environments (Hawley et al. 2014; Daims et al. 2015), from (alpha) proteobacteria might have engaged in syntrophic associations with AOA, which can oxidize ammonia only to nitrite. Alternatively, proto-mitochondria might have possessed enzymes for the oxidation of nitrite under aerobic conditions, e.g. the copper-containing *NirK* and nitrite reductase of the *NirB* type as in fungi for the reverse assimilatory reaction under anaerobic conditions. Genomic and bioinformatic analyses indicate that, so far, there is no extant alpha proteobacterium combining nitrogenase with NADH-dependent nitrite reductase for assimilatory production of ammonia, together with nitrite oxidoreductase for the oxidation of nitrite produced by a hypothetical syntrophic archaean. Consequently, it is necessary to consider a syntrophic consortium with *Nitrospira* organisms to support the hypothetical symbiosis between AOA and N_2-fixing, nitrate-assimilating alpha proteobacteria, of which there are several extant examples among currently described organisms (Table 12.1). Moreover, the archaean host might have provided carbon metabolites, e.g. pyruvate, as a fuel for N_2-fixing guests under both high and low concentrations of oxygen, given the versatility of carbon fixation pathways in current AOA (Zhalina et al. 2012). A similar exchange of carbon metabolites exists in the well-studied symbiosis between N_2-fixing bacteria and plants, as discussed below.

12.1.2 Part 2. Syntrophic Associations Between Bacteria and Plants Based on Nitrogen Compounds

Symbiosis of different bacteria and their hosts are widespread in nature and can enhance the carrying capacity of diverse environments (Martinez-Romero 2012). The natural association of N_2-fixing bacteria (diazotrophs) with photosynthetic organisms is fundamentally based on syntrophism, a term that refers to a type of mutualism where the consumption of excreted substances by symbiotic partners can accelerate the metabolism of the donor symbiont (Kouzuma et al. 2015). In common nitrogen-fixing symbiosis, carbon compounds produced by plants are exchanged for fixed nitrogen produced by bacteria, mainly in the form of ammonia (Ormeño-Orrillo et al. 2013). Among diazotrophs, rhizobia are well known for their nitrogen contribution to legumes in root or stem specialized structures called nodules. In such nodules, bacteria differentiate into bacteroides and produce the nitrogenase enzyme for N_2 fixation. In some legumes of temperate climates, there is a terminal differentiation process of bacteroides in response to plant peptides (Kondorosi et al. 2013). A review on the evolution and phylogenies of rhizobia was presented at the 13th Evolutionary Biology Meeting at Marseilles and reported in Martinez-Romero et al. (2010). Since then, a large diversity of new rhizobial lineages has been revealed in various genomic and metagenomic studies (Miranda-Sánchez et al. 2015; reviewed in Ormeño-Orrillo et al. 2015). This is the case, in particular, of *Rhizobium leguminosarum* exhibiting different genomic

lineages co-existing in a single limited habitat, with restricted genetic recombination. These diverse genomic lineages could be considered as different species, though the term 'biovar' was used for their designation instead (Kumar et al. 2015).

Rhizobium legume symbiosis is not an obligate one, and therefore, it has two important evolutionary characteristics: It follows precise recognition and colonization processes at the onset of each nodulation stage and, additionally, the microsymbiont genome does not undergo a reduction process as in the ancestors of mitochondria (Gray 2015). Both legumes and rhizobia can independently live without the corresponding partner. Seemingly, the lack of obligate dependence has hampered the development of N_2-fixing organelles in legumes. In contrast, an equivalent system has evolved in a diatom containing a cyanobacteria-derived organelle (with a reduced genome) that may not be grown in culture media (Prechtl et al. 2004; Kneip et al. 2007). The refined rhizobial colonization process has been extensively studied both in plants and in bacteria (Oldroyd and Downie 2008) and includes an attachment of rhizobia to roots followed by differential gene expression of both the bacterial and plant symbiotic partner in response to each other signals (Oldroyd and Downie 2008; Ramachandran et al. 2011; López- Guerrero et al. 2012).

The most remarkable molecular signal in rhizobial symbiosis is the Nod factor produced by rhizobia, which bear structural similarity to Myc factors produced by arbuscular mycorrhiza (Maillet et al. 2011). Recently, it has been reported that some of the *nod* genes might have emerged in a basal group of actinobacterial N_2-fixing *Frankia* organisms (Persson et al. 2015). *nod* genes resembling those from rhizobia are also found in some beta proteobacteria of the genus *Burkholderia* and *Cupriavidus*, which are capable of nodulating legumes (Andrus et al. 2012; Gyaneshwar et al. 2011). Presumably, *nod* genes, with the exception of *nodIJ* (Aoki et al. 2013), were laterally transferred from alpha to beta rhizobia (Moulin et al. 2011; Martinez-Romero et al. 2010). These genes are considered accessory in rhizobial genomes and consequently are prone to events of lateral gene transfer (López-Guerrero et al. 2012). Indeed, *nod* genes do not have a congruent evolutionary history with respect to housekeeping rhizobial genes (reviewed in Martinez-Romero 2009). Rhizobial *nod* genes are located in symbiosis plasmids or in symbiosis islands as other genes involved in the process of N_2 fixation, such as the *nif* genes responsible for the catalytic process of the nitrogenase reaction.

Rhizobia survive as free-living organisms in soils and associate with different legume and non-legume rhizospheres or may colonize plant internal tissues as endophytes of many diverse plants, including cereals (Chi et al. 2005; Chaintreuil et al. 2000), bananas (Martínez et al. 2003) and other plants (Doty et al. 2005). Consequently, they have a metabolic versatility, backed up by large genomes, which are not found in obligate endosymbionts. Indeed, we have found nitrate reduction genes in rhizobia and related N_2-fixing bacteria (see previous section and Table 12.1). Hence, nitrate assimilation and other pathways that are found in current genomes suggest that rhizobia have retained a reservoir of genes conferring metabolic adaptability (López-Guerrero et al. 2013), as well as syntrophic capacity

that originally emerged during the early evolution of alpha proteobacteria, from the root of which the ancestors of mitochondria diverged (Degli Esposti et al. 2015).

12.1.3 Part 3. How to Bridge the Lipid Divide Between Prokaryotes and Eukaryotes

The initial proposal for the three domains of life, Eukarya, Bacteria and Archaea (Woese and Fox 1977), was soon supported by one of the most remarkable features distinguishing Archaea from Bacteria and Eukarya, which is the lipid composition of the archaeal membrane. In archaea, isoprenoid hydrocarbon side chains are linked via ether bonds to the *sn*-glycerol-1-phosphate backbone. In contrast, in bacteria and eukarya, fatty acyl side chains are linked via ester bonds to the *sn*-glycerol-3-phosphate backbone. Nevertheless, the polar head groups of membrane lipids are globally shared by the three domains of life.

Eukaryotic membranes contain sphingolipids (SL), phosphatidylcholine (PC), phosphatidylethanolamine (PE) and cholesterol (CHO) as major (most abundant) lipids. Mitochondrial, lysosomal and smooth ER membranes have phosphatidylglycerol (PG) and cardiolipin (CL) in addition (Nelson and Cox 2013; Ridgway 2016). Among the minor lipids in eukaryotic membranes usually feature phosphatidic acid (PA), diacylglycerol (DAG), phosphatidylserine (PS), phosphatidylinositol (PI), its polyphosphorylated derivatives (PIPs) and numerous isoprenoid derivatives, i.e. ubiquinone. Another general hallmark of eukaryotic membrane lipids is the frequent presence of long (more than 19 carbons) and multiunsaturated fatty acyl residues. Additionally, there is a striking asymmetry in the distribution of phospholipids between the inner and the outer monolayer of the cytoplasmic membrane. Whereas PC and SL are mainly encountered in the outer monolayer, PE, PS, PA, PI and PIPs are predominantly orientated towards the cytoplasmic side of the membrane (Nelson and Cox 2013).

Phospholipids of bacteria share three key features with eukaryote phospholipids: (1) *sn*-glycerol-3-phosphate as the initial building block for lipid formation, (2) fatty acyl hydrocarbon moieties, that are; (3) linked as esters to the glycerol backbone. Although the model bacterium *Escherichia coli* has PG, CL and PE as major phospholipids in its membranes, it is clear by now that there is no such thing as a typical bacterial membrane lipid composition (Sohlenkamp and Geiger 2016). In addition to these three most common phospholipid classes, bacteria contain less frequently PC and PI, as well as a variety of other membrane lipids, such as ornithine lipids, glycolipids, SL, hopanoids or other isoprenoid-derived lipids among others (Sohlenkamp and Geiger 2016). Moreover, DAG, PA and PS are formed in most bacteria, often as intermediates in metabolic pathways (Geiger et al. 2013). Notably, CHO is absent from bacteria, but some of its functions might be taken over by hopanoids or other polyisoprenoids (Sohlenkamp and Geiger 2016; López and Kolter 2010). Fatty acyl chains encountered in bacteria are frequently saturated, monounsaturated or ramified (*iso*- or *anteiso*-) and seldom exceed 19

carbon units. Additionally, one should keep in mind that information on membrane lipid composition is scarce or absent for most of the bacterial phyla. In contrast, membrane phospholipids of archaea differ from the eukaryotic and bacterial versions in the three key features defining the so-called lipid divide (Koga 2011): (1) *sn*-glycerol-1-phosphate as the initial building block for lipid formation, (2) isoprenoid hydrocarbon moieties, that are; (3) linked as ethers to the glycerol backbone.

The opposite stereoconfiguration of the glycerol-phosphate backbone in archaea and bacteria is due to the action of the non-homologous enzymes, glycerol-1-phosphate dehydrogenase (encoded by *egsA* in archaea) (Daiyasu et al. 2002) and glycerol-3-phosphate dehydrogenase (encoded by *glpD* in bacteria), respectively. The C20 unit of geranylgeranyl pyrophosphate is formed from four isoprenoid units and used by geranylgeranyl glycerophosphate synthase to form geranylgeranyl glycerophosphate. In turn, the second alkylation is performed by bis-geranylgeranyl glycerophosphate synthase also using geranylgeranyl pyrophosphate to convert geranylgeranyl glycerophosphate into 2,3-bis-*O*-geranylgeranyl-*sn*-glyceryl-1-glycerophosphate (archaetidic acid). Again, the two enzymes forming the ether bonds are unique to archaea. Steps that follow the formation of archaeal phospholipids are the activation of CDP-archaeol by *CarS* (Jain et al. 2014a) and head group modification by enzymes of the CDP-alcohol phosphatidyltransferases and by enzymes that are homologues of their bacterial counterparts. As in bacteria, the alcohol head groups linked to archaetidic acid can be glycerol, ethanolamine, serine, choline or inositol. Activation of archaetidic acid and modification reactions with CDP-archaeol still occur with the tetra-unsaturated geranylgeranyl residues (Caforio et al. 2015). Some of these pre-existing double bonds might be used to form cyclopentane or cyclohexane ring moieties, which are encountered in some of the archaeal membrane lipids. Saturation of the isoprenoid double bonds probably occurs after the head group activation (Jain et al. 2014b). Tetraether lipid structures with varying numbers of cyclopentane moieties are widespread in archaea, and they are thought to be formed from saturated diethers via head-to-head condensation reactions (Jain et al. 2014b).

Biochemically, Thaumarchaeota are the best characterized organisms of the TACK superphylum from which the ancestral archaean that then formed the eukaryotic nucleus emerged (Sinninghe Damsté et al. 2002a; Pearson and Ingalls 2013; Dang et al. 2013; Zhalnina et al. 2012). Thaumarchaeota have a specific cyclopentane ring-containing dibiphytanyl glycerol tetraether membrane lipid, crenarchaeol (Sinninghe Damsté et al. 2002a). Although such tetraether lipids are unknown outside archaea, ladderane membrane lipids, isolated from the anammoxosome of Planctomycetes, can contain ester or ether bonds and harbour a similar concatenated cyclobutane rings in their hydrocarbon residues (Sinninghe Damsté et al. 2002b). With regard to some modification pathways of the phospholipids' polar head groups, archaea and bacteria are more similar to each other compared to eukaryotes. For example, in eukarya, PI is synthesized from CDP-diacylglycerol and inositol (Carman and Fischl 1992). In contrast, bacteria use inositol 1-phosphate and CDP-DAG to form phosphatidylinositol phosphate (PIP),

whereas in an analogous reaction archaea condense inositol 1-phosphate to CDP-archaeol, forming archaetidylinositol phosphate (AIP) (Michell 2013; Morii et al. 2014). In this context, it is interesting to note that PI-3-phosphate and other phosphorylated forms of PI play crucial roles in the complex membrane traffic machinery of eukaryotes (Lindmo and Stenmark 2006; Michell 2013) and might do so in archaean ancestors too, given the abundance of membrane trafficking proteins in Lokiarchaean genome (Spang et al. 2015). Subsequently, PIP and AIP are dephosphorylated to become PI and archaetidylinositol, respectively (Morii et al. 2014). Notably, eukaryotic PI synthase, bacterial PIP synthase, archael AIP synthase and PG phosphate synthase are homologues to each other (Morii et al. 2014).

Although the three main features mostly hold true to argue for the lipid divide between bacteria and archaea, there are some recent reports that certainly blur the picture. For example, archaea might be able to perform a simplified fatty acid biosynthesis that is independent of acyl carrier protein (Lombard et al. 2012). Another, possibly dogma-breaking example is provided by a novel pathway for inositol phospholipid synthesis in *Rhodothermus marinus* (Jorge et al. 2015). CDP-inositol and dialkylether glycerol are condensed by dialkylether glycerophosphoinositol synthase (BEPIS) to form *sn*-1,2-dialkylether-glycero-3-phosphoinositol (Jorge et al. 2015).

The limited knowledge of the energy metabolism of the proposed Lokiarchaean ancestor of eukaryotic cells (Spang et al. 2015) together with the uncertainty still surrounding the identity of the alpha proteobacterial lineage for the bacterial ancestors of mitochondria (see part 1 above) leaves space to an alternative scenario for the symbiotic event that originated eukaryogenesis (López-García and Moreira 2015). Namely, a small archaean might have become an endosymbiont of a large proteobacterium that subsequently was infected by alpha proteobacterial parasites which evolved into mitochondria (López-García and Moreira 2015 and references therein). Such a scenario would solve a fundamental problem of the popular alternative of an archaean host then ingesting proto-mitochondria: Eukaryotic cells predominantly have proteobacterial types of phospholipids in all their membranes (Nelson and Cox 2013), not just in mitochondria. Where are the relics of the ancestral archaean lipids in current eukaryotic cells? To answer this fundamental question, this article has presented a survey of the membrane lipids that may derive from either the archaean or the proteobacterial partner of the symbiotic event leading to eukaryogenesis. Based on such a survey, we conclude that presently there are no specific relics of ancestral archaean lipids in current eukaryotic cells. Considering that most data about the lipid composition in eukaryotic membranes have been generated several decades ago, it is well possible that minor lipids or difficult-to-detect lipids might have been missed. Therefore, it probably would be worthwhile to reinvestigate lipid composition from the different eukaryotic membranes by modern mass spectrometric technique searching specifically for lipids that are considered as markers of archaea.

Acknowledgements This work was sponsored by intramural funds at IIT and by CONACyT Grant No. 263876 to EMR for the sabbatical research period of MDE in Mexico.

References

Andrus AD, Andam C, Parker MA (2012) American origin of *Cupriavidus* bacteria associated with invasive *Mimosa* legumes in the Philippines. FEMS Microbiol Ecol 80:747–750. doi:10.1111/j.1574-6941.2012.01342.x

Aoki S, Ito M, Iwasaki W (2013) From β- to α-proteobacteria: the origin and evolution of rhizobial nodulation genes nodIJ. Mol Biol Evol 30:2494–2508. doi:10.1093/molbev/mst153

Berry EA, Huang LS, Saechao LK, Pon NG, Valkova-Valchanova M, Daldal F (2004) X-Ray structure of *Rhodobacter Capsulatus* Cytochrome bc (1): comparison with its mitochondrial and chloroplast counterparts. Photosynth Res 81:251–275

Brinkmann H, Philippe H (2007) The diversity of eukaryotes and the root of the eukaryotic tree. Adv Exp Med Biol 607:20–37

Brochier-Armanet C, Boussau B, Gribaldo S, Forterre P (2007) Mesophilic crenarchaeota: proposal for a third archaeal phylum, the Thaumarchaeota. Nat Rev Microbiol 6:245–252. doi:10.1038/nrmicro1852

Brochier-Armanet C, Gribaldo S, Forterre P (2012) Spotlight on the Thaumarchaeota. ISME J 6:227–230. doi:10.1038/ismej.2011.145

Caforio A, Jain S, Fodran P, Siliakus M, Minnaard AJ, van der Oost J, Driessen AJ (2015) Formation of the ether lipids archaetidylglycerol and archaetidylethanolamine in *Escherichia coli*. Biochem J 470:343–355. doi:10.1042/BJ20150626

Carman GM, Fischl AS (1992) Phosphatidylinositol synthase from yeast. Methods Enzymol 209:305–312

Chaintreuil C, Giraud E, Prin Y, Lorquin J, Bâ A, Gillis M, de Lajudie P, Dreyfus B (2000) Photosynthetic bradyrhizobia are natural endophytes of the African wild rice *Oryza breviligulata*. Appl Environ Microbiol 66:5437–5447

Chi Shen SH, Cheng HP, Jing YX, Yanni YG, Dazzo FB (2005) Ascending migration of endophytic rhizobia, from roots to leaves, inside rice plants and assessment of benefits to rice growth physiology. Appl Environ Microbiol 71:7271–7278

Daims H, Lebedeva EV, Pjevac P, Han P, Herbold C, Albertsen M, Jehmlich N, Palatinszky M, Vierheilig J, Bulaev A, Kirkegaard RH, von Bergen M, Rattei T, Bendinger B, Nielsen PH, Wagner M (2015) Complete nitrification by Nitrospira bacteria. Nature 528:504–509. doi:10.1038/nature16461

Daiyasu H, Hiroike T, Koga Y, Toh H (2002) Analysis of membrane stereochemistry with homology modeling of *sn*-glycerol-1-phosphate dehydrogenase. Protein Eng 15:987–995

Dang H, Zhou H, Yang J, Ge H, Jiao N, Luan X, Zhang C, Klotz MG (2013) Thaumarchaeotal signature gene distribution in sediments of the northern South China Sea: an indicator of the metabolic intersection of the marine carbon, nitrogen, and phosphorus cycles? Appl Environ Microbiol 79:2137–2147. doi:10.1128/AEM.03204-12

Degli Esposti M (2014) Bioenergetic evolution in proteobacteria and mitochondria. Genome Biol Evol 6:3238–3251. doi:10.1093/gbe/evu257

Degli Esposti M, Chouaia B, Comandatore F, Crotti E, Sassera D, Lievens PM, Daffonchio D, Bandi C (2014) Evolution of mitochondria reconstructed from the energy metabolism of living bacteria. PLoS ONE 9:e96566. doi:10.1371/journal.pone.0096566

Degli Esposti M, Rosas-Pérez T, Servín-Garcidueñas LE, Bolaños LM, Rosenblueth M, Martínez-Romero E (2015) Molecular evolution of cytochrome bd oxidases across proteobacterial genomes. Genome Biol Evol 7:801–820. doi:10.1093/gbe/evv032

Desjardins CA, Champion MD, Holder JW, Muszewska A, Goldberg J, Bailão AM, Brigido MM, Ferreira ME, Garcia AM, Grynberg M, Gujja S, Heiman DI, Henn MR, Kodira CD, León-Narváez H, Longo LV, Ma LJ, Malavazi I, Matsuo AL, Morais FV, Pereira M, Rodríguez-Brito S, Sakthikumar S, Salem-Izacc SM, Sykes SM, Teixeira MM, Vallejo MC, Walter ME, Yandava C, Young S, Zeng Q, Zucker J, Felipe MS, Goldman GH, Haas BJ, McEwen JG, Nino-Vega G, Puccia R, San-Blas G, Soares CM, Birren BW, Cuomo CA (2011)

Comparative genomic analysis of human fungal pathogens causing paracoccidioidomycosis. PLoS Genet 7:e1002345. doi:10.1371/journal.pgen.1002345

Doty SL, Dosher MR, Singleton GL, Moore AL, Aken B, van Stettler RF, Gordon MP (2005) Identification of an endophytic Rhizobium in stems of Populus. Symbiosis 39:27–35

Finnegan PM, Umbach AL, Wilce JA (2003) Prokaryotic origins for the mitochondrial alternative oxidase and plastid terminal oxidase nuclear genes. FEBS Lett 555:425–430

Gagliano AL, Tagliavia M, D'Alessandro W, Franzetti A, Parello F, Quatrini P (2016) So close, so different: geothermal flux shapes divergent soil microbial communities at neighbouring sites. Geobiology 14:150–62

Geiger O, López-Lara IM, Sohlenkamp C (2013) Phosphatidylcholine biosynthesis and function in bacteria. Biochim Biophys Acta 1831:503–513. doi:10.1016/j.bbalip.2012.08.009

Gray MW (2015) Mosaic nature of the mitochondrial proteome: implications for the origin and evolution of mitochondria. Proc Natl Acad Sci USA 112:10133–10138. doi:10.1073/pnas.1421379112

Guy L, Ettema TJ (2011) The archaeal 'TACK' superphylum and the origin of eukaryotes. Trends Microbiol 19:580–587. doi:10.1016/j.tim.2011.09.002

Gyaneshwar P, Hirsch AM, Moulin L, Chen WM, Elliott GN, Bontemps C, Estrada-de Los Santos P, Gross E, Dos Reis FB, Sprent JI, Young JP, James EK (2011) Legume-nodulating betaproteobacteria: diversity, host range, and future prospects. Mol Plant Microbe Interact 24:1276–1288. doi:10.1094/MPMI-06-11-0172

Hatzenpichler R (2012) Diversity, physiology, and niche differentiation of ammonia-oxidizing archaea. Appl Environ Microbiol 78:7501–7510. doi:10.1128/AEM.01960-12

Hawley AK, Brewer HM, Norbeck AD, Paša-Tolić L, Hallam SJ (2014) Metaproteomics reveals differential modes of metabolic coupling among ubiquitous oxygen minimum zone microbes. Proc Natl Acad Sci USA 111:11395–11400. doi:10.1073/pnas.1322132111

Jain S, Caforio A, Fodran P, Lolkema JS, Minnaard AJ, Driessen AJ (2014a) Identification of CDP-archaeol synthase, a missing link of ether lipid biosynthesis in Archaea. Chem Biol 21:1392–1401. doi:10.1016/j.chembiol.2014.07.022

Jain S, Caforio A, Driessen AJ (2014b) Biosynthesis of archaeal membrane ether lipids. Front Microbiol 5:641. doi:10.3389/fmicb.2014.00641

Johnstone IL, McCabe PC, Greaves P, Gurr SJ, Cole GE, Brow MA, Unkles SE, Clutterbuck AJ, Kinghorn JR, Innis MA (1990) Isolation and characterisation of the crnA-niiA-niaD gene cluster for nitrate assimilation in *Aspergillus nidulans*. Gene 90:181–192

Jorge CD, Borges N, Santos H (2015) A novel pathway for the synthesis of inositol phospholipids uses cytidine diphosphate (CDP)-inositol as donor of the polar head group. Environ Microbiol 17:2492–2504. doi:10.1111/1462-2920.12734

Kneip C, Lockhart P, Voss C, Maier UG (2007) Nitrogen fixation in eukaryotes–new models for symbiosis. BMC Evol Biol 7:55

Koga Y (2011) Early evolution of membrane lipids: how did the lipid divide occur? J Mol Evol 72:274–82. doi: 10.1007/s00239-011-9428-5

Kondorosi E, Mergaert P, Kereszt A (2013) A paradigm for endosymbiotic life: cell differentiation of Rhizobium bacteria provoked by host plant factors. Annu Rev Microbiol 67:611–628. doi:10.1146/annurev-micro-092412-155630

Koonin EV (2015) Origin of eukaryotes from within archaea, archaeal eukaryome and bursts of gene gain: eukaryogenesis just made easier? Philos Trans R Soc Lond B Biol Sci 370:20140333. doi:10.1098/rstb.2014.0333

Kouzuma A, Kato S, Watanabe K (2015) Microbial interspecies interactions: recent findings in syntrophic consortia. Front Microbiol 6:477. doi:10.3389/fmicb.2015.00477

Ku C, Nelson-Sathi S, Roettger M, Sousa FL, Lockhart PJ, Bryant D, Hazkani-Covo E, McInerney JO, Landan G, Martin WF (2016) Endosymbiotic origin and differential loss of eukaryotic genes. Nature 524:427–432. doi:10.1038/nature14963

Kumar N, Lad G, Giuntini E, Kaye ME, Udomwong P, Shamsani NJ, Young JP, Bailly X (2015) Bacterial genospecies that are not ecologically coherent: population genomics of Rhizobium leguminosarum. Open Biol 5:140133. doi:10.1098/rsob.140133

Lin JT, Goldman BS, Stewart V (1993) Structures of genes nasA and nasB, encoding assimilatory nitrate and nitrite reductases in Klebsiella pneumoniae M5al. J Bacteriol 175:2370–2378

Lindmo K, Stenmark H (2006) Regulation of membrane traffic by phosphoinositide 3-kinases. J Cell Sci 119:605–614

Lombard J, López-García P, Moreira D (2012) An ACP-independent fatty acid synthesis pathway in archaea: implications for the origin of phospholipids. Mol Biol Evol 29:3261–3265. doi:10.1093/molbev/mss160

López D, Kolter R (2010) Functional microdomains in bacterial membranes. Genes Dev 24:1893–1902. doi:10.1101/gad.1945010

López-García P, Moreira D (2015) Open questions on the origin of eukaryotes. Trends Ecol Evol 30:697–708. doi:10.1016/j.tree.2015

López-Guerrero M, Ormeño-Orrillo E, Acosta JL, Mendoza-Vargas A, Rogel MA, Ramírez MA, Rosenblueth M, Martínez-Romero J, Martínez-Romero E (2012) Rhizobial extrachromosomal replicon variability, stability and expression in natural niches. Plasmid 68:149–158. doi:10.1016/j.plasmid.2012.07.002

López-Guerrero M, Ormeño-Orrillo E, Rosenblueth M, Martínez J, Martínez-Romero E (2013) Buffet hypothesis for microbial nutrition at the rhizosphere. Front Plant Sci 4:1–4

Maillet F, Poinsot V, André O, Puech-Pagès V, Haouy A, Gueunier M, Cromer L, Giraudet D, Formey D, Niebel A, Martinez EA, Driguez H, Bécard G, Dénarié J (2011) Fungal lipochitooligosaccharide symbiotic signals in arbuscular mycorrhiza. Nature 469:58–63. doi:10.1038/nature09622

Mardanov AV, Beletsky AV, Kadnikov VV, Ignatov AN, Ravin NV (2014) Draft genome sequence of *sclerotinia borealis*, a psychrophilic plant pathogenic fungus. Genome Announc 2. pii: e01175-13. doi:10.1128/genomeA.01175-13

Martin WF, Garg S, Zimorski V (2015) Endosymbiotic theories for eukaryote origin. Philos Trans R Soc Lond B Biol Sci 370:20140330. doi:10.1098/rstb.2014.0330

Martinez D, Berka RM, Henrissat B, Saloheimo M, Arvas M, Baker SE, Chapman J, Chertkov O, Coutinho PM, Cullen D, Danchin EG, Grigoriev IV, Harris P, Jackson M, Kubicek CP, Han CS, Ho I, Larrondo LF, de Leon AL, Magnuson JK, Merino S, Misra M, Nelson B, Putnam N, Robbertse B, Salamov AA, Schmoll M, Terry A, Thayer N, Westerholm-Parvinen A, Schoch CL, Yao J, Barabote R, Nelson MA, Detter C, Bruce D, Kuske CR, Xie G, Richardson P, Rokhsar DS, Lucas SM, Rubin EM, Dunn-Coleman N, Ward M, Brettin TS (2008) Genome sequencing and analysis of the biomass-degrading fungus *Trichoderma reesei* (syn. *Hypocrea jecorina*). Nat Biotechnol 26:553–560. doi:10.1038/nbt1403

Martínez L, Caballero-Mellado J, Orozco J, Martínez-Romero E (2003) Diazotrophic bacteria associated with banana (*Musa* spp.). Plant Soil 257:35–47

Martinez-Romero E (2012) How do microbes enhance the carrying capacity of their habitats? Expert Opin Environ Biol 1:1–2

Martínez-Romero E (2009) Coevolution in Rhizobium-legume symbiosis? DNA Cell Biol 28:361–370. doi:10.1089/dna.2009.0863

Martínez-Romero JC, Ormeño-Orrillo E, Rogel-Hernández MA, López-López A, Martínez-Romero E (2010) Trends in rhizobial evolution and some taxonomic remarks. In: Pontarotti PT (ed) Evolutionary biology—concepts, molecular and morphological evolution. Springer, Berlin, pp 301–315

Michell RH (2013) Inositol lipids: from an archaeal origin to phosphatidylinositol 3,5-bisphosphate faults in human disease. FEBS J 280:6281–6294. doi:10.1111/febs.12452

Miranda-Sánchez F, Rivera J, Vinuesa P (2015) Diversity patterns of Rhizobiaceae communities inhabiting soils, root surfaces and nodules reveal a strong selection of rhizobial partners by legumes. Environ Microbiol. doi: 10.1111/1462-2920.13061 (Epub ahead of print)

Moreno-Vivián C, Ferguson SJ (1998) Definition and distinction between assimilatory, dissimilatory and respiratory pathways. Mol Microbiol 29:664–666

Morii H, Ogawa M, Fukuda K, Taniguchi H (2014) Ubiquitous distribution of phosphatidylinositol phosphate synthase and archaetidylinositol phosphate synthase in Bacteria and Archaea,

which contain inositol phospholipid. Biochem Biophys Res Commun 443:86–90. doi:10.1016/j.bbrc.2013.11.054

Moulin L, Munive A, Dreyfus B, Boivin-Masson C (2011) Nodulation of legumes by members of the beta-subclass of Proteobacteria. Nature 411:948–950

Muller F, Brissac T, Le Bris N, Felbeck H, Gros O (2012) First description of giant Archaea (*Thaumarchaeota*) associated with putative bacterial ectosymbionts in a sulfidic marine habitat. Environ Microbiol 12:2371–2383. doi:10.1111/j.1462-2920.2010.02309.x

Müller M, Mentel M, van Hellemond JJ, Henze K, Woehle C, Gould SB, Yu RY, van der Giezen M, Tielens AG, Martin WF (2012) Biochemistry and evolution of anaerobic energy metabolism in eukaryotes. Microbiol Mol Biol Rev 76:444–495. doi:10.1128/MMBR.05024-11

Nelson DL, Cox MM (2013) Lehninger: principles of biochemistry, 6th edn. WH Freeman and Company, New York

Oldroyd GE, Downie JA (2008) Coordinating nodule morphogenesis with rhizobial infection in legumes. Annu Rev Plant Biol 59:519–546. doi:10.1146/annurev.arplant.59.032607.092839

Ormeño-Orrillo E, Hungria M, Martínez-Romero E (2013) Dinitrogen-fixing prokaryotes. In: Rosenberg E et al (eds) The prokaryotes, prokaryotic physiology and biochemistry. Springer, Heidelberg, pp 427–451. ISBN: 978-3-642-30140-7

Ormeño-Orrillo E, Servín-Garcidueñas LE, Rogel MA, González V, Peralta H, Mora J, Martínez-Romero J, Martínez-Romero E (2015) Taxonomy of rhizobia and agrobacteria from the Rhizobiaceae family in light of genomics. Syst Appl Microbiol 38:287–291. doi:10.1016/j.syapm.2014.12.002

Pearson A, Ingalls AE (2013) Assessing the use of archaeal lipids as marine environmental proxies. Annu Rev Earth Planet Sci 41:359–384. doi:10.1146/annurev-earth-050212-123947

Persson T, Battenberg K, Demina IV, Vigil-Stenman T, Vanden Heuvel B, Pujic P, Facciotti MT, Wilbanks EG, O'Brien A, Fournier P, Cruz Hernandez MA, Mendoza Herrera A, Médigue C, Normand P, Pawlowski K, Berry AM (2015) Candidatus Frankia Datiscae Dg1, the Actinobacterial microsymbiont of Datisca glomerata, expresses the canonical nod genes nodABC in symbiosis with its host plant. PLoS ONE 10:e0127630. doi:10.1371/journal.pone.0127630

Pittis AA, Gabaldón T (2016) Late acquisition of mitochondria by a host with chimaeric prokaryotic ancestry. Nature 531:101–114

Prechtl J, Kneip C, Lockhart P, Wenderoth K, Maier UG (2004) Intracellular spheroid bodies of Rhopalodia gibba have nitrogen-fixing apparatus of cyanobacterial origin. Mol Biol Evol 21:1477–1481

Ramachandran VK, East AK, Karunakaran R, Downie JA, Poole PS (2011) Adaptation of Rhizobium leguminosarum to pea, alfalfa and sugar beet rhizospheres investigated by comparative transcriptomics. Genome Biol 12:R106. doi:10.1186/gb-2011-12-10-r106

Raymann K, Brochier-Armanet C, Gribaldo S (2015) The two-domain tree of life is linked to a new root for the Archaea. Proc Natl Acad Sci USA 112:6670–6675. doi:10.1073/pnas.1420858112

Ridgway ND (2016) Phospholipid synthesis in mammalian cells. In: Ridgway N, McLeod R (eds) Biochemistry of lipids, lipoproteins and membranes, 6th edn. Elsevier, Amsterdam, pp 209–236

Rodríguez-Ezpeleta N, Embley TM (2012) The SAR11 group of alpha-proteobacteria is not related to the origin of mitochondria. PLoS ONE 7:e30520. doi:10.1371/journal.pone.0030520

Sawana A, Adeolu M, Gupta RS (2013) Molecular signatures and phylogenomic analysis of the genus Burkholderia: proposal for division of this genus into the emended genus Burkholderia containing pathogenic organisms and a new genus Paraburkholderia gen. nov. harboring environmental species. Front Genet 5:429. doi:10.3389/fgene.2014.00429

Schweizer H, Larson TJ (1987) Cloning and characterization of the aerobic sn-glycerol-3-phosphate dehydrogenase structural gene glpD in Escherichia coli K-12. J Bacteriol 169:507–513

Searcy DG (2003) Metabolic integration during the evolutionary origin of mitochondria. Cell Res 13:229–238

Sinninghe Damsté JS, Schouten S, Hopmans EC, van Duin AC, Geenevasen JA (2002a) Crenarchaeol: the characteristic core glycerol dibiphytanyl glycerol tetraether membrane lipid of cosmopolitan pelagic crenarchaeota. J Lipid Res 43:1641–1651. doi:10.1194/jlr.m200148-jlr200

Sinninghe Damsté JS, Strous M, Rijpstra WI, Hopmans EC, Geenevasen JA, van Duin AC, van Niftrik LA, Jetten MS (2002b) Linearly concatenated cyclobutane lipids form a dense bacterial membrane. Nature 419:708–712

Sohlenkamp C, Geiger O (2016) Bacterial membrane lipids: diversity in structures and pathways. FEMS Microbiol Rev 40:133–159. doi:10.1093/femsre/fuv008

Spang A, Saw JH, Jørgensen SL, Zaremba-Niedzwiedzka K, Martijn J, Lind AE, van Eijk R, Schleper C, Guy L, Ettema TJ (2015) Complex archaea that bridge the gap between prokaryotes and eukaryotes. Nature 521:173–179. doi:10.1038/nature14447

Takasaki K, Shoun H, Nakamura A, Hoshino T, Takaya N (2004) Unusual transcription regulation of the niaD gene under anaerobic conditions supporting fungal ammonia fermentation. Biosci Biotechnol Biochem 68:978–980

Testa A, Oliver R, Hane J (2015) Overview of genomic and bioinformatic resources for *Zymoseptoria tritici*. Fungal Genet Biol 79:13–16. doi:10.1016/j.fgb.2015.04.011

Tetu SG, Breakwell K, Elbourne LD, Holmes AJ, Gillings MR, Paulsen IT (2013) Life in the dark: metagenomic evidence that a microbial slime community is driven by inorganic nitrogen metabolism. ISME J 7:1227–1236. doi:10.1038/ismej.2013.14

Woese CR, Fox GE (1977) Phylogenetic structure of the prokaryotic domain: the primary kingdoms. Proc Natl Acad Sci USA 74:5088–5090

Xavier BB, Miao VP, Jónsson ZO, Andrésson ÓS (2012) Mitochondrial genomes from the lichenized fungi Peltigera membranacea and Peltigera malacea: features and phylogeny. Fungal Biol 116:802–814. doi:10.1016/j.funbio.2012.04.013

Zhalnina KV, Dias R, Leonard MT, Dorr de Quadros P, Camargo FA, Drew JC, Farmerie WG, Daroub SH, Triplett EW (2012) Genome sequence of *Candidatus* Nitrososphaera evergladensis from group I.1b enriched from Everglades soil reveals novel genomic features of the ammonia-oxidizing archaea. PLoS One 9:e101648. doi:10.1371/journal.pone.0101648

Chapter 13
Genomic Analysis of Bacterial Outbreaks

Leonor Sánchez-Busó, Iñaki Comas, Beatriz Beamud, Neris García-González, Marta Pla-Díaz and Fernando González-Candelas

Abstract The study of outbreaks of infectious diseases has been revolutionized by the current availability of fast and efficient, high-throughput methods capable of yielding the nucleotide sequence of complete genomes of viruses and bacteria within a few days, or even hours. These methods are replacing previous molecular techniques which have been used for the past 30 years, although many of them are still the usual approach for many such investigations. Here we review the major technologies currently in use for high-throughput sequencing of bacterial genomes emphasizing their advantages and drawbacks for the analysis of outbreaks. The use of more efficient methods does not necessarily mean that all the problems in the study of outbreaks are automatically solved. In fact, because these methods are capable of revealing genetic variation at an unprecedented level and scale they also pose new challenges for the interpretation of the resulting data. We analyze some of these new challenges, especially those in which the short term evolution of microorganism plays an important role in understanding and interpreting the results and, quite often, in reconciling them with those derived from classical epidemio-

L. Sánchez-Busó · I. Comas · B. Beamud · N. García-González · M. Pla-Díaz · F. González-Candelas
Unidad Mixta "Infección y Salud Pública" FISABIO/CSISP—Universidad de Valencia/Instituto Cavanilles de Biodiversidad y Biología Evolutiva, Valencia, Spain

I. Comas · F. González-Candelas
CIBER en Epidemiología y Salud Pública, Valencia, Spain

L. Sánchez-Busó
Pathogen Genomics, The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Cambridgeshire, UK

I. Comas
Instituto de Biomedicina, CSCIC, Valencia, Spain

F. González-Candelas (✉)
Evolución y Salud—Instituto Cavanilles Biodiversidad y Biología Evolutiva, Universidad de Valencia, c/Catedrático José Beltrán 2, 46980 Valencia, Paterna, Spain
e-mail: fernando.gonzalez@uv.es

P. Pontarotti (ed.), *Evolutionary Biology*, DOI 10.1007/978-3-319-41324-2_13

logical analyses. Finally, we exemplify the advantages of using complete genome analysis with two cases, involving outbreaks of *Mycobacterium tuberculosis* and *Legionella pneumophila*.

13.1 Introduction

Outbreaks of infectious diseases often produce social alarms. These can be very local or reach every corner of every village and city on Earth. But all they share a need for a quick control and remediation that ensures the safety of the population. The identification and control of the source of an outbreak becomes a health priority, and many efforts are devoted to these activities in the first days and weeks after the detection and/or declaration of an outbreak (Mortimer 2003).

Outbreaks come in many shapes and flavors. For epidemiologists, an outbreak is simply an unusual increase in the prevalence of a disease in time and space. Hence, some outbreaks may be declared and last for years, while others are reduced to a few days or weeks; similarly, there might be an outbreak in a school or nursing home, but we talked a few years ago about an epidemic outbreak of "swine influenza" (Fraser et al. 2009; General Directorate of Epidemiology et al. 2009), and the WHO and other health organizations are currently worried about the spread of Zika virus. In some cases, the spread of the infectious pathogen occurs in a series of successive infections from one host to another, thus producing transmission chains or networks, depending on the topology of the resulting connections among infected persons.

One of the first tasks when an outbreak is suspected is to establish the basic parameters for controlling it. This can depend on the detection of a source, and the application of actions that prevent it from spreading the pathogen, or the characterization of the vector, so it can be controlled with chemical or biological agents, or the identification of the hereditary factors that allow the pathogen eluding previous, successful treatments and originate nosocomial outbreaks of multi-resistant strains. The advent of faster and cheaper gene sequencing techniques leads to the first systematic and general proposal of using a universal typing scheme that was reproducible, cheap, objective, and easily exchangeable among laboratories, known as multi-locus sequence typing (MLST) (Maiden et al. 1998). In this method, the nucleotide sequence of 6–7 loci is determined and used to derive an array of allele profiles in these loci. A new combination of allele profiles corresponds to a new sequence type which is uploaded to a Web server for easy access. Typing schemes, with detailed laboratory protocols, proficiency tests, and full information on identified sequences types are available for tens of bacterial species in general and specific Web servers (see, for instance, http://www.PubMLST.org).

For many pathogens, the availability of a MLST scheme represented a more than significant change in the analysis of outbreaks. This method quickly became the new "gold-standard" for typing pathogens and replaced previous methods. However, for a few but important pathogens, no MLST scheme revealing enough genetic variation for effectively distinguishing among non-epidemiologically linked isolates could be

designed. These pathogens include the causative agents of plague (*Yersinia pestis*), anthrax (*Bacillus anthracis*), tuberculosis (*Mycobacterium tuberculosis*), and leprosy (*Mycobacterium leprae*), among others, and are collectively known as "genetically monomorphic bacteria" (Achtman 2012). Specific typing methods such as insertion sequence RFLP and MIRU-VNTR were applied to *M. tuberculosis*, the pathogenic bacteria with the highest incidence and causing more deaths every year in the history of humankind. In these and other cases, the solutions adopted relied on very fast-evolving markers, which are usually prone to homoplastic changes, thus resulting in some false-positive identifications of phenotypic identities as indicative of very recent ancestry. Although this is not a problem in most settings, it became evident that the same logic applied in using MLST could be extended to the complete genome sequences to attain "perfect" accuracy by using all the genetic information in the isolates and not only a small sample from it.

This approach was first used in an outbreak setting in the investigation of the letters covered with anthrax spores in the aftermath of the 9/11 attacks in the USA. Complete genome sequences were obtained from a *B. anthracis* isolate derived from one of the victims and one reference strain, providing 60 SNPs that could be used subsequently to probe the common origin of the strain used in the bioterrorist attacks (Read et al. 2002). This work clearly showed that using the complete genome sequence was a more effective method for comparing isolates even in almost completely monomorphic species. However, Sanger sequencing is rather slow and painstaking as a result of the need to cut or amplify the genome in small pieces that are subsequently sequenced and assembled into a complete genome sequence. This situation changed dramatically with the introduction of new sequencing methods, then known as "next-generation sequencing" technologies. They offered several advantages over the traditional Sanger method (Medini et al. 2008). At the same time, other problems arose, such as the difficulties in handling and analyzing very large volumes of data, a myriad of programs and methods to analyze them, and new conceptual challenges in the interpretation of the results.

In this chapter, we provide a brief overview of the different next-generation sequencing platforms and methods currently available for deriving complete genome sequences from bacteria, the main results in terms of the epidemiological and evolutionary advances that have resulted from their application to bacterial outbreaks and transmission networks, and provide a more detailed analysis of two cases: the analysis of *Legionella pneumophila* outbreaks and of *M. tuberculosis* transmission networks.

13.2 High-throughput Sequencing Technologies in Outbreak Investigations

Several high-throughput sequencing platforms have been applied to the genomic study of both bacterial and virus pathogens. Encouraged by the increasing need of sequencing human genomes, three technologies were almost simultaneously

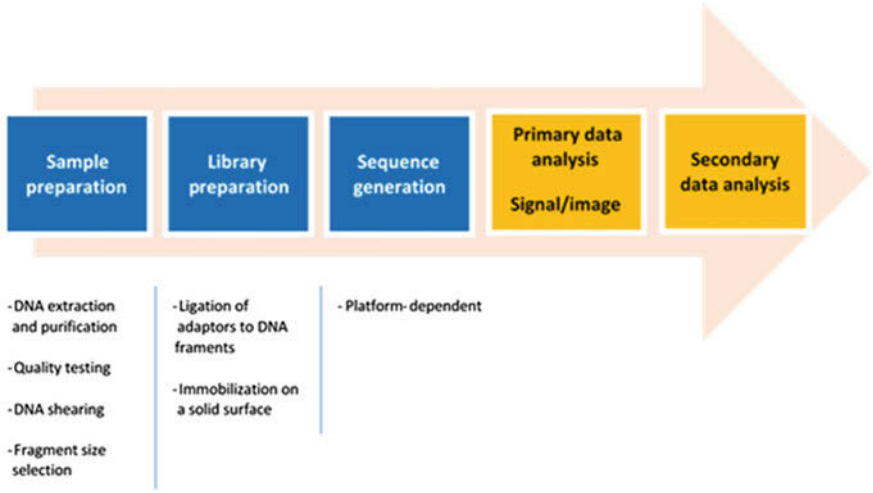


Fig. 13.1 General workflow followed during high-throughput sequencing (Metzker 2010)

released from different companies: 454 (Roche, introduced in 2005 and discontinued in 2016), Solexa (Illumina, introduced in 2006), and SOLiD (Life Technologies, introduced in 2006). These platforms share a general workflow, based on the idea of performing billions of sequencing reactions simultaneously. These are produced through molecular amplification of DNA fragments that are previously attached to a solid surface. These have been enhanced in their subsequent updates to increase both sequencing quality and throughput (Fig. 13.1).

Although 454 was the first released platform, its use has mainly been relegated to metagenomic studies (Schlüter et al. 2008a, b; Ghai et al. 2010) because of its long reads and relatively high error rates, which complicates the study of transmission chains or related cases during outbreak investigations. However, it has been used as the main technology in several studies (Lewis et al. 2010; Kennemann et al. 2011; Loman and Constantinidou 2013) and also following mixed strategies involving the usage of 454 reads as scaffolds and posterior error correction using Illumina (McAdam et al. 2012; Hasan et al. 2012). SOLiD has been the least used for outbreak investigations due to shorter and lower quality reads. As an example, it has been punctually applied in the investigation of *L. pneumophila* outbreaks in an endemic locality in Spain (Sánchez-Busó et al. 2014), *Mycobacterium abscessus* subsp. *bolletii* in Brazil and UK outbreaks (Davidson et al. 2013), or *Coccidioides immitis* producing coccidioidomycosis in transplanted patients in Los Angeles (Engelthaler et al. 2011). By far, Illumina has been the most widely used platform because of its high quality and sensible sized reads, which allow more accurate mapping and SNP calling. A thorough summary of the application of different sequencing technologies to analyze different mainly bacterial outbreaks is shown in Table 13.1.

Table 13.1 A summary of published works analyzing complete genome sequences of bacterial pathogens for the study of outbreaks and transmission chains

Pathogen	Genome size (Mb)	Sequencing strategy	References
Acinetobacter baumannii	4.11	Illumina HiSeq 2000 2 × 100 bp	Lewis et al. (2010), Hornsey et al. (2011), Kanamori et al. (2016), Fitzpatrick et al. (2016)
		Illumina MiSeq 2 × 150 bp, 2 × 250 bp	
		Roche 454 GS FLX	
		PacBio	
Bacillus anthracis	5.2	Sanger	Read et al. (2002)
Campylobacter jejuni	1.64	Illumina HiSeq 2000 76 bp	Cody et al. (2013)
Chlamydia trachomatis	≈1.0	Illumina GA	Harris et al. (2012), Seth-Smith et al. (2013)
		Illumina GAII PE 2 × 37 bp	
Clostridium difficile	4.0	Illumina GAII/GAIIx 2 × 51/100–108 bp	Didelot et al. (2012a), Eyre et al. (2012, 2013a, b), He et al. (2013), Knetsch et al. (2014)
		Illumina HiSeq 2000 2 × 100 bp; 2 × 54/108/76 bp	
Enterobacter cloacae	5.31	Illumina MiSeq	Reuter et al. (2013a)
Enterococcus faecium	2.9	Illumina MiSeq	Reuter et al. (2013a), Pinholt et al. (2015)
Escherichia coli	≈5.2	Roche 454 GS Junior	Mellmann et al. (2011), Brzuszkiewicz et al. (2011), Ahmed et al. (2012), Ju et al. (2012), Grad et al. (2012, 2013), Underwood et al. (2013b), Shah et al. (2014), Holmes et al. (2015)
		Roche 454 Titanium	
		Illumina MiSeq	
		Illumina Solexa	
		Illumina HiSeq 2000 2 × 101	
		Ilumina GAIIx	
		Ion Torrent PGM	
Helicobacter pylori	1.5–1.7	Roche 454	Kennemann et al. (2011)
Klebsiella pneumoniae	5.6	Illumina Hi Seq 2000	Snitkin et al. (2012), Espedido et al. (2013), Onori et al. (2015)
		Illumina MiSeq platform	
		Roche 454 Titanium XLR	

(continued)

Table 13.1 (continued)

Pathogen	Genome size (Mb)	Sequencing strategy	References
Legionella pneumophila	3.5	Illumina HiSeq 2 × 100 bp	Reuter et al. (2013a, b), Sánchez-Busó et al. (2014), Bartley et al. (2016)
		Illumina MiSeq 2 × 250 bp, 2 × 150 bp	
		SOLiD 5500XL SE 75 bp	
Listeria monocytogenes	3	Roche 454 GS FLX	Gilmour et al. (2010), Schmid et al. (2014), Kwong et al. (2016)
Mycobacterium abscessus	5–5.2	Illumina HiSeq 2 × 75 bp	Bryant et al. (2013b)
M. abscessus subsp. bolletii	≈5	Life Technologies SOLiD	Davidson et al. (2013)
Mycobacterium canettii	≈4.5	HiSeq 2000 MiSeq Illumina	Blouin et al. (2014)
Mycobacterium tuberculosis	4.4	Illumina GAII PE 2 × 36 bp; 2 × 50	Ioerger et al. (2010), Schürch et al. (2010), Gardy et al. (2011), Sandegren et al. (2011), Casali et al. (2012), Kato-Maeda et al. (2013), Bryant et al. (2013a), Roetzer et al. (2013), Köser et al. (2013), Török et al. (2013), Walker et al. (2013a, b), Pérez-Lago et al. (2014), Coscollá et al. (2015)
		Illumina GAIIx PE 2 × 76; 2 × 108	
		Illumina HiSeq PE 2 × 75 bp	
		Illumina MiSeq 150 bp	
		Roche 454 GS FLX 36 bp	
Neisseria gonorrhoeae	2.1	Illumina HiSeq	Grad et al. (2014)
Neisseria meningitidis	2.2	Illumina GAIIx PE 2 × 76	Jolley et al. (2012), Reuter et al. (2013a), Bennett et al. (2012)
		Illumina MiSeq	
Pseudomonas aeruginosa	6.26	Ion Torrent	Witney et al. (2014), Snyder et al. (2013)
Salmonella enterica	4.76	Illumina MiSeq	Holt et al. (2008), Lienau et al. (2011), Quick et al. (2015), Allard et al. (2012, 2013), Cao et al. (2013), Taylor et al. (2015), Bekal et al. (2016)
		Illumina HiSeq 2500	
		MinION	
		Roche 454	
Salmonella Typhimurium	4.7	Illumina GA II system	Okoro et al. (2012)
Shigella sonnei	5.06	Illumina GAII PE 2 × 54 bp	Holt et al. (2012, 2013), McDonnell et al. (2013)
		Illumina MiSeq	
		Illumina HiSeq 2000	

(continued)

Table 13.1 (continued)

Pathogen	Genome size (Mb)	Sequencing strategy	References
Staphylococcus aureus	2.8–3	Illumina MiSeq PE 2 × 150 bp	Harris et al. (2010, 2013), Eyre et al. (2012), McAdam et al. (2012), Young et al. (2012), Köser et al. (2012), Holden et al. (2013), Nübel et al. (2013), Price et al. (2014), Azarian et al. (2015), Paterson et al. (2015), Senn et al. (2016), Kinnevey et al. (2016), Reuter et al. (2016)
		Illumina GAIIx PE	
		Illumina GAII SE 150 bp	
		Illumina HiSeq 2000	
		Roche 454 GS FLX	
Streptococcus pneumoniae	1.98–2.19	Illumina HiSeq 2000 2 × 75 bp	Croucher et al. (2011, 2013), Loman et al. (2013), Chewapreecha et al. (2014)
		Illumina PE 2 × 54 bp	
		Roche 454	
Streptococcus pyogenes	1.85	Illumina HiSeq 2000	Zakour et al. (2012), Fittipaldi et al. (2013)
		Illumina GA1s	
		Roche 454	
Streptococcus suis	2.15	Roche 454 /GS 20	Holden et al. (2009)
		Solexa	
Vibrio cholerae	≈4	Illumina HiSeq	Mutreja et al. (2011), Hendriksen et al. (2011), Chin et al. (2011), Hasan et al. (2012), Shah et al. (2014), Schmid et al. (2014), Devault et al. (2014), Wagner et al. (2014), Knetsch et al. (2014)
		Illumina GAI	
		Illumina GAIIx	
		PacBio RS	
		Roche 454 GS FLX	
Yersinia pestis	5.46	Illumina	Cui et al. (2013), Wagner et al. (2014)

In 2010, the Ion Torrent (Life Technologies) platform, a new benchtop device with a different sequencing strategy was commercialized. This technology is based on monitoring pH changes in multi-well plates. A single reaction occurs per well so that when a hydrogen atom is released after the incorporation of each nucleotide during amplification, the pH in the media changes in a nucleotide-specific manner, so that the system is able to translate chemical into digital information. Reads produced by the Ion Torrent were of relatively good quality and were punctually applied to the study of *Escherichia coli* outbreaks (Mellmann et al. 2011; Holmes et al. 2015) and *Pseudomonas aeruginosa* (Snyder et al. 2013; Witney et al. 2014).

In early 2011, the PacBio RS system was also released, being the first platform performing Single Molecule Real Time (SMRT) sequencing, which is being increasingly applied to complete microbial genomes because of the long read lengths (Mutreja et al. 2011). But the definite current revolution in sequencing technologies with an impact in public health has been the release of the Oxford Nanopore MinION platform, currently in test mode, and scalable in the form of the

GridION platform. These contain a membrane with millions of embedded nanopores coupled with a polymerase. Changes in the electrical conductivity in the membrane as the different four bases pass through the nanopore are measured, allowing sequencing in real time. Specifically, the MinION platform is an USB-like device which can be connected directly to a computer and provides the sequences from extracted DNA in real time after a very simple library preparation. The portable MinION platform has been shown to be useful in real-time outbreak investigations, such as the 2015 Ebola virus disease epidemic in West Africa (Quick et al. 2016).

The different platforms differ in their sequencing strategy, which yields different throughputs and sequence qualities. Currently, the highest throughput can be achieved with the HiSeq X Ten Illumina platform, which can yield up to 3 billion of paired-end 150-bp sequences. This high-level throughput is mainly directed to population-scale human genome sequencing projects. In the case of microorganism sequencing, because their genomes are much smaller, sequencing throughput must depend on the depth of coverage required for each specific study. However, large-scale microbial sequencing projects can benefit from these high-throughput platforms by multiplexing different strains in the same run. Coverage depths of 50X–100X are usually sought for base call error correction, minimizing the rate of false-positive SNPs. Currently, the technologies with the lowest error rates are Illumina platforms, and the highest error rate from raw data is provided by Oxford Nanopore and PacBio platforms. However, bioinformatics pipelines for error correction during the post-processing of reads improve these rates, especially in the second case, in which the current final error rate can get as low as 1E-05. Multiple reviews on the characteristics of the different sequencing technologies, applications, advantages, and drawbacks have been published in the literature up to now (Metzker 2010; Casey et al. 2013; Ekblom and Wolf 2014).

Choosing the most appropriate sequencing technology depends on the scope of the study. High-throughput technologies can be applied in different steps during an outbreak investigation (Köser et al. 2012); from the detection and identification of the pathogen in direct uncultured samples (i.e., blood, sputum, etc.), epidemiological typing and detection of mutations associated with drug susceptibility to the study of transmission chains and potential super-spreaders.

13.3 Achievements and Limitations of NGS in Outbreak Investigations

Initial results. Although NGS techniques and devices became available around 2005 (Loman and Pallen 2015), it took a few more years until the new technologies were firstly applied to analyze an outbreak. This corresponded to an outbreak of methicillin-resistant *Staphylococcus aureus* (MRSA) (Harris et al. 2010). They analyzed a set of 63 isolates from two origins, a global collection of 43 samples

collected between 1982 and 2003, and 20 isolates from a Thai hospital sampled in a very short time period (months), suspected to correspond to a transmission chain. Their results provided evidence for the international spread of the resistant clone of *S. aureus* and the single origin of the samples from the hospital. But they also showed that bacteria can and do evolve rapidly. They estimated that in the core genome, the set of shared positions among all the studied isolates, the rate of divergence was about 1 SNP every 6 weeks. This explained the lack of identity among most hospital isolates, which differed in a few SNPs from each other, but it also revealed differences from the patterns of evolution revealed by other markers, such as *spa* and PFGE. Of note, the analysis of complete genomes showed that over a quarter of the homoplasies found among the isolates were directly related to the evolution of resistance to antibiotics.

At about the same time, Lewis et al. (2010) used complete genome sequences to establish relationships among otherwise indistinguishable strains of *Acinetobacter baumannii* which had cause a small outbreak at a British hospital. The SNPs found by WGS allowed the investigators to discriminate among alternative epidemiological hypotheses. These pioneering studies have been followed by many studies (Table 13.1) which have dealt with outbreaks and transmission networks of over 30 different bacteria species infecting humans. An even larger number of works have been published about viral infections (not included in this review), and a few have dealt with fungal infections. Two particular bacteria, *M. tuberculosis* and *L. pneumophila*, the main etiological agents of tuberculosis and legionellosis, respectively, are analyzed in more detail below, but some general patterns and conclusions have started to emerge from the analysis of more than 30 pathogenic bacteria that we briefly review next.

From retrospective to real-time analysis of outbreaks. We have previously commented that the molecular analysis of outbreaks and transmission networks is necessarily a complement to the epidemiological investigations leading to the identification and control of the source(s), vectors, or routes so to put a fast stop to ongoing processes. Hence, it is very important that the information obtained from the molecular analyses can be shared with the epidemiology team for a better evaluation of the total evidence available thus far and more appropriate and accurate decisions can be adopted. The initial methodologies available for WGS were very labor intensive, and the shortest time since a sample was obtained until its complete sequence could be determined was in the order of weeks. Too long for a pressing demand of action. However, the advent of new technologies, such as Ion Torrent PGM and, more recently, MinION, has changed this situation. Both methods can deliver sequence information within a few hours of gaining access to the sample, thus allowing a very rapid communication of results to field workers.

The first case in which these new technologies were applied during the investigation of the source of an outbreak was that an enteroaggregative *E. coli* O104:H4 strain that affected several European countries in the spring of 2011 (Mellmann et al. 2011). Complete genome sequences were obtained from a representative isolate of the outbreak and a reference strain which produced similar clinical features in just 62 h. The comparison revealed key differences in plasmid and gene

contents between the strains, indicating that the outbreak was due to a new and not a previously circulating strain of the bacterium. It also allowed the design of a test to be applied for quick diagnostic in any laboratory.

Loss of identity as hallmark of relatedness. One consequence of using complete genome sequences for the analysis of outbreaks and transmission chains is the necessary dismissal of complete identity as the proof of charge in considering two or more isolates as linked to the same transmission event or episode. This was usually the case for most previous markers which explored only a minor fraction of the nucleotides in the genome of the pathogenic bacteria. Except for a few rapidly evolving markers, usually associated with tandem repeats, the number of differences expected between two isolates depends on three factors: the mutation rate per site, the number of sites being compared, and the time since they diverged from their last common ancestor. When the number of generations since divergence is relatively small, as in outbreaks and most transmission networks, and the number of sites being sampled is also small, the probabilities of finding a SNP (or a different allele in the case of MLST) are also very small. However, using complete genome sequences, and assuming that the previous assumptions remain identical, will increase those probabilities in a three-fold factor or more, because the number of sites interrogated is now in the order of millions instead of tens or hundreds.

Within-host evolution. In addition, the exploration of complete genome sequences of long- or chronically infecting bacteria has shown that evolution does occur within hosts at relevant rates for being reflected in some nucleotide changes (Didelot et al. 2016). Even for pathogens that produce acute infections, a low per site mutation rate is compensated by the large number of nucleotides present in a genome and the different random and directional processes that occur in an infected individual, thus leading to some new mutations arising in many newly replicated genomes (Kennemann et al. 2011; Mathers et al. 2015). If the infection lasts longer or becomes chronic, the chances that changes occur in the pathogen are very high and additional evolutionary processes such as compartmentalization may contribute to within patient differentiation of bacterial subpopulations.

These processes have important consequences at different levels. On the one hand, a variable population can adapt more rapidly to new environmental conditions which might include new treatments or an adaptive immune response by the host (Mwangi et al. 2007). On the other hand, a variable population will result in different initial compositions in successive transmission events, which will be reflected in differences among the populations established in the new hosts. The analysis of transmission networks becomes more complicated because using a single genome sequence per host cannot reveal the whole range of variation present within it (Worby et al. 2014). Under these circumstances, the use of evolutionary methods to reveal the common ancestry of isolates derived from patients presumably included in the same network becomes an absolute necessity.

Mutation patterns and processes. Apart from revealing larger amounts of variation than anticipated from previous studies with just a few gene sequences, whole-genome sequences have also informed about the types and distribution of mutational changes occurring at different timescales. A few years ago, the

contribution of homologous recombination and horizontal gene transfer to genetic variation in bacterial genomes was found to be considerably more important than previously thought (Doolittle 1998). But this was thought to be the result of millions of generations in which a generally rare process might have been acting. In shorter timescales, months or years, the impact of processes generating variation other than point mutation was thought to be negligible except for loci including repeat units, such as in MIRU-VNTRs in *M. tuberculosis*, in which slippage and mispairing during replication often lead to new alleles.

Recent analyses at the complete genome level have shown that this view is incorrect, at least for some bacteria such as *Neisseria gonorrhoeae*, *Salmonella enterica,* or *L. pneumophila* (Didelot and Maiden 2010; Sánchez-Busó et al. 2014). In fact, a comparison of the relative effects of recombination and point mutation in almost 50 bacterial species revealed variation of three orders of magnitude (Vos and Didelot 2009). Although there are not quantitative estimates yet, horizontal gene transfer, with or without final stabilization in the receiving genome, is also known to play a significant role in the short-term evolution of many bacteria, as unfortunately shown by the ease of spread of many antibiotic resistance genes across species. The additional variation introduced by these processes has to be considered when analyzing large transmission networks or long-lasting outbreaks, because the incorporation of these new variants may confound inferences of recent ancestry based on overall similarity or on a few loci.

Rates of evolution. The increased availability of complete genome sequences from bacteria with a more or less direct epidemiological link has also provided an opportunity for a more detailed study of evolutionary processes at the population genomic level. Apart from the different types of variants introduced in these populations, the access to asynchronically sampled isolates allows the application of Bayesian methods to estimate evolutionary rates (Drummond et al. 2006). These methods can accommodate strict and relaxed clock models, different demographic regimes, as well as variation in rates among lineages, thus allowing the estimation of relevant evolutionary parameters from organisms with different natural and evolutionary histories. Most often they are applied to rapidly evolving organisms, collectively known as measurably evolving populations (Drummond et al. 2003; Biek et al. 2015), which mainly include viruses along with some bacteria. But the methods are also valid for more slowly evolving organisms with sampling dates different enough as to provide estimates of the evolutionary rate. Recently, this approach has been used with bacterial genomes obtained from ancient samples (Schuenemann et al. 2013; Bos et al. 2014, 2016; Mendum et al. 2014; Rasmussen et al. 2015; Maixner et al. 2016).

One apparent feature of the estimates of bacterial evolutionary rates is the negative correlation between the time to the most recent common ancestor of the sample studied and the inferred evolutionary rate (Fig. 13.1). Higher evolutionary rates at short times can be explained by the relative inefficiency of natural selection and/or genetic drift in the removal of neutral or quasi-neutral polymorphisms which are continuously arising in bacterial populations. Hence, transitional polymorphisms contribute significantly to the apparent acceleration of evolutionary rates in

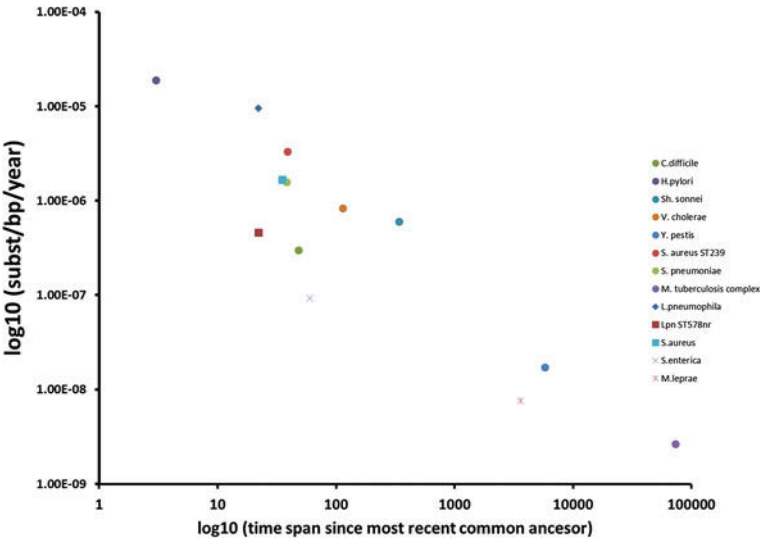


Fig. 13.2 Estimates of evolutionary rate for different bacterial species and its relationship to the time elapsed since the most recent common ancestor of the isolates used to determine the rate. *Sources of data H. pylori* (Kennemann et al. 2011), *C. difficile* (Didelot et al. 2012b), *Sh. sonnei* (Holt et al. 2012), *Y. pestis* (Rasmussen et al. 2015), *S. aureus* (Harris et al. 2010, Ward et al. 2014), *S. pneumoniae* (Croucher et al. 2011), *L. pneumophila* (Sánchez-Busó et al. 2014), *M. tuberculosis* (Comas et al. 2013), *M. leprae* (Schuenemann et al. 2013), *S. enterica* (Zhou et al. 2013)

short timescales. At the same time, they also provide a wealth of variation what might have an adaptive value if the circumstances are appropriate. On the long run, many of these transient variants will have disappeared and evolutionary rates are reduced correspondingly. This negative correlation has to be taken into account when comparing rates across studies, even for the same species, and in the inference of other evolutionary parameters (Biek et al. 2015).

The analysis of (almost) complete genome data. One of the main advantages of MLST or SBT over alternative methods for the analysis of pathogenic bacteria in the context of outbreaks and transmission chains is the objectivity and simplicity in the specification of the variants found in any isolate. The nucleotide sequences obtained for each locus are compared to a predetermined database in which previous homologous sequences have been deposited. If there is a perfect match, the newly determined variant received the same identifier as the pre-existing one. If that is not the case, curators of the database will assign a new code to the variant. The combination of allele codes in the loci included in the typing scheme is summarized in a sequence type (ST) with a different number of each combination of variants. This procedure is easily communicated because it requires the identification of nucleotide variants, usually through Sanger sequencing, in just 6 or 7 loci. However, the advent of NGS and the determination of complete genome sequences makes this procedure of denoting the variants impractical (Fig. 13.2).

Several alternatives have already been proposed for the identification of complete genome sequences for epidemiological analysis. One method consists in extending the MLST naming scheme to more loci, eventually all the loci in the genome of the corresponding species, thus leading to "whole-genome MLST" (wgMLST) schemes (Cody et al. 2013). The first proposal of wgMSLT was done for *Campylobacter* isolates and the initial MLST scheme based on 7 loci was extended to 1667 loci, although this number was reduced to 1026 when only those present in all the isolates analyzed were considered. This represents the "core genome" of the species, which is complemented by the "auxiliary genome," the set of loci which are present in some but not all the isolates of a species. In light of the very large genome plasticity of many bacterial species, fixed compositions of the core and auxiliary genomes are almost impossible, which creates an additional problem for the stability of the scheme. Nevertheless, this approach has gained some popularity and cgMLST ("core genome MLST," a reduced version of wgMLST as described above) schemes are now available for several pathogens including *S. aureus*, *Listeria monocytogenes*, *Enterococcus faecium* (De Been et al. 2015), and *S. enterica* (Taylor et al. 2015), among others.

To prevent the proliferation of STs which inevitably accompanies wgMLST or cgMLST, a first-level classification of STs into clusters or clonal groups is usually performed (Cody et al. 2013; Qin et al. 2016). These can be based on an extension of the BURST method (Feil et al. 2001, 2004), which considers as variants of the same clonal group to those that differ in one single locus of the original MLST scheme, or use more sophisticated approaches based on the population genetic analysis of the actual SNPs detected in the loci included in the wgMLST or cgMLST (Qin et al. 2016) with different molecular population methods such as BAPS (Corander and Tang 2007) or STRUCTURE (Rosenberg et al. 2002). These methods share the advantage of portability, thus allowing comparisons among different laboratories and needs. However, they also discard important information, eventually crucial, contained in the auxiliary genome. Hence, although standard typing schemes are useful, whole-genome sequence information should not be reduced to a ST number or complex under a wgMLST and the complete data should still be available for future use by the scientific community.

13.4 Outbreak Investigation in *Mycobacterium tuberculosis*: The Genome as an Epidemiological Marker

Mycobacterium tuberculosis is the main causative agent of human tuberculosis in the world. Every year, more than 1.5 million persons die of tuberculosis, more than of any other infectious disease (WHO 2014). The epidemiology of the disease has to take into account the natural history of the bacteria. It is an obligate human pathogen with very effective airborne transmission and that typically infects the

lungs. It is estimated that one-third of the human population is infected by the bacilli, and this explains why every year around 9 million new cases are declared. In most cases, the initial infection derives in an asymptomatic state called latency in which the bacteria have not been eliminated but are controlled by the immune system. In 5–10 % of the latent cases, the disease progresses to an active state in which the bacteria actively replicate and cause pulmonary disease. Only an active tuberculosis case can transmit the disease, and thus, in tuberculosis, disease and transmission are linked. The typical window of progress to active disease after infection is two years but the bacteria may remain latent for years or even decades.

Mycobacterium tuberculosis has been traditionally regarded as a monomorphic organism due to the low genetic diversity found among representative strains datasets (Achtman 2008). Thus, epidemiological tools were developed based on fast-evolving genetic elements (Barnes and Cave 2003). Typing of the insertion sequence IS6110 by RFLP and typing of minisatellites, called MIRU-VNTR, are the two gold standards in tuberculosis molecular epidemiology and, together with spoligotyping, based on the CRISPR region of the bacteria, have allowed to define successful *M. tuberculosis* clones. Among these clones, the identification of a hypervirulent clade, called Beijing family, has attracted much attention (Parwati et al. 2010). Strains from the Beijing family are more common in East Asia but can be identified across the globe. Experimental and epidemiological research has identified Beijing strains as hypervirulent in the mice model of infection and with frequent association to drug resistance in humans. In South Africa, Beijing strains have been on the rise for the last 40 years (Cowley et al. 2008). Beijing strains belong to one of the seven lineages of human tuberculosis strains (Comas et al. 2013). The most common is lineage 4, which is highly frequent in Africa, Europe, and America. There is a strong association between lineages and their geographic origin, being the most extreme cases the two lineages of *Mycobacterium africanum,* that can only be found in West Africa (De Jong et al. 2010), and Lineage 7 recently described in Ethiopia (Comas et al. 2013). Regardless the lineage, drug resistance to first- and second-line treatments has been identified (Farhat et al. 2013). The mutations responsible for drug resistance are always chromosomal mutations because there is no ongoing horizontal gene transfer in *M. tuberculosis*. Although ecological theory predicts that drug resistance mutations have a fitness cost, experimental evolution and molecular epidemiology have shown that different drug resistance mutations have different fitness costs (Comas et al. 2012). As a consequence, multidrug-resistant cases (MDR-TB) among people never treated before, and therefore due to transmission, are on the rise and in some particular areas represent more than 50 % of the tuberculosis burden of the region. Although not part of this review, whole-genome sequencing is allowing to define the set of mutations associated with resistance to the different antibiotics but also the genotype of highly successful MDR-TB strains.

The first study that showed the potential of the genome as an epidemiological marker dates back to 2009 (Niemann et al. 2009). In this study, three strains which looked almost identical using traditional molecular epidemiology markers such as restriction fragment length polymorphisms (RFLP) and minisatellite

(MIRU-VNTR) were shown to differ in more than 100 SNPs. Later on, Gardy et al. (2011) used genome comparison techniques to solve an ongoing outbreak in British Columbia suspected to have started in the early 1990s. By combining genomic, epidemiological, and social contact data, the authors showed that it can be gained a better resolution of the transmission events within transmission clusters. Such events are very difficult to identify with traditional molecular epidemiology markers. This work already defined index cases associated with multiple secondary cases, also denoted as super-spreaders. Super-spreaders are becoming a common topic when analyzing large transmission clusters (Walker et al. 2013b) instead of the traditional view of a stepwise "chain" of transmission.

From 2010, NGS has been successfully applied to deeply resolve tuberculosis outbreaks. Considerably attention has been paid to understand those outbreaks that have been ongoing over years. For example, a large outbreak in Hamburg, Germany, was identified by classical genotyping data in 1996 (Roetzer et al. 2013). However, clustering data not always correlated with epidemiological and geographical information leading to the suspicion that the outbreak was more complex than previously anticipated. By whole-genome sequencing of 86 strains from the outbreak (1996–2011), Roetzer et al. (2013) were able to identify an independent transmission network, thus confirming the non-clonality of the outbreak. Two clusters were determined, one starting in 1997 and the other starting in 2010, much more in agreement with epidemiological investigations. Therefore, one important application of whole-genome sequencing to investigate tuberculosis outbreaks is the ability to assign with higher confidence cases to the outbreak and exclude those that, albeit genetically close, correspond to a different chain of events.

Similarly, in Bern, Switzerland, a genotype detected by RFLP profiling caused a large number of tuberculosis cases during the 1990s (Stucki et al. 2015). The cases were associated with the typical risk factors in local populations found in European cities such as HIV infection or alcoholism. Stucki et al. (2015) sequenced the complete genome of strains belonging to the original outbreak along with local control strains. By comparing outbreak and control strains, they designed a real-time SNP typing assay based on the detection of genome position with a polymorphism specific to the outbreak strains. Next, they typed a retrospective collection of isolates of the Canton of Bern from 1993 to 2011. They were able to identify 68 additional cases of the outbreak based on the presence of the mentioned SNP including cases from 2011. Therefore, the combination of whole-genome sequencing and SNP typing allowed them to identify cases associated with the outbreak and find that the outbreak that started in early 1990s was still ongoing at the time of investigation. In addition, they obtained the whole-genome sequence of all the isolates assigned to the outbreak. With this information, they were able to resolve the individual transmission patterns for 75 % of the strains. Importantly, 66 out of the 68 strains had exactly the same RFLP pattern. Furthermore, the analysis of the transmission network together with the epidemiological information revealed two different sub-outbreaks initiated by two different "super-spreaders".

Therefore, next-generation sequencing of the Hamburg (Roetzer et al. 2013), the Bern outbreak (Stucki et al. 2015), and others (Török and Peacock 2012; Smit et al. 2015; Lee et al. 2015) have revealed the complexity of tuberculosis outbreaks. Given that tuberculosis is not an acute disease and that a tuberculosis case can be latent, asymptomatic for years, the true extent of tuberculosis outbreaks can only be revealed by sustained genotyping efforts over years. Furthermore, as in the case of the Bern outbreak, whole-genome sequence data can be used to design new diagnostics and/or surveillance tools. A similar approach has been used to prospectively identify new outbreak-associated cases in sputum samples (Pérez-Lago et al. 2015).

Apart from specific outbreaks, genomic epidemiology has been used in a population-based scale to evaluate its utility for surveillance and diagnostics. In a series of publications starting in 2012, Public Health England has applied next-generation sequencing to incorporate whole-genome sequencing as the default typing method of *Mycobacterium tuberculosis* in the UK (Walker and Beatson 2012; Walker et al. 2014). They have shown that the genome data allow to delineate outbreaks better than MIRU-VNTR analyses. Furthermore, in an attempt to derive a rule of thumb to identify a transmission event between two cases, they also sequence several isolates from the same patient and known household contacts. They were able to identify a threshold of five SNPs when the cases had a confirmed epidemiological link and they proposed a threshold of up to 12 SNPs for casual transmission in the community (Walker and Beatson 2012). Other studies have found a similar distribution of SNPs when analyzing transmission events in populations (Bryant et al. 2013a; Casali et al. 2014).

However, we are still blinded about how these thresholds apply to different clinical settings than the low-burden countries of Europe. In high-burden countries, delineating transmission clusters should be more difficult if public health interventions cannot stop transmission events (Yates et al. 2016). Thus, the circulating strains may be participating at the same time in several clusters. The only population-based study published in a high-burden country shows that the threshold described in (Pérez-Lago et al. 2015) may be useful, although more work will be needed to generalize the results to, for example, large urban areas.

There are several factors that may distort the proposed threshold values. One of these factors is mixed infections. The true extent of co-infections in high-burden countries is not clear, and there is hope that whole-genome data can distinguish between relapses and re-infections (Bryant et al. 2013a; Guerra-Assunção et al. 2015). This issue is critical to delineate transmission in high-burden countries but also for clinical trials investigations because relapse is one of the end points of those investigations. However, it is the diversity that can be found during infection from a single strain what is attracting more research and attention. From drug susceptibility clinical data, it has been clear for decades that several populations may coexist in the same patient. These subpopulations were flagged due to inconsistent results in drug resistance susceptibility tests between isolates of the same patient (Rinder et al. 2001). Whole-genome sequencing has shown that, in fact, this is the case and what is recovered from a sputum sample is often a mix of different subpopulations (Sun et al. 2012). These subpopulations can be revealed by looking at positions in which

a mutant and a wild-type allele can be identified at the same time. In the context of drug resistance, it has been shown that several drug-resistant subpopulations may coexist and compete and that their frequencies may change over time (Liu et al. 2015). A similar phenomenon has been shown outside the context of drug resistance. The issue of within patient diversity not only has clinical and diagnostic implications. If several subpopulations coexist and accumulate a different number of SNPs, then chances are that the epidemiological investigation of outbreaks may be distorted by the isolate chosen for the analysis (Walker et al. 2013a, b). An analysis of cases in which higher than expected diversity was expected confirmed that, although the thresholds proposed to delineate a transmission event are in general valid, there are epidemiological cases in which a larger than expected number of SNPs can be found (Pérez-Lago et al. 2014). How frequent are those "outliers" is a matter of ongoing investigation.

13.5 High-throughput Investigation of *Legionella Pneumophila* Outbreaks

High-throughput sequencing can also be used to study organisms with higher level of polymorphism and strictly environmental, contrary to Mycobacterium tuberculosis. This is the case of *L. pneumophila*, causative agent of Legionellosis, and for which there is only one report of a possible person-to-person transmission (Correia et al. 2016) up to date. This opportunistic pathogen can produce pneumonia after inhalation of aerosols with enough bacterial load, with the highest burden in warm water-related environments. The first reported outbreak dates from 1976 when more than hundred legionnaires were infected in a convention in Philadelphia (Fraser et al. 1977). A legionellosis outbreak is defined as a cluster of more than three cases occurring at the same place and time, and the epidemiological investigation is crucial to find the environmental sources.

The investigation of legionellosis outbreaks has traditionally been conducted by using biochemical or molecular methods that allows comparing the clinical isolates with the strains obtained from the environment (Fields et al. 2002). Broad techniques such as serogrouping benefited from genetic methods that provided improved resolution in the so-called sequence-based typing (SBT) (Gaia et al. 2003, 2005), based on MLST approach (Urwin and Maiden 2003) but incorporating virulence genes in the scheme to increase the discrimination power among strains.

However, although SBT provided researchers with a tool that allowed the classification of strains into groups (sequence types, STs), the introduction of high-throughput sequencing techniques for microbial analysis and outbreak investigations in other species derived in its application to legionellosis outbreaks because of its increased discrimination power. The first published work was indeed a pilot study to test the potential of whole-genome sequencing (WGS) on the discrimination between isolates from an outbreak produced in the UK in 2003 and

non-outbreak-related strains (Reuter et al. 2013b). From this point, a number of other outbreaks have been analyzed using WGS, as for example an outbreak of ST62 associated with a cooling tower in Quebec City in 2012 (Lévesque et al. 2014) or a massive outbreak that occurred in Edinburgh (UK 2012) related to multiple STs and including mixed infections (McAdam et al. 2014). WGS has also been used to investigate the persistent infection history of ST23 in a hotel in Spain in 2012 (Sánchez-Busó et al. 2016) and the eradication of *L. pneumophila* associated with a hospital in Australia that have been responsible of nosocomial cases (Bartley et al. 2016).

The environmental source of legionellosis cases has been historically difficult to trace, and because of the high social and economic impact of this kind of outbreaks on the affected populations, public health interventions are obliged to be rapid and accurate. WGS has shown further variability within many STs (Underwood et al. 2013a; Sánchez-Busó et al. 2014), showing evidence that at least some of them are not clonal. This observation complicates the study of legionellosis outbreaks and was the leading aim in the study by Sánchez-Busó et al. (2014). In this work, 69 isolates including strains associated with 13 different outbreaks and sporadic cases occurred in a single locality (Alcoy, Spain) during more than 10 years (1999–2010) were analyzed by high-throughput sequencing. Different STs were included, with special interest on ST578 cases, which had been recurrently reported as the causing ST of most of those outbreaks (Coscollá et al. 2010).

The analysis showed two main lineages within the endemic ST578, more than 1000 SNPs apart from each other. Not all the strains from the same outbreak clustered together, revealing the non-clonality of the isolates, as these were phylogenetically grouped independently of their source (clinical or environmental), sampling date or outbreak. Because ST578 is known to be endemic in the area of Alcoy, these results suggest that it is indeed very complicated to find an infectious source using just molecular data in endemic areas. These should be used together with the epidemiological investigation to be able to draw the accurate conclusions that public health interventions require.

Other interesting fact that this work shows is that the genomic data can reflect public health actions along time. As an example, using Bayesian inference, an estimate of the ST578 population dynamics revealed a decreased population size between 2006 and 2008, which correlated with a moment in which public health measures were taken in the city by removing high-risk installation from the city center.

In the case of organisms where person-to-person transmission is very rare or even inexistent, whole-genome sequencing can provide the most discriminant tool to link clinical cases with environmental sources, providing the accuracy that public health interventions require in these cases. But, moreover, it can help understand how outbreaks occur, which is the starting line to be able to predict and even prevent their occurrence.

13.6 Conclusion

Complete genome analysis of bacterial pathogens is still far from being the usual method for analyzing outbreaks and transmission networks, although it will not take long before it does so. The increasing speed, ease, and reliability as well as the reduced costs associated with new high-throughput sequencing technologies point to that direction. But gaining information is only a part of the process. More data also mean an increased need for interpretative tools at all levels, from the mere analysis of reads to the inference of the evolutionary and genealogical relationships among the isolates. Progress is still pending at all levels, from the technology to obtain, fast and cheap, complete genome sequence data of a specific pathogen from an infected individual or a potential vector o source to analytical tools capable of extracting the relevant information from the deluge of data generated by high-throughput sequencers and for the integration of this information with the clinical, epidemiological, and evolutionary information which are needed when they have to be interpreted in the appropriate context.

Acknowledgements We thank Dr. Pierre Pontarotti for his kind invitation to write this chapter. This work has been funded by project BFU2014-58656-R from MINECO (Spanish Government) to FGC. IC is supported by Ramón y Cajal Spanish research grant RYC-2012-10627, MINECO research grant SAF2013-43521-R, and the European Research Council (ERC) (638553-TB-ACCELERATE). BB has been recipient of a Beca de Colaboración from the Spanish Ministerio de Educación y Cultura.

References

Achtman M (2008) Evolution, population structure, and phylogeography of genetically monomorphic bacterial pathogens. Annu Rev Microbiol 62:53–70

Achtman M (2012) Insights from genomic comparisons of genetically monomorphic bacterial pathogens. Phil Trans R Soc B 367:860–867

Ahmed SA, Awosika J, Baldwin C, Bishop-Lilly KA, Biswas B, Broomall S, Chain PSG, Chertkov O, Chokoshvili O, Coyne S, Davenport K, Detter JC, Dorman W, Erkkila TH, Folster JP, Frey KG, George M, Gleasner C, Henry M, Hill KK, Hubbard K, Insalaco J, Johnson S, Kitzmiller A, Krepps M, Lo CC, Luu T, McNew LA, Minogue T, Munk CA, Osborne B, Patel M, Reitenga KG, Rosenzweig CN, Shea A, Shen X, Strockbine N, Tarr C, Teshima H, Van Gieson E, Verratti K, Wolcott M, Xie G, Sozhamannan S, Gibbons HS, Threat Characterization Consortium (2012) Genomic comparison of *Escherichia coli* O104: H4 isolates from 2009 and 2011 reveals plasmid, and prophage heterogeneity, including shiga toxin encoding phage stx2. PLoS ONE 7:e48228

Allard MW, Luo Y, Strain E, Li C, Keys CE, Son I, Stones R, Musser SM, Brown EW (2012) High resolution clustering of *Salmonella enterica* serovar Montevideo strains using a next-generation sequencing approach. BMC Genom 13:1

Allard MW, Luo Y, Strain E, Pettengill J, Timme R, Wang C, Li C, Keys CE, Zheng J, Stones R, Wilson MR, Musser SM, Brown EW (2013) On the evolutionary history, population genetics and diversity among isolates of *Salmonella* enteritidis PFGE pattern JEGX01.0004. PLoS ONE 8:e55254

Azarian T, Cook RL, Johnson JA, Guzman N, McCarter YS, Gomez N, Rathore MH, Morris JGJ, Salemi M (2015) Whole-genome sequencing for outbreak investigations of Methicillin-resistant *Staphylococcus aureus* in the neonatal intensive care unit: time for routine practice? Infect Control Hosp Epidemiol 36:777–785

Barnes PF, Cave MD (2003) Molecular epidemiology of tuberculosis. N Engl J Med 349:1149–1156

Bartley PB, Ben Zakour NL, Stanton-Cook M, Muguli R, Prado L, Garnys V, Taylor K, Barnett TC, Pinna G, Robson J, Paterson DL, Walker MJ, Schembri MA, Beatson SA (2016) Hospital-wide eradication of a nosocomial *Legionella pneumophila* serogroup 1 outbreak. Clin Infect Dis 62:273–279

Bekal S, Berry C, Reimer AR, Van Domselaar G, Beaudry G, Fournier E, Doualla-Bell F, Levac E, Gaulin C, Ramsay D, Huot C, WAlker M, Sieffert C, Tremblay C (2016) Usefulness of high-quality core genome single-nucleotide variant analysis for subtyping the highly clonal and the most prevalent *Salmonella enterica* serovar Heidelberg clone in the context of outbreak investigations. J Clin Microbiol 54:289–295

Bennett JS, Jolley KA, Earle SG, Corton C, Bentley SD, Parkhill J, Maiden MCJ (2012) A genomic approach to bacterial taxonomy: an examination and proposed reclassification of species within the genus Neisseria. Microbiology 158:1570–1580

Biek R, Pybus OG, Lloyd-Smith JO, Didelot X (2015) Measurably evolving pathogens in the genomic era. Trends Ecol Evol 30:306–313

Blouin Y, Cazajous G, Dehan C, Soler C, Vong R, Hassan MO, Hauck Y, Boulais C, Andriamanantena D, Martinaud C, Martin É, Pourcel C, Vergnaud G (2014) Progenitor "*Mycobacterium canettii*" clone responsible for lymph node tuberculosis epidemic, Djibouti. Emerg Infect Dis 20:21–28

Bos KI, Harkins KM, Herbig A, Coscolla M, Weber N, Comas I, Forrest SA, Bryant JM, Harris SR, Schuenemann VJ (2014) Pre-columbian mycobacterial genomes reveal seals as a source of new world human tuberculosis. Nature 514:494–497

Bos KI, Herbig A, Sahl J, Waglechner N, Fourment M, Forrest SA, Klunk J, Schuenemann VJ, Poinar D, Kuch M, Golding GB, Dutour O, Keim P, Wagner DM, Holmes EC, Krause J, Poinar HN (2016) Eighteenth century Yersinia pestis genomes reveal the long-term persistence of an historical plague focus. eLife Sci e12994

Bryant J, Schurch A, Van Deutekom H, Harris S, de Beer J, de Jager V, Kremer K, Van Hijum S, Siezen R, Borgdorff M, Bentley S, Parkhill J, Van Soolingen D (2013a) Inferring patient to patient transmission of *Mycobacterium tuberculosis* from whole genome sequencing data. BMC Infect Dis 13:110

Bryant JM, Grogono DM, Greaves D, Foweraker J, Roddick I, Inns T, Reacher M, Haworth CS, Curran MD, Harris SR, Peacock SJ, Parkhill J, Floto RA (2013b) Whole-genome sequencing to identify transmission of *Mycobacterium abscessus* between patients with cystic fibrosis: a retrospective cohort study. The Lancet 381:1551–1560

Brzuszkiewicz E, Thürmer A, Schuldes J, Leimbach A, Liesegang H, Meyer FD, Boelter J, Petersen H, Gottschalk G, Daniel R (2011) Genome sequence analyses of two isolates from the recent *Escherichia coli* outbreak in Germany reveal the emergence of a new pathotype: Entero-Aggregative-Haemorrhagic *Escherichia coli* (EAHEC). Arch Microbiol 193:883–891

Cao G, Meng J, Strain E, Stones R, Pettengill J, Zhao S, McDermott P, Brown E, Allard M (2013) Phylogenetics and differentiation of *Salmonella* newport lineages by whole genome sequencing. PLoS ONE 8:e55687

Casali N, Nikolayevskyy V, Balabanova Y, Ignatyeva O, Kontsevaya I, Harris SR, Bentley SD, Parkhill J, Nejentsev S, Hoffner SE, Horstmann RD, Brown T, Drobniewski F (2012) Microevolution of extensively drug-resistant tuberculosis in Russia. Genome Res 22:735–745

Casali N, Nikolayevskyy V, Balabanova Y, Harris SR, Ignatyeva O, Kontsevaya I, Corander J, Bryant J, Parkhill J, Nejentsev S, Horstmann RD, Brown T, Drobniewski F (2014) Evolution and transmission of drug resistant tuberculosis in a Russian population. Nat Genet 46:279–286

Casey G, Conti D, Haile R, Duggan D (2013) Next generation sequencing and a new era of medicine. Gut 62:920–932

Chewapreecha C, Harris SR, Croucher NJ, Turner C, Marttinen P, Cheng L, Pessia A, Aanensen DM, Mather AE, Page AJ, Salter SJ, Harris D, Nosten F, Goldblatt D, Corander J, Parkhill J, Turner P, Bentley SD (2014) Dense genomic sampling identifies highways of pneumococcal recombination. Nat Genet 46:305–309

Chin CS, Sorenson J, Harris JB, Robins WP, Charles RC, Jean-Charles RR, Bullard J, Webster DR, Kasarskis A, Peluso P, Paxinos EE, Yamaichi Y, Calderwood SB, Mekalanos JJ, Schadt EE, Waldor MK (2011) The origin of the Haitian cholera outbreak strain. N Engl J Med 364:33–42

Cody AJ, McCarthy ND, Jansen van Rensburg M, Isinkaye T, Bentley SD, Parkhill J, Dingle KE, Bowler ICJW, Jolley KA, Maiden MCJ (2013) Real-time genomic epidemiological evaluation of human *Campylobacter* isolates by use of whole-genome multilocus sequence typing. J Clin Microbiol 51:2526–2534

Comas I, Borrell S, Roetzer A, Rose G, Malla B, Kato-Maeda M, Galagan J, Niemann S, Gagneux S (2012) Whole-genome sequencing of rifampicin-resistant *Mycobacterium tuberculosis* strains identifies compensatory mutations in RNA polymerase genes. Nat Genet 44:106–110

Comas I, Coscolla M, Luo T, Borrell S, Holt KE, Kato-Maeda M, Parkhill J, Malla B, Berg S, Thwaites G, Yeboah-Manu D, Bothamley G, Mei J, Wei L, Bentley S, Harris SR, Niemann S, Diel R, Aseffa A, Gao Q, Young D, Gagneux S (2013) Out-of-Africa migration and Neolithic coexpansion of *Mycobacterium tuberculosis* with modern humans. Nat Genet 45:1176–1182

Corander J, Tang J (2007) Bayesian analysis of population structure based on linked molecular information. Math Biosci 205:19–31

Correia AM, Ferreira JS, Borges V, Nunes A, Gomes B, Capucho R, Gonçalves J, Antunes DM, Almeida S, Mendes A, Guerreiro M, Sampaio DA, Vieira L, Machado J, Simões MJ, Gonçalves P, Gomes JP (2016) Probable person-to-person transmission of Legionnaires' disease. N Engl J Med 374:497–498

Coscollá M, Fenollar J, Escribano I, González-Candelas F (2010) Legionellosis outbreak associated with asphalt paving machine, Spain, 2009. Emerg Infect Dis 16:1381–1387

Coscollá M, Barry PM, Oeltmann JE, Koshinsky H, Shaw T, Cilnis M, Posey J, Rose J, Weber T, Fofanov VY, Gagneux S, Kato-Maeda M, Metcalfe JZ (2015) Genomic epidemiology of multidrug-resistant *Mycobacterium tuberculosis* during transcontinental spread. J Infect Dis

Cowley D, Govender D, February B, Wolfe M, Steyn L, Evans J, Wilkinson RJ, Nicol MP (2008) Recent and rapid emergence of W-Beijing strains of *Mycobacterium tuberculosis* in Cape Town, South Africa. Clin Infect Dis 47:1252–1259

Croucher NJ, Harris SR, Fraser C, Quail MA, Burton J, Van der Linden M, McGee L, von Gottberg A, Song JH, Ko KS, Pichon B, Baker S, Parry CM, Lambertsen LM, Shahinas D, Pillai DR, Mitchell TJ, Dougan G, Tomasz A, Klugman KP, Parkhill J, Hanage WP, Bentley SD (2011) Rapid pneumococcal evolution in response to clinical interventions. Science 331:430–434

Croucher NJ, Finkelstein JA, Pelton SI, Mitchell PK, Lee GM, Parkhill J, Bentley SD, Hanage WP, Lipsitch M (2013) Population genomics of post-vaccine changes in pneumococcal epidemiology. Nat Genet 45:656–663

Cui Y, Yu C, Yan Y, Li D, Li Y, Jombart T, Weinert LA, Wang Z, Guo Z, Xu L, Zhang Y, Zheng H, Qin N, Xiao X, Wu M, Wang X, Zhou D, Qi Z, Du Z, Wu H, Yang X, Cao H, Wang H, Wang J, Yao S, Rakin A, Li Y, Falush D, Balloux F, Achtman M, Song Y, Wang J, Yang R (2013) Historical variations in mutation rate in an epidemic pathogen, *Yersinia pestis*. Proc Nat Acad Sci U.S.A. 110:577–582

Davidson RM, Hasan NA, de Moura VCN, Duarte RS, Jackson M, Strong M (2013) Phylogenomics of Brazilian epidemic isolates of *Mycobacterium abscessus* subsp. *bolletii* reveals relationships of global outbreak strains. Infect Genet Evol 20:292–297

De Been M, Pinholt M, Top J, Bletz S, Mellmann A, Van Schaik W, Brouwer E, Rogers M, Kraat Y, Bonten M, Corander J, Westh H, Harmsen D, Willems RJL (2015) Core genome multilocus sequence typing scheme for high-resolution typing of *Enterococcus faecium*. J Clin Microbiol 53:3788–3797

De Jong BC, Antonio M, Gagneux S (2010) *Mycobacterium africanum*—review of an important cause of human tuberculosis in West Africa. PLoS Negl Trop Dis 4:e744

Devault AM, Golding GB, Waglechner N, Enk JM, Kuch M, Tien JH, Shi M, Fisman DN, Dhody AN, Forrest S, Bos KI, Earn DJD, Holmes EC, Poinar HN (2014) Second-pandemic strain of *Vibrio cholerae* from the Philadelphia cholera outbreak of 1849. N Engl J Med 370:334–340

Didelot X, Maiden MCJ (2010) Impact of recombination on bacterial evolution. Trends Microbiol 18:315–322

Didelot X, Eyre D, Cule M, Ip C, Ansari A, Griffiths D, Vaughan A, O'Connor L, Golubchik T, Batty E, Piazza P, Wilson D, Bowden R, Donnelly P, Dingle K, Wilcox M, Walker S, Crook D, Peto T, Harding R (2012a) Microevolutionary analysis of *Clostridium difficile* genomes to investigate transmission. Genome Biol 13:R118

Didelot X, Meric G, Falush D, Darling A (2012b) Impact of homologous and non-homologous recombination in the genomic evolution of *Escherichia coli*. BMC Genom 13:256

Didelot X, Walker AS, Peto TE, Crook DW, Wilson DJ (2016) Within-host evolution of bacterial pathogens. Nat Rev Micro 14:150–162

Doolittle WF (1998) You are what you eat: a gene transfer ratchet could account for bacterial genes in eukaryotic nuclear genomes. Trends Genet 14:307–311

Drummond AJ, Pybus OG, Rambaut A, Forsberg R, Rodrigo AG (2003) Measurably evolving populations. Trends Ecol Evol 18:481–488

Drummond AJ, Ho SYW, Phillips MJ, Rambaut A (2006) Relaxed phylogenetics and dating with confidence. PLoS Biol 4:e88

Ekblom R, Wolf JBW (2014) A field guide to whole-genome sequencing, assembly and annotation. Evol Appl 7:1026–1042

Engelthaler DM, Chiller T, Schupp JA, Colvin J, Beckstrom-Sternbergg SM, Driebe EM, Moses T, Tembe W, Sinari S, Beckstrom-Sternbergg JS, Christoforides A, Pearson JV, Capten J, Keim P, Peterson A, Tersahita D, Arunmozhi B (2011) Next-generation sequencing of *Coccidioides immitis* isolated during cluster investigation. Emerg Infect Dis 17:227–232

Espedido BA, Steen JA, Ziochos H, Grimmond SM, Cooper MA, Gosbell IB, Van Hal SJ, Jensen SO (2013) Whole genome sequence analysis of the first australian OXA-48-producing outbreak-associated *Klebsiella pneumoniae* isolates: the resistome and in vivo evolution. PLoS ONE 8:e59920

Eyre DW, Golubchik T, Gordon NC, Bowden R, Piazza P, Batty EM, Ip CL, Wilson DJ, Didelot X, O'Connor L (2012) A pilot study of rapid benchtop sequencing of *Staphylococcus aureus* and *Clostridium difficile* for outbreak detection and surveillance. BMJ Open 2:e001124

Eyre DW, Cule ML, Griffiths D, Crook DW, Peto TEA, Walker AS, Wilson DJ (2013a) Detection of mixed infection from bacterial whole genome sequence data allows assessment of its role in *Clostridium difficile* transmission. PLoS Comput Biol 9:e1003059

Eyre DW, Cule ML, Wilson DJ, Griffiths D, Vaughan A, O'Connor L, Ip CLC, Golubchik T, Batty EM, Finney JM, Wyllie DH, Didelot X, Piazza P, Bowden R, Dingle KE, Harding RM, Crook DW, Wilcox MH, Peto TEA, Walker AS (2013b) Diverse sources of *C. difficile* infection identified on whole-genome sequencing. N Engl J Med 369:1195–1205

Farhat MR, Shapiro BJ, Kieser KJ, Sultana R, Jacobson KR, Victor TC et al (2013) Genomic analysis identifies targets of convergent positive selection in drug-resistant *Mycobacterium tuberculosis*. Nat Genet 45:1183–1189

Feil EJ, Holmes EC, Bessen DE, Chan MS, Day NPJ, Enright MC, Goldstein R, Hood DW, Kalia A, Moore CE, Zhou J, Spratt BG (2001) Recombination within natural populations of pathogenic bacteria: short-term empirical estimates and long-term phylogenetic consequences. Proc Nat Acad Sci U.S.A. 98:182–187

Feil EJ, Li BC, Aanensen DM, Hanage WP, Spratt BG (2004) eBURST: inferring patterns of evolutionary descent among clusters of related bacterial genotypes from multilocus sequence typing data. J Bacteriol 186:1518–1530

Fields BS, Benson RF, Besser RE (2002) *Legionella* and Legionnaires' disease: 25 years of investigation. Clin Microbiol Rev 15:506–526

Fittipaldi N, Tyrrell GJ, Low DE, Martin I, Lin D, Hari KL, Musser JM (2013) Integrated whole-genome sequencing and temporospatial analysis of a continuing group A *Streptococcus* epidemic. Emerg Microbes Infect 2:e13

Fitzpatrick MA, Ozer EA, Hauser AR (2016) Utility of whole-genome sequencing in characterizing *Acinetobacter* epidemiology and analyzing hospital outbreaks. J Clin Microbiol 54:593–612

Fraser DW, Tsai TR, Orenstein W, Parkin WE, Beecham HJ, Sharrar RG, Harris J, Mallison GF, Martin SM, McDade JE, Shepard CC, Brachman PS (1977) Legionnaires' disease: description of an epidemic of pneumonia. N Engl J Med 297:1189–1197

Fraser C, Donnelly CA, Cauchemez S, Hanage WP, Van Kerkhove MD, Hollingsworth TD, Griffin J, Baggaley RF, Jenkins HE, Lyons EJ (2009) Pandemic potential of a strain of influenza A (H1N1): early findings. Science 324:1557–1561

Gaia V, Fry NK, Harrison TJ, Peduzzi R (2003) Sequence-based typing of *Legionella pneumophila* serogroup 1 offers the potential for true portability in legionellosis outbreak investigation. J Clin Microbiol 41:2932–2939

Gaia V, Fry NK, Afshar B, Luck PC, Meugnier H, Etienne J, Peduzzi R, Harrison TJ (2005) Consensus sequence-based scheme for epidemiological typing of clinical and environmental isolates of *Legionella pneumophila*. J Clin Microbiol 43:2047–2052

Gardy JL, Johnston JC, Sui SJH, Cook VJ, Shah L, Brodkin E, Rempel S, Moore R, Zhao Y, Holt R, Varhol R, Birol I, Lem M, Sharma MK, Elwood K, Jones SJM, Brinkman FSL, Brunham RC, Tang P (2011) Whole-genome squencing and social-network analysis of a tuberculosis outbreak. N Engl J Med 364:730–739

General Directorate of Epidemiology MoHM, Pan American Health Organization, World Health Organization, Public Health Agency of Canada, CDC (United States) (2009) Outbreak of swine-origin Influenza A (H1N1) virus infection-Mexico, March–April 2009. Morb Mortal Wkly Rep 58:467–470

Ghai R, Martin-Cuadrado AB, Molto AG, Heredia IG, Cabrera R, Martin J, Verdu M, Deschamps P, Moreira D, Lopez-Garcia P, Mira A, Rodriguez-Valera F (2010) Metagenome of the Mediterranean deep chlorophyll maximum studied by direct and fosmid library 454 pyrosequencing. ISME J 4:1154–1166

Gilmour M, Graham M, Van Domselaar G, Tyler S, Kent H, Trout-Yakel K, Larios O, Allen V, Lee B, Nadon C (2010) High-throughput genome sequencing of two *Listeria monocytogenes* clinical isolates during a large foodborne outbreak. BMC Genom 11:120

Grad YH, Lipsitch M, Feldgarden M, Arachchi HM, Cerqueira GC, FitzGerald M, Godfrey P, Haas BJ, Murphy CI, Russ C (2012) Genomic epidemiology of the *Escherichia coli* O104: H4 outbreaks in Europe, 2011. Proc Nat Acad Sci U.S.A. 109:3065–3070

Grad YH, Godfrey P, Cerquiera GC, Mariani-Kurkdjian P, Gouali M, Bingen E, Shea TP, Haas BJ, Griggs A, Young S, Zeng Q, Lipsitch M, Waldor MK, Weill FX, Wortman JR, Hanage WP (2013) Comparative genomics of recent shiga toxin-producing *Escherichia coli* O104: H4: short-term evolution of an emerging pathogen. mBio 4

Grad YH, Kirkcaldy RD, Trees D, Dordel J, Harris SR, Goldstein E, Weinstock H, Parkhill J, Hanage WP, Bentley S, Lipsitch M (2014) Genomic epidemiology of *Neisseria gonorrhoeae* with reduced susceptibility to cefixime in the USA: a retrospective observational study. Lancet Infect Dis 14:220–226

Guerra-Assunção JA, Crampin AC, Houben RMGJ, Mzembe T, Mallard K, Coll F, Khan P, Banda L, Chiwaya A, Pereira RPA, McNerney R, Fine PEM, Parkhill J, Clark TG, Glynn JR (2015) Large-scale whole genome sequencing of *M. tuberculosis* provides insights into transmission in a high prevalence area. eLife Sci 4:e05166

Harris SR, Feil EJ, Holden MTG, Quail MA, Nickerson EK, Chantratita N, Gardete S, Tavares A, Day N, Lindsay JA, Edgeworth JD, de Lencastre H, Parkhill J, Peacock SJ, Bentley SD (2010) Evolution of MRSA during hospital transmission and intercontinental spread. Science 327:469–474

Harris SR, Clarke IN, Seth-Smith HMB, Solomon AW, Cutcliffe LT, Marsh P, Skilton RJ, Holland MJ, Mabey D, Peeling RW, Lewis DA, Spratt BG, Unemo M, Persson K, Bjartling C,

Brunham R, de Vries HJC, Morre SA, Speksnijder A, Bebear CM, Clerc M, de Barbeyrac B, Parkhill J, Thomson NR (2012) Whole-genome analysis of diverse *Chlamydia trachomatis* strains identifies phylogenetic relationships masked by current clinical typing. Nat Genet 44:413–419

Harris SR, Cartwright EJ, Török ME, Holden MT, Brown NM, Ogilvy-Stuart AL, Ellington MJ, Quail MA, Bentley SD, Parkhill J, Peacock SJ (2013) Whole-genome sequencing for analysis of an outbreak of meticillin-resistant *Staphylococcus aureus*: a descriptive study. Lancet Infect Dis 13:130–136

Hasan NA, Choi SY, Eppinger M, Clark PW, Chen A, Alam M, Haley BJ, Taviani E, Hine E, Su Q, Tallon LJ, Prosper JB, Furth K, Hoq MM, Li H, Fraser-Liggett CM, Cravioto A, Huq A, Ravel J, Cebula TA, Colwell RR (2012) Genomic diversity of 2010 Haitian cholera outbreak strains. PNAS 109:E2010–E2017

He M, Miyajima F, Roberts P, Ellison L, Pickard DJ, Martin MJ, Connor TR, Harris SR, Fairley D, Bamford KB, D'Arc S, Brazier J, Brown D, Coia JE, Douce G, Gerding D, Kim HJ, Koh TH, Kato H, Senoh M, Louie T, Michell S, Butt E, Peacock SJ, Brown NM, Riley T, Songer G, Wilcox M, Pirmohamed M, Kuijper E, Hawkey P, Wren BW, Dougan G, Parkhill J, Lawley TD (2013) Emergence and global spread of epidemic healthcare-associated *Clostridium difficile*. Nat Genet 45:109–113

Hendriksen RS, Price LB, Schupp JM, Gillece JD, Kaas RS, Engelthaler DM, Bortolaia V, Pearson T, Waters AE, Upadhyay BP (2011) Population genetics of *Vibrio cholerae* from Nepal in 2010: evidence on the origin of the Haitian outbreak. mBio 2:e00157–11

Holden MTG, Hauser H, Sanders M, Ngo TH, Cherevach I, Cronin A, Goodhead I, Mungall K, Quail MA, Price C, Rabbinowitsch E, Sharp S, Croucher NJ, Chieu TB, Thi Hoang Mai N, Diep TS, Chinh NT, Kehoe M, Leigh JA, Ward PN, Dowson CG, Whatmore AM, Chanter N, Iversen P, Gottschalk M, Slater JD, Smith HE, Spratt BG, Xu J, Ye C, Bentley S, Barrell BG, Schultsz C, Maskell DJ, Parkhill J (2009) Rapid evolution of virulence and drug resistance in the emerging zoonotic pathogen *Streptococcus suis*. PLoS ONE 4:e6072

Holden MTG, Hsu LY, Kurt K, Weinert LA, Mather AE, Harris SR, Strommenger B, Layer F, Witte W, de Lencastre H, Skov R, Westh H, Zemlickova H, Coombs G, Kearns AM, Hill RLR, Edgeworth J, Gould I, Gant V, Cooke J, Edwards GF, McAdam PR, Templeton KE, McCann A, Zhou Z, Castillo-Ramirez S, Feil EJ, Hudson LO, Enright MC, Balloux F, Aanensen DM, Spratt BG, Fitzgerald JR, Parkhill J, Achtman M, Bentley SD, Nübel U (2013) A genomic portrait of the emergence, evolution and global spread of a methicillin resistant *Staphylococcus aureus* pandemic. Genome Res 23:653–664

Holmes A, Allison L, Ward M, Dallman TJ, Clark R, Fawkes A, Murphy L, Hanson M (2015) Utility of whole-genome sequencing of *Escherichia coli* O157 for outbreak detection and epidemiological surveillance. J Clin Microbiol 53:3565–3573

Holt KE, Parkhill J, Mazzoni CJ, Roumagnac P, Weill FX, Goodhead I, Rance R, Baker S, Maskell DJ, Wain J, Dolecek C, Achtman M, Dougan G (2008) High-throughput sequencing provides insights into genome variation and evolution in *Salmonella* Typhi. Nat Genet 40:987–993

Holt KE, Baker S, Weill FX, Holmes EC, Kitchen A, Yu J, Sangal V, Brown DJ, Coia JE, Kim DW, Choi SY, Kim SH, da Silveira WD, Pickard DJ, Farrar JJ, Parkhill J, Dougan G, Thomson NR (2012) *Shigella sonnei* genome sequencing and phylogenetic analysis indicate recent global dissemination from Europe. Nat Genet 44:1056–1059

Holt KE, Thieu Nga TV, Thanh DP, Vinh H, Kim DW, Vu Tra MP, Campbell JI, Hoang NVM, Vinh NT, Minh PV, Thuy CT, Nga TTT, Thompson C, Dung TTN, Nhu NTK, Vinh PV, Tuyet PTN, Phuc HL, Lien NTN, Phu BD, Ai NTT, Tien NM, Dong N, Parry CM, Hien TT, Farrar JJ, Parkhill J, Dougan G, Thomson NR, Baker S (2013) Tracking the establishment of local endemic populations of an emergent enteric pathogen. Proc Nat Acad Sci U.S.A. 110:17522–17527

Hornsey M, Loman N, Wareham DW, Ellington MJ, Pallen MJ, Turton JF, Underwood A, Gaulton T, Thomas CP, Doumith M (2011) Whole-genome comparison of two *Acinetobacter*

baumannii isolates from a single patient, where resistance developed during tigecycline therapy. J Antimicrob Chemother 66:1499–1503

Ioerger TR, Feng Y, Chen X, Dobos KM, Victor TC, Streicher EM, Warren RM, Van Pittius NCG, Helden PD, Sacchettini JC (2010) The non-clonality of drug resistance in Beijing-genotype isolates of *Mycobacterium tuberculosis* from the Western Cape of South Africa. BMC Genom 11:1

Jolley KA, Hill DMC, Bratcher HB, Harrison OB, Feavers IM, Parkhill J, Maiden MCJ (2012) Resolution of a meningococcal disease outbreak from whole-genome sequence data with rapid web-based analysis methods. J Clin Microbiol 50:3046–3053

Ju W, Cao G, Rump L, Strain E, Luo Y, Timme R, Allard M, Zhao S, Brown E, Meng J (2012) Phylogenetic analysis of non-O157 Shiga toxin-producing *Escherichia coli* by whole genome sequencing. J Clin Microbiol

Kanamori H, Parobek CM, Weber DJ, Van Duin D, Rutala WA, Cairns BA, Juliano JJ (2016) Next-generation sequencing and comparative analysis of sequential outbreaks caused by multidrug-resistant *Acinetobacter baumannii* at a large academic burn center. Antimicrob Agents Chemother 60:1249–1257

Kato-Maeda M, Ho C, Passarelli B, Banaei N, Grinsdale J, Flores L, Anderson J, Murray M, Rose G, Kawamura LM, Pourmand N, Tariq MA, Gagneux S, Hopewell PC (2013) Use of whole genome sequencing to determine the microevolution of *Mycobacterium tuberculosis* during an outbreak. PLoS ONE 8:e58235

Kennemann L, Didelot X, Aebischer T, Kuhn S, Drescher B, Droege M, Reinhardt R, Correa P, Meyer TF, Josenhans C, Falush D, Suerbaum S (2011) *Helicobacter pylori* genome evolution during human infection. Proc Nat Acad Sci U.S.A. 108:5033–5038

Kinnevey PM, Shore AC, Mac Aogáin M, Creamer E, Brennan GI, Humphreys H, Rogers TR, O'Connell B, Coleman DC (2016) Enhanced tracking of nosocomial transmission of endemic sequence type 22 methicillin-resistant *Staphylococcus aureus* type iv isolates among patients and environmental sites by use of whole-genome sequencing. J Clin Microbiol 54:445–448

Knetsch CW, Connor TR, Mutreja A, Van Dorp SM, Sanders IM, Browne HP, Harris D, Lipman L, Keessen EC, Corver J (2014) Whole genome sequencing reveals potential spread of *Clostridium difficile* between humans and farm animals in the Netherlands, 2002 to 2011. EuroSurveillance 19:30–41

Köser CU, Holden MTG, Ellington MJ, Cartwright EJP, Brown NM, Ogilvy-Stuart AL, Hsu LY, Chewapreecha C, Croucher NJ, Harris SR (2012) Rapid whole-genome sequencing for investigation of a neonatal MRSA outbreak. N Engl J Med 366:2267–2275

Köser CU, Bryant JM, Becq J, Török ME, Ellington MJ, Marti-Renom MA, Carmichael AJ, Parkhill J, Smith GP, Peacock SJ (2013) Whole-genome sequencing for rapid susceptibility testing of *M. tuberculosis*. N Engl J Med 369:290–292

Kwong JC, Mercoulia K, Tomita T, Easton M, Li HY, Bulach DM, Stinear TP, Seemann T, Howden BP (2016) Prospective whole-genome sequencing enhances national surveillance of *Listeria monocytogenes*. J Clin Microbiol 54:333–342

Lee RS, Radomski N, Proulx JF, Manry J, McIntosh F, Desjardins F, Soualhine H, Domenech P, Reed MB, Menzies D, Behr MA (2015) Reemergence and amplification of tuberculosis in the Canadian Arctic. J Infect Dis 211:1905–1914

Lévesque S, Plante PL, Mendis N, Cantin P, Marchand G, Charest H, Raymond F, Huot C, Goupil-Sormany I, Desbiens F (2014) Genomic characterization of a large outbreak of *Legionella pneumophila* serogroup 1 strains in Quebec City, 2012. PLoS ONE 9:e103852

Lewis T, Loman NJ, Bingle L, Jumaa P, Weinstock GM, Mortiboy D, Pallen MJ (2010) High-throughput whole-genome sequencing to dissect the epidemiology of *Acinetobacter baumannii* isolates from a hospital outbreak. J Hosp Infect 75:37–41

Lienau EK, Strain E, Wang C, Zheng J, Ottesen AR, Keys CE, Hammack TS, Musser SM, Brown EW, Allard MW, Cao G, Meng J, Stones R (2011) Identification of a salmonellosis outbreak by means of molecular sequencing. N Engl J Med 364:981–982

Liu Q, Via LE, Luo T, Liang L, Liu X, Wu S, Shen Q, Wei W, Ruan X, Yuan X, Zhang G, Barry CE, Gao Q (2015) Within patient microevolution of *Mycobacterium tuberculosis* correlates with heterogeneous responses to treatment. Scientific reports 5:17507

Loman NJ, Constantinidou C (2013) A culture-independent sequence-based metagenomics approach to the investigation of an outbreak of shiga-toxigenic *Escherichia coli* O104: H4. JAMA 309:1502–1510

Loman NJ, Pallen MJ (2015). Twenty years of bacterial genome sequencing. Nat Rev Micro

Loman NJ, Gladstone RA, Constantinidou C, Tocheva AS, Jefferies JM, Faust SN, OΓÇÖConnor L, Chan J, Pallen MJ, Clarke SC (2013) Clonal expansion within pneumococcal serotype 6C after use of seven-valent vaccine. PLoS ONE 8:e64731

Maiden MC, Bygraves JA, Feil E, Morelli G, Russell JE, Urwin R, Zhang Q, Zhou J, Zurth K, Caugant DA, Feavers IM, Achtman M, Spratt BG (1998) Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. Proc Nat Acad Sci U.S.A. 95:3140–3145

Maixner F, Krause-Kyora B, Turaev D, Herbig A, Hoopmann MR, Hallows JL, Kusebauch U, Vigl EE, Malfertheiner P, Megraud F, OΓÇÖSullivan N, Cipollini G, Coia V, Samadelli M, Engstrand L, Linz B, Moritz RL, Grimm R, Krause J, Nebel A, Moodley Y, Rattei T, Zink A (2016) The 5300-year-old Helicobacter pylori genome of the Iceman. Science 351:162–165

Mathers AJ, Peirano G, Pitout JDD (2015) The role of epidemic resistance plasmids and international high-risk clones in the spread of multidrug-resistant enterobacteriaceae. Clin Microbiol Rev 28:565–591

McAdam PR, Templeton KE, Edwards GF, Holden MTG, Feil EJ, Aanensen DM, Bargawi HJA, Spratt BG, Bentley SD, Parkhill J, Enright MC, Holmes A, Girvan EK, Godfrey PA, Feldgarden M, Kearns AM, Rambaut A, Robinson DA, Fitzgerald JR (2012) Molecular tracing of the emergence, adaptation, and transmission of hospital-associated methicillin-resistant *Staphylococcus aureus*. Proc Nat Acad Sci U.S.A. 109:9107–9112

McAdam P, Vander broek C, Lindsay D, Ward M, Hanson M, Gillies M, Watson M, Stevens J, Edwards G, Fitzgerald R (2014) Gene flow in environmental *Legionella pneumophila* leads to genetic and pathogenic heterogeneity within a Legionnaires' disease outbreak. Genome Biol 15:504

McDonnell J, Dallman T, Atkin S, Turbitt DA, Connor TR, Grant KA, Thomson NR, Jenkins C (2013) Retrospective analysis of whole genome sequencing compared to prospective typing data in further informing the epidemiological investigation of an outbreak of *Shigella sonnei* in the UK. Epidemiol Infect 141:2568–2575

Medini D, Serruto D, Parkhill J, Relman DA, Donati C, Moxon R, Falkow S, Rappuoli R (2008) Microbiology in the post-genomic era. Nat Rev Micro 6:419–430

Mellmann A, Harmsen D, Cummings CA, Zentz EB, Leopold SR, Rico A, Prior K, Szczepanowski R, Ji Y, Zhang W (2011) Prospective genomic characterization of the German enterohemorrhagic *Escherichia coli* O104: H4 outbreak by rapid next generation sequencing technology. PLoS ONE 6:e22751

Mendum T, Schuenemann V, Roffey S, Taylor G, Wu H, Singh P, Tucker K, Hinds J, Cole S, Kierzek A, Nieselt K, Krause J, Stewart G (2014) *Mycobacterium leprae* genomes from a British medieval leprosy hospital: towards understanding an ancient epidemic. BMC Genom 15:270

Metzker ML (2010) Sequencing technologies—the next generation. Nat Rev Genet 11:31–46

Mortimer PP (2003) Five postulates for resolving outbreaks of infectious disease. J Med Microbiol 52:447–451

Mutreja A, Kim DW, Thomson NR, Connor TR, Lee JH, Kariuki S, Croucher NJ, Choi SY, Harris SR, Lebens M, Niyogi SK, Kim EJ, Ramamurthy T, Chun J, Wood JLN, Clemens JD, Czerkinsky C, Nair GB, Holmgren J, Parkhill J, Dougan G (2011) Evidence for several waves of global transmission in the seventh cholera pandemic. Nature 477:462–465

Mwangi MM, Wu SW, Zhou Y, Sieradzki K, de Lencastre H, Richardson P, Bruce D, Rubin E, Myers E, Siggia ED, Tomasz A (2007) Tracking the in vivo evolution of multidrug resistance

in *Staphylococcus aureus* by whole-genome sequencing. Proc Nat Acad Sci U.S.A. 104:9451–9456

Niemann S, Köser CU, Gagneux S, Plinke C, Homolka S, Bignell H, Carter RJ, Cheetham RK, Cox A, Gormley NA, Kokko-Gonzales P, Murray LJ, Rigatti R, Smith VP, Arends FPM, Cox HS, Smith G, Archer JAC (2009) Genomic diversity among drug sensitive and multidrug resistant isolates of *Mycobacterium tuberculosis* with identical DNA fingerprints. PLoS ONE 4:e7407

Nübel U, Nachtnebel M, Falkenhorst G, Benzler J, Hecht J, Kube M, Bröcker F, Moelling K, Bührer C, Gastmeier P, Piening B, Behnke M, Dehnert M, Layer F, Witte W, Eckmanns T (2013) MRSA transmission on a neonatal intensive care unit: epidemiological and genome-based phylogenetic analyses. PLoS ONE 8:e54898

Okoro CK, Kingsley RA, Connor TR, Harris SR, Parry CM, Al-Mashhadani MN, Kariuki S, Msefula CL, Gordon MA, de Pinna E (2012) Intracontinental spread of human invasive *Salmonella* Typhimurium pathovariants in sub-Saharan Africa. Nat Genet 44:1215–1223

Onori R, Gaiarsa S, Comandatore F, Pongolini S, Brisse S, Colombo A, Cassani G, Marone P, Grossi P, Minoja G, Bandi C, Sassera D, Toniolo A (2015) Tracking nosocomial *Klebsiella pneumoniae* infections and outbreaks by whole-genome analysis: small-scale Italian scenario within a single hospital. J Clin Microbiol 53:2861–2868

Parwati I, Van Crevel R, Van Soolingen D (2010) Possible underlying mechanisms for successful emergence of the *Mycobacterium tuberculosis* Beijing genotype strains. Lancet Infect Dis 10:103–111

Paterson GK, Harrison EM, Murray GGR, Welch JJ, Warland JH, Holden MTG, Morgan FJE, Ba X, Koop G, Harris SR, Maskell DJ, Peacock SJ, Herrtage ME, Parkhill J, Holmes MA (2015) Capturing the cloud of diversity reveals complexity and heterogeneity of MRSA carriage, infection and transmission. Nat Commun 6:6560

Pérez-Lago L, Comas I, Navarro Y, González-Candelas F, Herranz M, Bouza E, García de Viedma D (2014) Whole genome sequencing analysis of intrapatient microevolution in *Mycobacterium tuberculosis*: potential impact on the inference of tuberculosis transmission. J Infect Dis 209:98–108

Pérez-Lago L, Martínez Lirola M, Herranz M, Comas I, Bouza E, García-de-Viedma D (2015) Fast and low-cost decentralized surveillance of transmission of tuberculosis based on strain-specific PCRs tailored from whole genome sequencing data: a pilot study. Clin Microbiol Infect 21:249

Pinholt M, Larner-Svensson H, Littauer P, Moser CE, Pedersen M, Lemming LE, Ejlertsen T, Söndergaard TS, Holzknecht BJ, Justesen US, Dzajic E, Olsen SS, Nielsen JB, Worning P, Hammerum AM, Westh H, Jakobsen L (2015) Multiple hospital outbreaks of *vanA Enterococcus faecium* in Denmark, 2012–13, investigated by WGS, MLST and PFGE. J Antimicrob Chemother 70:2474–2482

Price JR, Golubchik T, Cole K, Wilson DJ, Crook DW, Thwaites GE, Bowden R, Sarah Walker A, Peto TEA, Paul J, Llewelyn MJ (2014) Whole-genome sequencing shows that patient-to-patient transmission rarely accounts for acquisition of *Staphylococcus aureus* on an intensive care unit. Clin Infect Dis 58:609–618

Qin T, Zhang W, Liu W, Zhou H, Ren H, Shao Z, Lan R, Xu J (2016) Population structure and minimum core genome typing of *Legionella pneumophila*. Sci Rep 6:21356

Quick J, Ashton P, Calus S, Chatt C, Gossain S, Hawker J, Nair S, Neal K, Nye K, Peters T, de Pinna E, Robinson E, Struthers K, Webber M, Catto A, Dallman T, Hawkey P, Loman N (2015) Rapid draft sequencing and real-time nanopore sequencing in a hospital outbreak of *Salmonella*. Genome Biol 16:114

Quick J, Loman NJ, Duraffour S, Simpson JT, Severi E, Cowley L, Bore JA, Koundouno R, Dudas G, Mikhail A, Ouédraogo N, Afrough B, Bah A, Baum JHJ, Becker-Ziaja B, Boettcher JP, Cabeza-Cabrerizo M, Camino-Sánchez Á, Carter LL, Doerrbecker J, Enkirch T, Dorival IGa, Hetzelt N, Hinzmann J, Holm T, Kafetzopoulou LE, Koropogui M, Kosgey A, Kuisma E, Logue CH, Mazzarelli A, Meisel S, Mertens M, Michel J, Ngabo D, Nitzsche K, Pallasch E, Patrono LV, Portmann J, Repits JG, Rickett NY, Sachse A, Singethan K, Vitoriano

Is, Yemanaberhan RL, Zekeng EG, Racine T, Bello A, Sall AA, Faye O, Faye O, Magassouba N, Williams CV, Amburgey V, Winona L, Davis E, Gerlach J, Washington F, Monteil V, Jourdain M, Bererd M, Camara A, Somlare H, Camara A, Gerard M, Bado G, Baillet B, Delaune D, Nebie KY, Diarra A, Savane Y, Pallawo RB, Gutierrez GJ, Milhano N, Roger I, Williams CJ, Yattara F, Lewandowski K, Taylor J, Rachwal P, Turner J, Pollakis G, Hiscox JA, Matthews DA, Shea MKO, Johnston AM, Wilson D, Hutley E, Smit E, Di Caro A, Wölfel R, Stoecker K, Fleischmann E, Gabriel M, Weller SA, Koivogui L, Diallo B, Keïta S, Rambaut A, Formenty P, Günther S, Carroll MW (2016). Real-time, portable genome sequencing for Ebola surveillance. Nature

Rasmussen S, Allentoft ME, Nielsen K, Orlando L, Sikora M, Sjögren KG, Pedersen AG, Schubert M, Van Dam A, Kapel CMO, Nielsen HB, Brunak S, Avetisyan P, Epimakhov A, Khalyapin MV, Gnuni A, Kriiska A, Lasak I, Metspalu M, Moiseyev V, Gromov A, Pokutta D, Saag L, Varul L, Yepiskoposyan L, Sicheritz-Pontén T, Foley RA, Lahr MM, Nielsen R, Kristiansen K, Willerslev E (2015) Early divergent strains of *Yersinia pestis* in Eurasia 5000 years ago. Cell 163:571–582

Read TD, Salzberg SL, Pop M, Shumway M, Umayam L, Jiang L, Holtzapple E, Busch JD, Smith KL, Schupp JM, Solomon D, Keim P, Fraser CM (2002) Comparative genome sequencing for discovery of novel polymorphisms in *Bacillus anthracis*. Science 296:2028–2033

Reuter S, Ellington MJ, Cartwright EP (2013a) Rapid bacterial whole-genome sequencing to enhance diagnostic and public health microbiology. JAMA Intern Med 173:1397–1404

Reuter S, Harrison TG, Köser CU, Ellington MJ, Smith GP, Parkhill J, Peacock SJ, Bentley SD, Török ME (2013b) A pilot study of rapid whole-genome sequencing for the investigation of a *Legionella* outbreak. BMJ Open 3

Reuter S, Török ME, Holden MTG, Reynolds R, Raven KE, Blane B, Donker T, Bentley SD, Aanensen DM, Grundmann H, Feil EJ, Spratt BG, Parkhill J, Peacock SJ (2016) Building a genomic framework for prospective MRSA surveillance in the United Kingdom and the Republic of Ireland. Genome Res 26:263–270

Rinder H, Mieskes KT, Löscher T (2001) Heteroresistance in *Mycobacterium tuberculosis*. Int J Tuberc Lung Dis 5:339–345

Roetzer A, Diel R, Kohl TA, Rückert C, Nübel U, Blom J, Wirth T, Jaenicke S, Schuback S, Rüsch-Gerdes S, Supply P, Kalinowski J, Niemann S (2013) Whole genome sequencing versus traditional genotyping for investigation of a *Mycobacterium tuberculosis* outbreak: a longitudinal molecular epidemiological study. PLoS Med 10:e1001387

Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, Zhivotovsky LA, Feldman MW (2002) Genetic structure of human populations. Science 298:2381–2385

Sánchez-Busó L, Comas I, Jorques G, González-Candelas F (2014) Recombination drives genome evolution in outbreak-related *Legionella pneumophila* isolates. Nat Genet 46:1205–1211

Sánchez-Busó L, Guiral S, Crespi S, Moya V, Camaró ML, Olmos P, Adrián F, Morera V, González Morán F, Vanaclocha H, González-Candelas F (2016) Genomic investigation of a legionellosis outbreak in a persistently colonized hotel. Front Microbiol 6:1556

Sandegren L, Groenheit R, Koivula T, Ghebremichael S, Advani A, Castro E, Pennhag A, Hoffner S, Mazurek J, Pawlowski A, Kan B, Bruchfeld J, Melefors Ö, Källenius G (2011) Genomic stability over 9 Years of an Isoniazid resistant *Mycobacterium tuberculosis* outbreak strain in Sweden. PLoS ONE 6:e16647

Schlüter A, Bekel T, Diaz NN, Dondrup M, Eichenlaub R, Gartemann KH, Krahn I, Krause L, Krömeke H, Kruse O, Mussgnug JH, Neuweger H, Niehaus K, Pühler A, Runte KJ, Szczepanowski R, Tauch A, Tilker A, Viehöver P, Goesmann A (2008a) The metagenome of a biogas-producing microbial community of a production-scale biogas plant fermenter analysed by the 454-pyrosequencing technology. J Biotechnol 136:77–90

Schlüter A, Krause L, Szczepanowski R, Goesmann A, Pühler A (2008b) Genetic diversity and composition of a plasmid metagenome from a wastewater treatment plant. J Biotechnol 136:65–76

Schmid D, Allerberger F, Huhulescu S, Pietzka A, Amar C, Kleta S, Prager R, Preussel K, Aichinger E, Mellmann A (2014) Whole genome sequencing as a tool to investigate a cluster of seven cases of listeriosis in Austria and Germany, 2011–2013. Clin Microbiol Infect 20: 431–436

Schuenemann VJ, Singh P, Mendum TA, Krause-Kyora B, Jäger G, Bos KI, Herbig A, Economou C, Benjak A, Busso P, Nebel A, Boldsen JL, Kjellström A, Wu H, Stewart GR, Taylor GM, Bauer P, Lee OYC, Wu HHT, Minnikin DE, Besra GS, Tucker K, Roffey S, Sow SO, Cole ST, Nieselt K, Krause J (2013) Genome-wide comparison of medieval and modern *Mycobacterium leprae*. Science 341:179–183

Schürch AC, Kremer K, Kiers A, Daviena O, Boeree MJ, Siezen RJ, Smith NH, Van Soolingen D (2010) The tempo and mode of molecular evolution of *Mycobacterium tuberculosis* at patient-to-patient scale. Infect Genet Evol 10:108–114

Senn L, Clerc O, Zanetti G, Basset P, Prod'hom G, Gordon NC, Sheppard AE, Crook DW, James R, Thorpe HA, Feil EJ, Blanc DS (2016) The stealthy superbug: the role of asymptomatic enteric carriage in maintaining a long-term hospital outbreak of ST228 Methicillin-Resistant Staphylococcus aureus. mBio 7

Seth-Smith HMB, Harris SR, Skilton RJ, Radebe FM, Golparian D, Shipitsyna E, Duy PT, Scott P, Cutcliffe LT, O'Neill C, Parmar S, Pitt R, Baker S, Ison CA, Marsh P, Jalal H, Lewis DA, Unemo M, Clarke IN, Parkhill J, Thomson NR (2013) Whole-genome sequences of *Chlamydia trachomatis* directly from clinical samples without culture. Genome Res 23:855–866

Shah MA, Mutreja A, Thimson N, Baker S, Parkhill J, Dougan G, Bokhari H, Wren BW (2014) Genomic epidemiology of *Vibrio cholerae* O1 associated with floods, Pakistan, 2010. Emerg Infect Dis 20:13–20

Smit PW, Vasankari T, Aaltonen H, Haanperä M, Casali N, Marttila H, Marttila J, Ojanen P, Ruohola A, Ruutu P, Drobniewski F, Lyytikäinen O, Soini H (2015) Enhanced tuberculosis outbreak investigation using whole genome sequencing and IGRA. Eur Respir J 45:276–279

Snitkin ES, Zelazny AM, Thomas PJ, Stock F, NISC Comparative Sequencing Program, Henderson DK, Palmore TN, Segre JA (2012) Tracking a hospital outbreak of carbapenem-resistant *Klebsiella pneumoniae* with whole-genome sequencing. Sci Transl Med 4:148ra116

Snyder LA, Loman NJ, Faraj LA, Levi K, Weinstock G, Boswell TC, Pallen MJ, Ala'Aldeen A (2013) Epidemiological investigation of *Pseudomonas aeruginosa* isolates from a six-year-long hospital outbreak using high-throughput whole genome sequencing. EuroSurveillance 18:20611

Stucki D, Ballif M, Bodmer T, Coscolla M, Maurer AM, Droz S, Butz C, Borrell S, Längle C, Feldmann J, Furrer H, Mordasini C, Helbling P, Rieder HL, Egger M, Gagneux S, Fenner L (2015) Tracking a tuberculosis outbreak over 21 years: strain-specific single-nucleotide polymorphism typing combined with targeted whole-genome sequencing. J Infect Dis 211:1306–1316

Sun G, Luo T, Yang C, Dong X, Li J, Zhu Y, Zheng H, Tian W, Wang S, Barry CE, Mei J, Gao Q (2012) Dynamic population changes in *Mycobacterium tuberculosis* during acquisition and fixation of drug resistance in patients. J Infect Dis 206:1724–1733

Taylor AJ, Lappi V, Wolfgang WJ, Lapierre P, Palumbo MJ, Medus C, Boxrud D (2015) Characterization of foodborne outbreaks of *Salmonella enterica* serovar enteritidis with whole-genome sequencing single nucleotide polymorphism-based analysis for surveillance and outbreak detection. J Clin Microbiol 53:3334–3340

Török ME, Peacock SJ (2012) Rapid whole-genome sequencing of bacterial pathogens in the clinical microbiology laboratory-pipe dream or reality? J Antimicrob Chemother 67:2307–2308

Török ME, Reuter S, Bryant J, Köser CU, Stinchcombe SV, Nazareth B, Ellington MJ, Bentley SD, Smith GP, Parkhill J, Peacock SJ (2013) Rapid whole-genome sequencing for investigation of a suspected tuberculosis outbreak. J Clin Microbiol 51:611–614

Underwood A, Jones G, Mentasti M, Fry N, Harrison T (2013a) Comparison of the *Legionella pneumophila* population structure as determined by sequence-based typing and whole genome sequencing. BMC Microbiol 13:302

Underwood AP, Dallman T, Thomson NR, Williams M, Harker K, Perry N, Adak B, Willshaw G, Cheasty T, Green J (2013b) Public health value of next-generation DNA sequencing of enterohemorrhagic *Escherichia coli* isolates from an outbreak. J Clin Microbiol 51:232–237

Urwin R, Maiden MCJ (2003) Multi-locus sequence typing: a tool for global epidemiology. Trends Microbiol 11:479–487

Vos M, Didelot X (2009) A comparison of homologous recombination rates in bacteria and archaea. ISME J 3:199–208

Wagner DM, Klunk J, Harbeck M, Devault A, Waglechner N, Sahl JW, Enk J, Birdsell DN, Kuch M, Lumibao C, Poinar D, Pearson T, Fourment M, Golding B, Riehm JM, Earn DJD, DeWitte S, Rouillard JM, Grupe G, Wiechmann I, Bliska JB, Keim PS, Scholz HC, Holmes EC, Poinar H (2014) *Yersinia pestis* and the Plague of Justinian 541–543 AD: a genomic analysis. Lancet Infect Dis 14:319–326

Walker MJ, Beatson SA (2012) Outsmarting outbreaks. Science 338:1161–1162

Walker TM, Monk P, Grace Smith E, Peto TEA (2013a) Contact investigations for outbreaks of *Mycobacterium tuberculosis*: advances through whole genome sequencing. Clin Microbiol Infect 19:796–802

Walker TM, Ip CL, Harrell RH, Evans JT, Kapatai G, Dedicoat MJ, Eyre DW, Wilson DJ, Hawkey PM, Crook DW, Parkhill J, Harris D, Walker AS, Bowden R, Monk P, Smith EG, Peto TE (2013b) Whole-genome sequencing to delineate *Mycobacterium tuberculosis* outbreaks: a retrospective observational study. Lancet Infect Dis 13:137–146

Walker TM, Lalor MK, Broda A, Ortega LS, Morgan M, Parker L, Churchill S, Bennett K, Golubchik T, Giess AP, Del Ojo Elias C, Jeffery KJ, Bowler ICJW, Laurenson IF, Barrett A, Drobniewski F, McCarthy ND, Anderson LF, Abubakar I, Thomas HL, Monk P, Smith EG, Walker AS, Crook DW, Peto TEA, Conlon CP (2014) Assessment of *Mycobacterium tuberculosis* transmission in Oxfordshire, UK, 2007–12, with whole pathogen genome sequences: an observational study. Lancet Respir Med 2:285–292

Ward MJ, Gibbons CL, McAdam PR, Van Bunnik BAD, Girvan EK, Edwards GF, Fitzgerald JR, Woolhouse MEJ (2014) Time-scaled evolutionary analysis of the transmission and antibiotic resistance dynamics of *Staphylococcus aureus* clonal complex 398. Appl Environ Microbiol 80:7275–7282

WHO (2014). Global tubreculosis report, 2014

Witney AA, Gould KA, Pope CF, Bolt F, Stoker NG, Cubbon MD, Bradley CR, Fraise A, Breathnach AS, Butcher PD, Planche TD, Hinds J (2014) Genome sequencing and characterization of an XDR ST111 serotype O12 hospital outbreak strain of *Pseudomonas aeruginosa*. Clin Microbiol Infect

Worby CJ, Lipsitch M, Hanage WP (2014) Within-host bacterial diversity hinders accurate reconstruction of transmission networks from genomic distance data. PLoS Comput Biol 10: e1003549

Yates TA, Khan PY, Knight GM, Taylor JG, McHugh TD, Lipman M, White RG, Cohen T, Cobelens FG, Wood R, Moore DAJ, Abubakar I (2016) The transmission of *Mycobacterium tuberculosis* in high burden settings. Lancet Infect Dis 16:227–238

Young BC, Golubchik T, Batty EM, Fung R, Larner-Svensson H, Votintseva AA, Miller RR, Godwin H, Knox K, Everitt RG (2012) Evolutionary dynamics of *Staphylococcus aureus* during progression from carriage to disease. Proc Nat Acad Sci U.S.A. 109:4550–4555

Zakour NLB, Venturini C, Beatson SA, Walker MJ (2012) Analysis of a *Streptococcus pyogenes* puerperal sepsis cluster by use of whole-genome sequencing. J Clin Microbiol 50:2224–2228

Zhou Y, Gao H, Mihindukulasuriya K, Rosa P, Wylie K, Vishnivetskaya T, Podar M, Warner B, Tarr P, Nelson D, Fortenberry JD, Holland M, Burr S, Shannon W, Sodergren E, Weinstock G (2013) Biogeography of the ecosystems of the healthy human body. Genome Biol 14:R1

Chapter 14
Three-dimensional Genomic Organization of Genes' Function in Eukaryotes

Alon Diament and Tamir Tuller

Abstract It is well known that in prokaryotes, genes are organized in transcription units called operons. Since each operon includes genes which are related to the same pathway, a relation between genomic proximity and functionality can be easily observed. In eukaryotes, usually there are no operons; however, in the last few decades, there have been growing evidence that the organization of eukaryotic genes is not random: Evolution shapes gene organization in eukaryotes in a way that will improve the organism's fitness. In this chapter, we will review how previous studies in the field employed sophisticated experiments and analysis tools to decipher the way genes are organized in eukaryotic genomes.

14.1 Introduction

Various studies in recent years demonstrated that all known biological network are modular—these networks can be divided into modules that include components with similar/shared functions and with relatively many interactions (Costanzo et al. 2010; Girvan and Newman 2002; Mitra et al. 2013; Pritykin and Singh 2013; Ravasz et al. 2002; Stuart et al. 2003). Specifically, it was shown that proteins within the same functional modules evolve at more similar rates than those between different modules (Chen and Dokholyan 2006). The expression level of such functionally related proteins also coevolves more strongly than that between different modules (Chen and Dokholyan 2006). Thus, we expect to see this modularity reflected at least to some extent in the organization of genomes.

A. Diament · T. Tuller
Department of Biomedical Engineering, Tel Aviv University (TAU), Tel Aviv, Israel

T. Tuller (✉)
The Sagol School of Neuroscience, Tel-Aviv University (TAU), Tel-Aviv, Israel
e-mail: tamirtul@post.tau.ac.il

P. Pontarotti (ed.), *Evolutionary Biology*, DOI 10.1007/978-3-319-41324-2_14

Understanding the importance of genome architecture, the arrangement of genes within the genome, and how this organization evolved has been intensively studied in recent years. It has become evident that the genomic architecture and thus the three-dimensional organization of genes in the genome are far from random. As opposed to displaying a homogenous distribution of genes along chromosomes, genomes appear to be specifically organized for gene regulation (Kosak and Groudine 2004). Coordinated cell growth and development require that cells regulate the expression of large sets of genes in an appropriate manner, and increasing evidence suggests that genomic organization has a major role in this regulation.

The aim of this book chapter was to briefly review the research and knowledge from the last few decades regarding the relation between genes' function and their (specifically 3D) genomic organization. It is organized as follows: We start with a brief discussion related to the differences between prokaryotes and eukaryotes in terms of genomic organization of genes; we continue with a short review about pioneer studies of the genomic organization of genes based on their functionality that were based on analyses of "linear" (1D) information about gene order. Next, we describe novel techniques that enable the estimation of the three-dimensional (3D) organizations of genomes and recent studies that used these data for understanding the 3D organization of genes based on their functionality. We conclude with discussion and future directions related to the study of the genomic organization of genes.

14.2 Genome and Gene Structure in Prokaryotes Versus Eukaryotes

The aim of this chapter was to review the genomic organization of genes in eukaryotes; thus, we would like to start with a brief comparison between eukaryotes and prokaryotes in term of their genomic features, gene structure, and gene expression (see also Table 14.1).

Perhaps the most striking difference between eukaryotes and prokaryotes is compartmentalization of the cell. Eukaryote cells contain a number of organelles, including the nucleus which contains most of the DNA in the cell (aside from the mitochondria and the chloroplast). Within the nucleus, processes are also compartmentalized, including transcription, replication, and DNA repair taking place in preferred nuclear regions (Misteli 2007).

Prokaryote cells are smaller in size compared with eukaryotes, and they contain a single chromosome but may contain many additional plasmids of small circular DNA. By contrast, eukaryote cells usually contain multiple chromosomes, with multiple copies of each chromosome (most often they are diploid). While prokaryotes are mostly unicellular, eukaryotes are usually multicellular, having tissues comprising of distinct cell types showing different phenotypes such as typical gene expression profiles and active pathways. While the genome in every

Table 14.1 Summary of differences between eukaryotic and prokaryotic genome/gene structure and gene expression

	Prokaryotes	Eukaryotes
Cell size	1–10 μm	10–100 μm
Number of cells	Mostly unicellular	Mostly multicellular, with multiple cell types
Genome size	0.6–10 Mbp	2.9–4000 Mbp
Number of genes	1000–7000	4000–30,000
Gene coding sequence (without introns) length (average)	924 bp	1346 bp
Percentage of genome containing genes	88 % (*E. coli*)	Incl. introns: 32 % (*H. sapiens*) w/o introns: 1.1 %
Number of chromosomes	One (additional plasmids)	More than one; usually contain 2 copies or more of each chromosome (diploid)
Chromosome packaging	Supercoils around HU proteins	DNA is wrapped around histones, coiled in a 30-nm fiber, and arranged in scaffolds, loops and domains
Nucleus	Absent	Present
Operons	mRNA is polycistronic	Rare. Each eukaryote gene is transcribed separately (monocistronic), with separate transcriptional controls on each gene
Introns	Very rare	Very common
Compartmentalization	Single compartment	Highly compartmentalized, containing multiple organelles (mitochondria, chloroplast, ER, etc.)
Translation	Shine–Dalgarno initiation sequence Co-transcriptional translation	5' cap initiation Transcription and translation occur in different compartments
Transcription	One type of RNA polymerase	Separate RNA polymerase for each type of RNA
Sexual reproduction	No meiosis	Involves meiosis
Selective pressure/effective population size	Strong/large	Weak/small

cell of the organism is identical, 3D genomic organization can change according to context and possibly have a role in regulating tissue-specific expression patterns.

Many earlier studies emphasized the higher levels of genomic organization in prokaryotes compared with eukaryotes; for example, it is known that prokaryotes (unlike most eukaryotes) tend to contain operons (Fig. 14.1a): clusters of adjacent, cotranscribed genes with related function and common regulation (Salgado et al. 2000); for example, many times an operon can include the genes encoding the proteins of a certain metabolic pathway. Thus, linear organization in prokaryotes is clearly detectable and serves a clear functional purpose. While in prokaryotes,

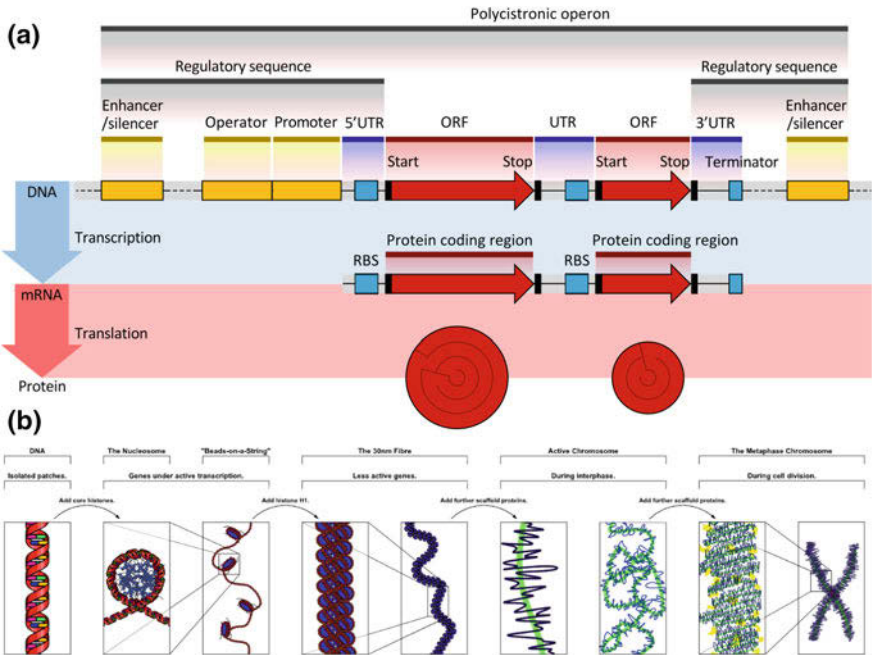


Fig. 14.1 **a** The structure of a prokaryotic operon of protein-coding genes. Regulatory sequence controls when expression occurs for the multiple protein-coding regions (*red*). Promoter, operator, and enhancer regions (*yellow*) regulate the transcription of the gene into an mRNA. The mRNA untranslated regions (*blue*) regulate translation into the final protein products (*red circles*). (Thomas Shafee /CC BY-SA 4.0) **b** Illustration of eukaryotes genomic conformation at various scales (Wikimedia Commons/CC BY-SA 3.0.)

multiple genes in operons are transcribed as a single (polycistronic) molecule, eukaryote genes are transcribed individually. Eukaryote genes usually contain many noncoding introns that are removed during splicing following transcription and prior to translation.

Eukaryote genomes are typically considerably larger than prokaryote genomes and are compactly packed within the nucleus. Eukaryote chromosome packaging includes several hierarchical folding stages (Fig. 14.1b): Wrapped DNA around histones forms nucleosomes, often described as "beads on a string," which is further coiled in a 30-nm fiber and so on.

An important feature of the eukaryotic nucleus is the organization of chromatin into distinct chromosome territories (Cremer et al. 2006) which are surrounded by the interchromatin compartment that is necessary for transport of regulatory molecules to the targeted DNA (Bártová and Kozubek 2006). The inner structure of the chromosome territories, as well as the arrangement of the chromosomes within the interphase nuclei, has been found to be non-randomly organized. It has been observed that genomes tend to have specific conformations and typical organization during different steps of the cell cycle, differentiation (Kim et al. 2004; Kosak et al.

2002), and during tumor cell transformation (Bártová and Kozubek 2006). Thus, eukaryotes show a complex genomic architecture that has yet to be fully understood. The next sections describe recent findings that shed light on eukaryotic gene organization in 1D and 3D.

14.3 Models Based on the Analysis of Linear Genomes

Early studies that analyzed genomic organization in eukaryotes focused on the linear organization of genes. That is, how closely they are positioned on the DNA sequence, and whether this one-dimensional (1D) order displays specific patterns. This approach was motivated by the abundance of identified operons in prokaryotes, as well as by the availability of genomic sequences and lack of large-scale methods to study higher dimensions of genomic organization.

It has long been observed that many co-expressed or similarly expressed genes, often with related function (Tuller et al. 2009) (Fig. 14.2a), appear to be linearly clustered in eukaryote genomes (Kosak and Groudine 2004) such as *S. cerevisiae* (Cohen et al. 2000), *C. elegans* (Miller et al. 2004; Roy et al. 2002), *D. melanogaster* (Weber and Hurst 2011), and human (Lercher et al. 2002). Many yeast genes coding for subunits of stable protein complexes are located within 10–30 kb of each other (Teichmann and Veitia 2004). The nematode worm *C. elegans* in particular was the subject of many studies due to the existence of polycistronic, prokaryote-like operons in its genome (Blumenthal et al. 2002). Furthermore, it was shown that multiple aspects of chromosome organization are related: For example, human housekeeping genes tend to be clustered in highly expressed, GC-rich regions (Martin and Lercher 2003).

These results have suggested that linear gene organization in eukaryotes has a functional role and that selection possibly acts to optimize it (Hurst et al. 2004; Sémon and Duret 2006). In support this hypothesis, it has been observed that gene expression clusters tend to contain fewer chromosomal breakpoints between human and mouse (Singer et al. 2005). Alternatively, it was suggested that coevolution between neighboring genes may drive their expression level in the same direction once one of the genes changes its activity, thus promoting the formation of co-expressed domains, as has been observed in a study of differential expression between mouse and human (Sémon and Duret 2006).

The organization of genes is indeed dynamic and changes over the course evolution. A number of known evolutionary events demonstrate the functional possibilities of repositioning of genes on chromosomes in coordination with a rewiring of their function (Field and Osbourn 2008; Lee and Sonnhammer 2003; Osbourn and Field 2009; Slot and Rokas 2010), for example, along a metabolic pathway (Wong and Wolfe 2005) (Fig. 14.2b). Interacting genes that appear separated in one genome may even become fused together into a single gene (Marcotte et al. 1999) (Fig. 14.2c).

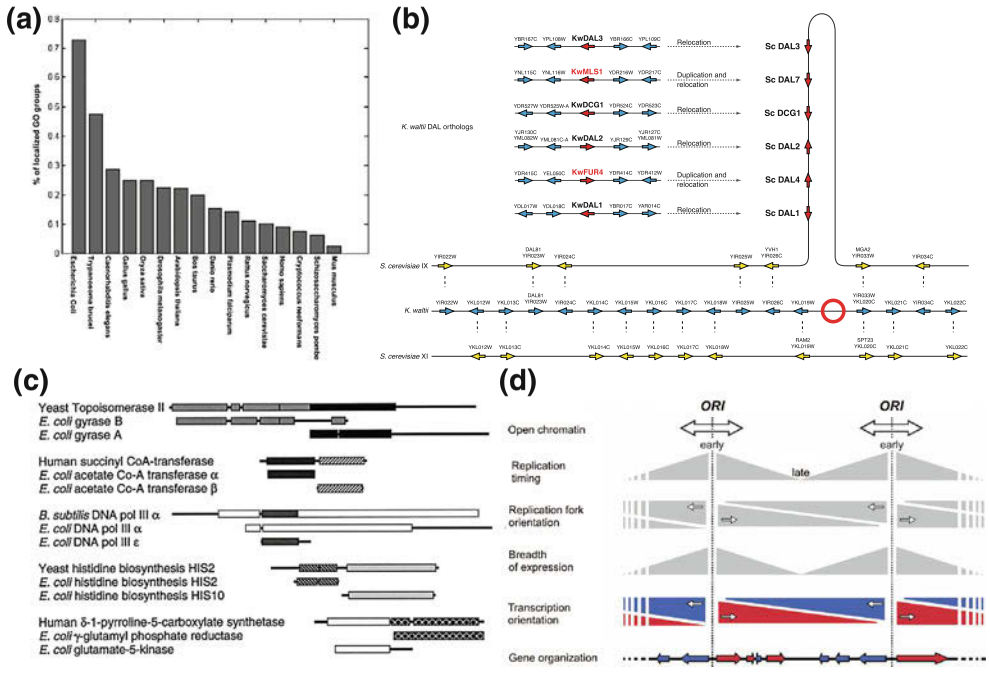


Fig. 14.2 **a** Linear clustering of GO groups in various organisms. (Tuller et al. Journal of Computational Biology, 2009.) **b** Locations of the rearranged metabolic pathway DAL genes in *K. waltii* as compared with *S. cerevisiae*. (Wong and Wolfe, Nature Genetics, 2005.) **c** Domain fusion analysis: The second and third proteins in each example are predicted to interact because their homologs are fused in the first protein. (Marcotte et al., Science 1999.) **d** Model of gene organization coordinated by replication and transcription (Huvet et al. Genome Research, 2007.)

Genomic organization also shows some level of conservation. It has been shown that the adjacency between pairs of genes is typically conserved in linages of *S. cerevisiae* (baker's yeast) for genes that are positioned significantly closer along the DNA sequence (Poyatos and Hurst 2007). Gene clustering and organization tends to be more similar between organisms in evolutionary proximity (Tuller et al. 2009). It has been also observed that eukaryotic genomes are enriched with tandem gene arrays, containing duplicated genes in linear proximity to one another (Kosak and Groudine 2004). In addition, conservation of organization appears to be correlated with conservation of expression patterns (Sémon and Duret 2006).

It is not only functionally related genes that cluster in eukaryotic genomes and specifically mammalian genomes: There also seems to be general clustering of apparently unrelated genes and that correlates with chromatin state (Sproul et al. 2005). Mammalian genomes display striking modularity in terms of gene expression, with the linear distribution of expression along chromosomes organized into regions of high and low levels of gene activity (Caron et al. 2001), with highly active regions separated by large regions of low activity (Kosak and Groudine 2004). This modularity is also related to regulation by heritable modifications known as epigenetic factors such as DNA methylation, histone modifications, and nucleosome remodeling. Such epigenetic factors are strongly related to chromatin-mediated control of gene expression (Bártová and Kozubek 2006). A partial explanation for this type gene clustering may be related to noise reduction in gene expression, by observing that essential, and possibly other noise-sensitive, genes tend to be located in open chromatin domains with low nucleosome occupancy and a stable level of expression (Batada and Hurst 2007). Essential genes have also been noted to cluster in regions with low recombination activity in yeast (Pál and Hurst 2003), and clusters of genetically linked genes have been observed in inbred mice strains (Petkov et al. 2007). Moreover, many additional forces take part in the shaping of the modular genomic organization, such as the coordination of replication and transcription (Huvet et al. 2007) (Fig. 14.2d).

This early evidence supports the notion that, at least in some cases, genomes are carefully organized and that selection acts on the positioning of genes within genomes. These earlier studies also demonstrated that it is possible to detect signals related to genomic organization of genes in eukaryotes via the *linear* analysis of the genomes. Despite the aforementioned evidence, gene clustering is still much less prevalent in eukaryotes compared with prokaryotes (Sémon and Duret 2006), and the mechanisms that maintain co-expression clusters are not well understood. Linear organization can only partially account for the modularity observed in expression patterns and cellular processes, as some pathways and functions are highly clustered, while many others are not (Field and Osbourn 2008; Lee and Sonnhammer 2003) and some reports maintain that clusters are often not conserved (Weber and Hurst 2011).

14.4 Experimental Techniques for Analysis of 3D Genomic Data

The size and density of chromosomes during interphase render them untraceable under a microscope. In contrast, the dense form that chromosomes take during cell division is more readily visualized and has biased research until late in the twentieth century toward this phase in the cell cycle. More elaborate approaches are needed to study chromosome organization in interphase. Fluorescent in situ hybridization (FISH) (Langer-Safer et al. 1982) and similar methods have been instrumental in gaining an early understanding of genomic organization in 3D and remain a common approach. The method utilizes target-specific DNA probes that hybridize with loci in the genome and then detected optically, using antibodies linked with fluorophore. FISH enables, using a combination of probes and dyes, to monitor a number of loci simultaneously, from whole chromosomes (Speicher et al. 1996) to specific sites of interests (Schmidt et al. 1994), and to estimate their typical organization in the nucleus, distances between loci and the dynamics of this organization in time (Belmont 2001). However, the main limitation of FISH is that the number loci that can be tracked is limited to a few (van der Ploeg 2000), thus hindering large-scale studies of genomic organization at the resolution of genes or, e.g., regulatory binding sites.

A more recent technology has revolutionized the study of 3D genomic organization by taking an indirect approach (compared with imaging) to monitor distances between loci. Chromosome conformation capture (3C) (Dekker et al. 2002) enabled monitoring the interaction frequency between genomic loci, i.e., how often they are observed in proximity in vivo in cell populations (unlike FISH, which is a single-cell approach). The method consists of the following steps (Fig. 14.3): Formaldehyde is used for cross-linking genomic loci that are in proximity in the nucleus. Then, the genome is digested using a restriction enzyme, followed by intramolecular ligation. Finally, cross-links are reversed and products are detected and quantified by the polymerase chain reaction (PCR) using locus-specific primers. Based on a set of loci distributed across the genome, a matrix of interaction frequencies between and within chromosomes can be generated and analyzed. It has been shown that the detected interaction frequency between pairs of loci is related to their linear distance on the chromosome (Dekker et al. 2002) as well as to their 3D measured distance via FISH (Lieberman-Aiden et al. 2009; Tanizawa et al. 2010), thus enabling the study of the 3D organization at unprecedented scale and resolution. However, in the 3C protocol, the detection step (using semiquantitative PCR) requires that the sites of interest will be predefined, thus severely limiting the throughput of the method. To alleviate this, modifications to the protocol have been suggested, including among others the circularized chromosome conformation capture (4C) (Simonis et al. 2006), which allows the study of contacts between one region of interest to all other regions of the genome. Another proposed modification was carbon copy chromosome conformation capture (5C) (Dostie et al. 2006) that

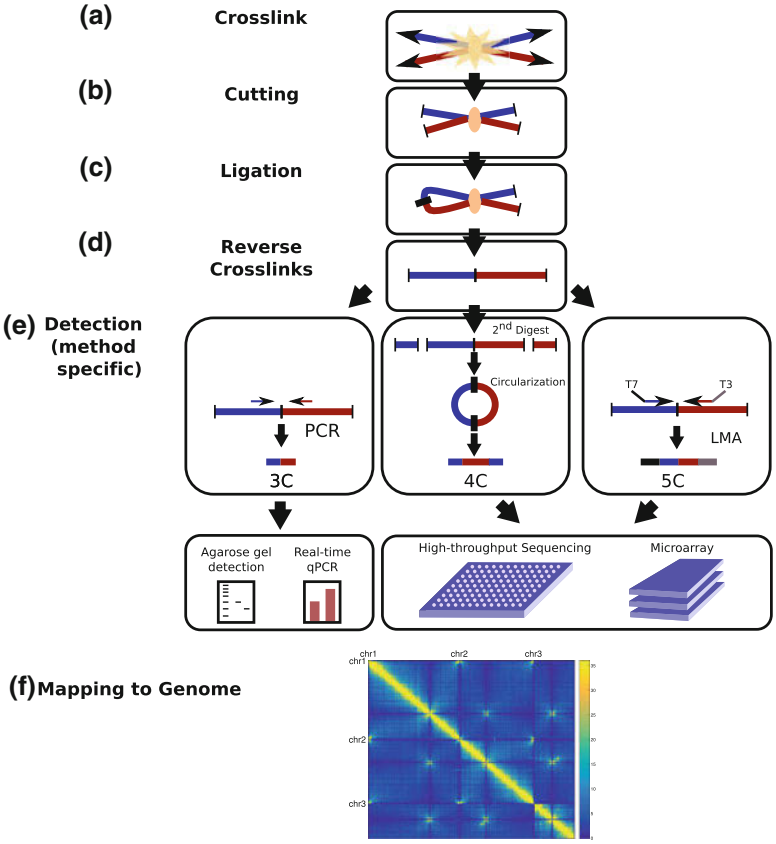


Fig. 14.3 Illustration of the generic protocol for estimating 3D distances among DNA fragments. **a** Formaldehyde is used to cross-link DNA regions that are in proximity in 3D. **b** Restriction enzymes are used to cut the DNA into fragments. **c** Cross-linked fragments ends (restriction sites) are ligated. **d** Cross-links are reversed. **e** Ligation products are detected by various means, depending on the specific protocol (e.g., sequencing), and the two interacting loci from both ends of the restriction site are identified. (Panels A-E from Wikimedia Commons /in public domain.) **f** Reads are mapped to a reference genome and the frequency by which each pair of regions interacts is recorded, here showing the 3 chromosomes of *S. pombe*, based on data from (Mizuguchi et al. 2014)

employs microarrays or quantitative DNA sequencing using 454-technology as detection methods to increase the throughput of 3C.

Whole-genome contact mapping reached a new high when Hi-C was proposed (Lieberman-Aiden et al. 2009), taking advantage of high-throughput paired-end sequencing of the ligation products to increase the coverage in orders of magnitude.

Recently, several additional enhancements to 3C protocols have been proposed to address their limitations. One major limitation is that 3C technologies measure contact frequencies in cell populations. Single-cell Hi-C (Nagano et al. 2013) was

able to overcome this and to identify structures that are maintained/variable across cells and to bridge gaps between imaging techniques and sequencing-based techniques. Another limitation is that usually chromosomes are found in multiple copies in the cell, and to this end, an allele-specific 4C variant has been developed (Splinter et al. 2011). Other modifications were aimed at improving signal-to-noise ratio (Kalhor et al. 2012), increasing whole-genome mapping resolution to a level of single chromatin loops (Rao et al. 2014), monitoring large-scale array of regulatory elements (Mifsud et al. 2015) or single nucleosomes (Hsieh et al. 2015).

In addition to the experimental techniques mentioned, complementary approaches to study genomic architecture have been suggested, such as mapping of genomic regions that are attached to the nuclear lamina (lamina associated domains, or LADs) (Guelen et al. 2008), and ChIA-PET (Fullwood and Ruan 2009) which combines sonication-based chromatin fragmentation, chromatin immunoprecipitation, chromatin proximity ligation, and paired-end tag high-throughput sequencing to detect chromatin interactions.

14.5 Models Based on the Analysis of 3D Genomic Data

One of the hallmarks of eukaryotic nuclear architecture is the tendency of chromosomes to interact much more frequently with *cis* elements on the same chromosome than with *trans* elements on other chromosomes, with little mixing between chromosomes. The term "chromosome territories" was coined in the literature to describe this tendency (Cremer et al. 2006). It was proposed, based on analysis Hi-C contact frequencies and their relation to linear distances on the chromosome, that a fractal globule model for the folding of chromosomes can explain both chromosome territories and the observed global properties of contacts, that decrease as a power law with linear distance (Lieberman-Aiden et al. 2009). This model for polymer (chromosome) folding describes hierarchical local folding that forms a continuous fractal trajectory that densely fills 3D space without crossing itself (no knots). It is clear that additional factors may determine the properties of chromosome folding; for example, a recent study of chromosome dynamics in the nucleus using FISH has shown that chromosome movement is more constrained than previously thought and proposed a model where the observed polymer diffusion is related to cross-linking of chromosomes by lamin A that restricts their movement (Bronshtein et al. 2015).

Hi-C analysis also suggested that regions in the genome are associated with one of two compartments in the nucleus—a transcriptionally active one and an inactive one (Lieberman-Aiden et al. 2009) (Fig. 14.4a). The association with either compartment was based on studying the principle components of the contact frequency matrix. The active compartment was shown to be enriched with active genes, and vice versa. Recently, high-resolution Hi-C analysis suggested that each compartment is further divided into sub-compartments (Rao et al. 2014). Compartments were also shown to be cell type dependent and related to changes in expression.

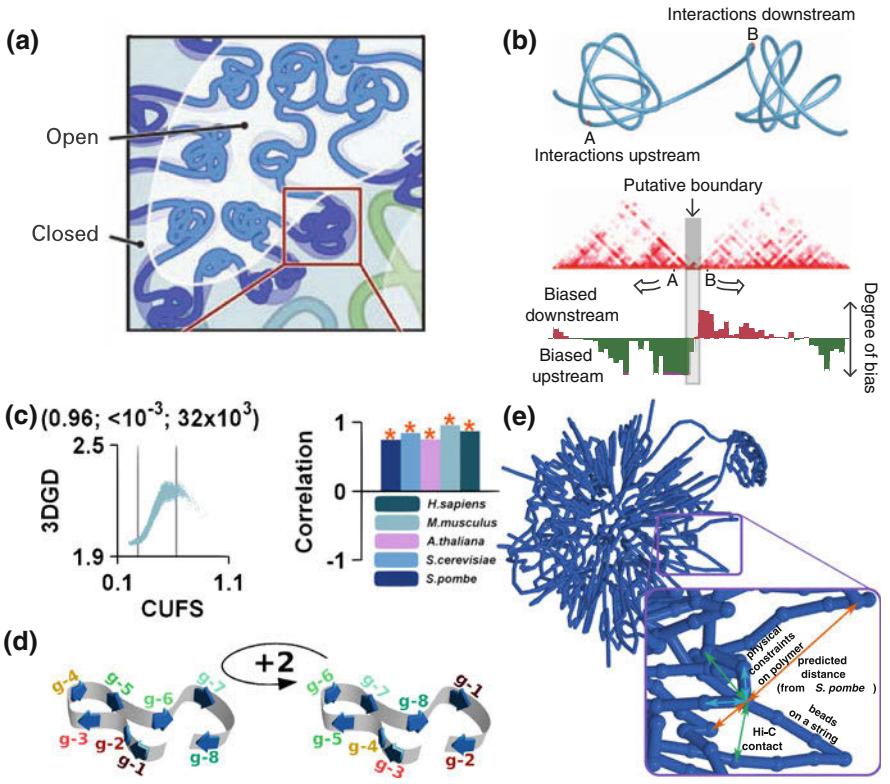


Fig. 14.4 **a** An illustration of chromosome 3D organization into an active compartment of open chromatin and into a silenced compartment (Lieberman-Aiden et al. 2009). **b** An illustration of two TADs and the corresponding enrichment in intradomain interactions within the domains (Dixon et al. 2012). **c** A relation between genes' 3D distance and their functionality, as measured by their codon usage frequency similarity (CUFS). This proposed distance metric serves as a proxy for functional and expression similarity between genes. (Diament et al. 2014). **d** The cyclic chromosome shift test (Diament et al. 2014). **e** An improved 3D reconstruction approach using distances from an evolutionary-related organism (Diament and Tuller 2015)

FISH imaging has suggested the existence of transcription factories, nuclear foci with ongoing active transcription (Misteli 2007; Osborne et al. 2004). A model was proposed, where loops containing active genes colocalize in such factories that are RNA polymerase type dependent. Analysis of Hi-C data in yeast has shown that binding sites of most transcription factors tend to colocalize in 3D more significantly than they are linearly clustered (Ben-Elazar et al. 2013; Diament and Tuller 2015; Kruse et al. 2013), and specific colocalization was reported in the case of tRNA genes (Diament and Tuller 2015; Duan et al. 2010). It was further reported that TFs with spatially colocalized targets are also expressed higher than TFs that are not colocalized, suggesting that regulatory activity is correlated with the presence of transcription factories (Ben-Elazar et al. 2013). Similarly, the hypothesis of

replication factories (Cook 2002) has found supporting evidence in enrichment of Hi-C contact frequencies between early-firing replication origins in yeast (Diament and Tuller 2015; Duan et al. 2010).

Topologically associated domains (TADs) (Dixon et al. 2012) are one of the novel structural features in the organization eukaryotic genomes that were first identified in Hi-C experiments. TADs consist of large, 100 kbp–1 Mbp sized local chromatin interaction domains, where loci are enriched with intradomain interactions and are depleted from interdomain interactions (Fig. 14.4b). They have been identified in higher eukaryotes, from *Drosophila* (Sexton et al. 2012) to mammals. The domains are highly conserved across species (Vietri Rudan et al. 2015) and are stable across different cell types, although intra- and interdomain interactions are known to change during differentiation in accordance with gene regulation (Dixon et al. 2015). They have also been shown to be maintained across individual cells (Nagano et al. 2013). The formation of TADs was proposed to be related to the insulator protein CTCF (Dixon et al. 2012), which in turn was shown to be correlated with the formation of long-distance loops of chromatin in the human genome (Rao et al. 2014). A recently proposed biophysical model suggests that loops form through the extrusion of flexible chromatin fibers by a CTCF- and cohesin-associated complex, and was able to predict the formation of observed contact domains from CTCF binding data (Sanborn et al. 2015). Biophysical models for the effect of chromatin loops on gene expression have also been recently suggested (Doyle et al. 2014). A comparative study of TADs' organization between four mammals revealed that conserved CTCF binding sites are enriched at TAD boundaries and that divergent CTCF binding between species is correlated with divergence of internal domain structure (Vietri Rudan et al. 2015).

The aforementioned models describe the formation of *local* organization, but do not consider the *global* organization of genes and its functional implications. A number of recent studies attempted to address this question. Utilizing a novel proposed metric for functional and expression similarity between genes based on their sequence, an analysis of Hi-C data from 5 eukaryotes has shown a strong relation between functional similarity and genes' 3D distances (Diament et al. 2014) (Fig. 14.4c). Specifically, an empirical null model was utilized in order to show that the observed relation is significantly due to the 3D conformation and not only to linear distances between genes (Fig. 14.4d). Furthermore, genes encoding interacting proteins, genes that form protein complexes, and genes along the same pathway were shown to be colocalized in 3D in human (Thévenin et al. 2014). The colocalization of genes on the same pathway has been recently shown to be cell type specific in multiple human cell lines, and the spatial proximity of related pathways has been reported, indicating a higher order of organization (Karathia et al. 2015). A correlation between distances on a protein–protein interaction network and 3D distances in yeast has been reported (Diament and Tuller 2015). 3D organization was also shown to be related to Gene Ontology functional terminology and to the co-expression of genes in yeast (Homouz and Kudlicki 2013) and in *P. falciparum* (Ay et al. 2014b).

The availability of high-throughput whole-genome measurements has enabled the development of models for the global organization of genomes. For example, a polymer folding simulation in yeast has suggested that many earlier results—including features of the contact maps, the colocalization of early-firing replication origins and genomic location of tRNA genes—can be explained by random configurations of chromosomes that are tethered to a number of sites in the nucleus (Tjong et al. 2012). Such random models offer insights into the possible mechanisms that give rise to the complex genomic architecture. There have been a number of attempts to interpret Hi-C data by generating non-random 3D reconstructions of complete chromosomes and/or genomes based on distance constraints obtained from Hi-C contact maps (Ay et al. 2014b; Ben-Elazar et al. 2013; Duan et al. 2010; Nagano et al. 2013; Tanizawa et al. 2010). Recently, an improved approach for 3D reconstructions was proposed that integrates predicted 3D distances into the reconstruction process in order to guide the optimization into a better solution (Diament and Tuller 2015). Predicted 3D distances were based on measured distances in an evolutionary-related organism or on 3D distance prediction from estimated expression/functional similarity between genes (Fig. 14.4e). Some reconstruction methods are based on stochastic generation of ensemble models, to account for the variability in organization in cell populations (Nagano et al. 2013; Rousseau et al. 2011; Tjong et al. 2012). Additional approaches to 3D reconstruction, both data-driven (based on Hi-C measurements) and de novo biophysical simulations, have been proposed in recent years (Imakaev et al. 2015).

14.6 Conclusions and Future Directions

As discussed in this chapter, various gene expression principles are expected to be the driving force for evolutionary selection for genomic gene organization by function. The studies described here have demonstrated that gene order and their 3D organization in eukaryotic genomes are not random and are correlated with their function.

Nevertheless, there are many open questions and challenges in the field. One open question is related to the causality and directionality of the reported relation between gene functionality and gene 3D location. For example, it is possible that 3D proximity directly affects the expression of close genes, making it more coordinated and matches their similar function. On the other hand, it is possible that the genomic proximity of genes affects their functionality (e.g., due to a more similar mutation rate) such that close genes eventually converge to have a more similar function(s). The resultant correlation can of course be due to both explanations and/or due to additional ones. It may be possible to partially answer this question in the near future with the advance of synthetic biology approaches and accurate large-scale methods for gene expression measurement. For example, an experiment to study this question may include various variants/mutant of the genome of a eukaryotic model organism (e.g., *S. cerevisiae*) where in each mutant only the order

of the genes in the genome is permuted. For each variant Hi-C, gene expression and fitness measurements should be performed.

Some of the aforementioned models and the regulatory role of TADs have been demonstrated in experiments in the mouse X inactivation center (Nora et al. 2012) and in several loci on human chromosome 8 (Sanborn et al. 2015) and were recently shown in relation to disease (Lupiáñez et al. 2015; Grubert et al. 2015). Using CRISPR/Cas genome editing and disrupting the structure of TADs (by altering CTCF binding sites at domain boundaries), the authors were able to show the formation of new enhancer–promoter interactions which consequently led to disease-related changes in the expression of genes (Lupiáñez et al. 2015). While TAD structure is essential, it has been suggested that nuclear positioning of TADs and transcriptional control are not causally related but independently controlled by locally associated *trans*-acting factors (Geeven et al. 2015; Therizols et al. 2014). In a recent study, sub-domains within TADs (SubTADs) were shown to reorganize in response to protein recruitment and form new interdomain interactions while staying intact (Wijchers et al. 2016). In this experiment, transcriptional activity was not always accompanied by changes in positioning of the relevant sub-domain into a different nuclear/expression compartment, and vice versa. Thus, many open questions remain for large-scale synthetic experiments.

An additional interesting research direction is related to cancer evolution. It is well known that cancer cells undergo various local and global genomic aberrations and reorganizations (Bickmore and Teague 2002; Forment et al. 2012; Kim et al. 2013). The cancerous genomic organization affects among others gene organization and clearly affects the 3D organization of the genome. Thus, if indeed there is a causal relation between gene order and gene expression (and functionality), it is possible that cancer evolution is directly acting on 3D gene organization in the cancerous cell. Indeed, recently a few pioneer studies about 3D genomic organization in cancer have been performed, showing a strong relation between chromosome translocations and spatial proximity between regions (Chiarle et al. 2011; Hakim et al. 2012; Oliveira et al. 2012; Zhang et al. 2012), including tissue-specific cancerous translocations (Meaburn et al. 2007). In addition, viral cancer-causing insertions have been shown to cluster in 3D hot spots (Babaei et al. 2015), CTCF/cohesin binding sites have been found to be a major mutational hotspot in the noncoding cancer genome (Katainen et al. 2015), and genomic organization in cancer has been shown to be related to position-specific mutation rates (Schuster-Böckler and Lehner 2012). Two recent reviews have been dedicated to this subject (Corces and Corces 2016; Valton and Dekker 2016). However, additional large-scale study of 3D organization of genes (using techniques such as Hi-C) in various types of cancer cells should provide additional evidence related to the topic.

Finally, there are many experimental, statistical, and computational challenges in the field, needed to be solved in order to be able to develop accurate models of 3D gene organization and to understand the evolution of this feature. The challenges in the field include among others the relatively low resolution in most of the recent Hi-C experiments, specifically in multicellular organisms such as mammals (high resolution is possible (Rao et al. 2014) but currently very expensive), making the modeling

very challenging and noisy (Ay et al. 2014a; Kruse et al. 2013). On the other hand, new higher resolution datasets that are expected to be generated in the near future make the analysis and storage of these data in multiple organisms, tissues, and conditions very challenging as the size of a single high-resolution (e.g., 10-kbp) Hi-C map of the human genome is approximately 14 GB and usually contains about 0.8 billion interactions at variable frequencies (Rao et al. 2014). In addition, Hi-C data, similarly to all NGS-based protocols, introduce various non-trivial biases (Cournac et al. 2012; Imakaev et al. 2012; Nagano et al. 2015; Yaffe and Tanay 2011); the accurate filtering of these biases is far from being trivial as they are related to factors such as the biomolecular protocol used, the 3D conformation and other properties of the genome, the NGS protocol used, and more. In addition, the study of most eukaryotes is complicated by the presence of multiple copies from each chromosome, with different spatial interactions, that are not trivial to demultiplex, measured simultaneously. Finally, most of the previous Hi-C experiments were based on the analysis of a *population* of cells and not single cells; thus, the results Hi-C map is an *average* over many cells. Since we expect that the 3D conformation varies among the different cells, it is not trivial (or impossible) to deduce 3D genomic models of single cells based on such data. Recently, a few single-cell Hi-C experiments have been performed (Nagano et al. 2013); however, the amount of data obtained from each such experiment is low and the number of such experiments reported so far is also very low. Performing and analyzing additional such experiments in the future may enable developing more accurate 3D genomic organization models and promoting molecular evolution studies on the topic.

Acknowledgments AD is grateful to the Azrieli Foundation for the award of an Azrieli Fellowship. This study was supported in part by a fellowship from the Edmond J. Safra Center for Bioinformatics at Tel Aviv University.

References

Ay F, Bailey TL, Noble WS (2014a) Statistical confidence estimation for Hi-C data reveals regulatory chromatin contacts. Genome Res 2014a Feb 5; gr. 160374.113

Ay F, Bunnik EM, Varoquaux N, Bol SM, Prudhomme J, Vert J-P et al (2014b) Three-dimensional modeling of the *P. falciparum* genome during the erythrocytic cycle reveals a strong connection between genome architecture and gene expression. Genome Res 24 (6):974–988

Babaei S, Akhtar W, de Jong J, Reinders M, de Ridder J (2015) 3D hotspots of recurrent retroviral insertions reveal long-range interactions with cancer genes. Nat Commun 27(6):6381

Bártová E, Kozubek S (2006) Nuclear architecture in the light of gene expression and cell differentiation studies. Biol Cell 98(6):323–336

Batada NN, Hurst LD (2007) Evolution of chromosome organization driven by selection for reduced gene expression noise. Nat Genet 39(8):945–949

Belmont AS (2001) Visualizing chromosome dynamics with GFP. Trends Cell Biol 11(6): 250–257

Ben-Elazar S, Yakhini Z, Yanai I (2013) Spatial localization of co-regulated genes exceeds genomic gene clustering in the *Saccharomyces cerevisiae* genome. Nucleic Acids Res 41 (4):2191–2201

Bickmore WA, Teague P (2002) Influences of chromosome size, gene density and nuclear position on the frequency of constitutional translocations in the human population. Chromosome Res 10 (8):707–715

Blumenthal T, Evans D, Link CD, Guffanti A, Lawson D, Thierry-Mieg J et al (2002) A global analysis of *Caenorhabditis elegans* operons. Nature 417(6891):851–854

Bronshtein I, Kepten E, Kanter I, Berezin S, Lindner M, Redwood AB et al (2015) Loss of lamin a function increases chromatin dynamics in the nuclear interior. Nat Commun 24(6):8044

Caron H, van Schaik B, van der Mee M, Baas F, Riggins G, van Sluis P et al (2001) The human transcriptome map: clustering of highly expressed genes in chromosomal domains. Science 291(5507):1289–1292

Chen Y, Dokholyan NV (2006) The coordinated evolution of yeast proteins is constrained by functional modularity. Trends Genet 22(8):416–419

Chiarle R, Zhang Y, Frock RL, Lewis SM, Molinie B, Ho Y-J et al (2011) Genome-wide translocation sequencing reveals mechanisms of chromosome breaks and rearrangements in B cells. Cell 147(1):107–119

Cohen BA, Mitra RD, Hughes JD, Church GM (2000) A computational analysis of whole-genome expression data reveals chromosomal domains of gene expression. Nat Genet 26(2):183–186

Cook PR (2002) Predicting three-dimensional genome structure from transcriptional activity. Nat Genet 32(3):347–352

Corces MR, Corces VG (2016) The three-dimensional cancer genome. Curr Opin Genet Dev 36: 1–7

Costanzo M, Baryshnikova A, Bellay J, Kim Y, Spear ED, Sevier CS et al (2010) The genetic landscape of a cell. Science 327(5964):425–431

Cournac A, Marie-Nelly H, Marbouty M, Koszul R, Mozziconacci J (2012) Normalization of a chromosomal contact map. BMC Genom 13(1):436

Cremer T, Cremer M, Dietzel S, Müller S, Solovei I, Fakan S (2006) Chromosome territories—a functional nuclear landscape. Curr Opin Cell Biol 18(3):307–316

Dekker J, Rippe K, Dekker M, Kleckner N (2002) Capturing chromosome conformation. Science 295(5558):1306–1311

Diament A, Tuller T (2015) Improving 3D genome reconstructions using orthologous and functional constraints. PLoS Comput Biol 11(5):e1004298

Diament A, Pinter RY, Tuller T (2014) Three-dimensional eukaryotic genomic organization is strongly correlated with codon usage expression and function. Nat Commun 5:5876

Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y et al (2012) Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature 485(7398):376–380

Dixon JR, Jung I, Selvaraj S, Shen Y, Antosiewicz-Bourget JE, Lee AY et al (2015) Chromatin architecture reorganization during stem cell differentiation. Nature 518(7539):331–336

Dostie J, Richmond TA, Arnaout RA, Selzer RR, Lee WL, Honan TA et al (2006) Chromosome conformation capture carbon copy (5C): a massively parallel solution for mapping interactions between genomic elements. Genome Res 16(10):1299–1309

Doyle B, Fudenberg G, Imakaev M, Mirny LA (2014) Chromatin loops as allosteric modulators of enhancer-promoter interactions. PLoS Comput Biol 10(10):e1003867

Duan Z, Andronescu M, Schutz K, McIlwain S, Kim YJ, Lee C et al (2010) A three-dimensional model of the yeast genome. Nature 465(7296):363–367

Field B, Osbourn AE (2008) Metabolic diversification—independent assembly of operon-like gene clusters in different plants. Science 320(5875):543–547

Forment JV, Kaidi A, Jackson SP (2012) Chromothripsis and cancer: causes and consequences of chromosome shattering. Nat Rev Cancer 12(10):663–670

Fullwood MJ, Ruan Y (2009) ChIP-based methods for the identification of long-range chromatin interactions. J Cell Biochem 107(1):30–39

Geeven G, Zhu Y, Kim BJ, Bartholdy BA, Yang S-M, Macfarlan TS et al (2015) Local compartment changes and regulatory landscape alterations in histone H1-depleted cells. Genome Biol 16:289

Girvan M, Newman MEJ (2002) Community structure in social and biological networks. PNAS 99 (12):7821–7826

Grubert F, Zaugg JB, Kasowski M, Ursu O, Spacek DV, Martin AR et al (2015) Genetic control of chromatin states in humans involves local and distal chromosomal interactions. Cell 162 (5):1051–1065

Guelen L, Pagie L, Brasset E, Meuleman W, Faza MB, Talhout W et al (2008) Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. Nature 453(7197):948–951

Hakim O, Resch W, Yamane A, Klein I, Kieffer-Kwon K-R, Jankovic M et al (2012) DNA damage defines sites of recurrent chromosomal translocations in B lymphocytes. Nature 484 (7392):69–74

Homouz D, Kudlicki AS (2013) The 3D organization of the yeast genome correlates with co-expression and reflects functional relations between genes. PLoS ONE 8(1):e54699

Hsieh T-HS, Weiner A, Lajoie B, Dekker J, Friedman N, Rando OJ (2015) Mapping nucleosome resolution chromosome folding in yeast by micro-C. Cell 162(1):108–119

Hurst LD, Pál C, Lercher MJ (2004) The evolutionary dynamics of eukaryotic gene order. Nat Rev Genet 5(4):299–310

Huvet M, Nicolay S, Touchon M, Audit B, d'Aubenton-Carafa Y, Arneodo A et al (2007) Human gene organization driven by the coordination of replication and transcription. Genome Res 17 (9):000–000

Imakaev M, Fudenberg G, McCord RP, Naumova N, Goloborodko A, Lajoie BR et al (2012) Iterative correction of Hi-C data reveals hallmarks of chromosome organization. Nat Methods 9(10):999–1003

Imakaev MV, Fudenberg G, Mirny LA (2015) Modeling chromosomes: Beyond pretty pictures. FEBS Lett 589(20, Part A):3031–3036

Kalhor R, Tjong H, Jayathilaka N, Alber F, Chen L (2012) Genome architectures revealed by tethered chromosome conformation capture and population-based modeling. Nat Biotech 30 (1):90–98

Karathia H, Kingsford C, Girvan M, Hannenhalli S (2015) A pathway-centric view of spatial proximity in the 3D nucleome across cell lines. BioRxiv 027045

Katainen R, Dave K, Pitkänen E, Palin K, Kivioja T, Välimäki N et al (2015) CTCF/cohesin-binding sites are frequently mutated in cancer. Nat Genet 47(7):818–821

Kim T-M, Xi R, Luquette LJ, Park RW, Johnson MD, Park PJ (2013) Functional genomic analysis of chromosomal aberrations in a compendium of 8000 cancer genomes. Genome Res 23 (2):217–227

Kim SH, McQueen PG, Lichtman MK, Shevach EM, Parada LA, Misteli T (2004) Spatial genome organization during T-cell differentiation. Cytogenet Genome Res 105(2–4):292–301

Kosak ST, Groudine M (2004) Gene order and dynamic domains. Science 306(5696):644–647

Kosak ST, Skok JA, Medina KL, Riblet R, Beau MML, Fisher AG et al (2002) Subnuclear compartmentalization of immunoglobulin loci during lymphocyte development. Science 296 (5565):158–162

Kruse K, Sewitz S, Babu MM (2013) A complex network framework for unbiased statistical analyses of DNA–DNA contact maps. Nucl Acids Res 41(2):701–710

Langer-Safer PR, Levine M, Ward DC (1982) Immunological method for mapping genes on Drosophila polytene chromosomes. Proc Natl Acad Sci USA 79(14):4381–4385

Lee JM, Sonnhammer ELL (2003) Genomic gene clustering analysis of pathways in eukaryotes. Genome Res 13(5):875–882

Lercher MJ, Urrutia AO, Hurst LD (2002) Clustering of housekeeping genes provides a unified model of gene order in the human genome. Nat Genet 31(2):180–183

Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A et al (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science 326(5950):289–293

Lupiáñez DG, Kraft K, Heinrich V, Krawitz P, Brancati F, Klopocki E et al (2015) Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. Cell 161(5):1012–1025

Marcotte EM, Pellegrini M, Ng H-L, Rice DW, Yeates TO, Eisenberg D (1999) Detecting protein function and protein-protein interactions from genome sequences. Science 285(5428):751–753

Martin J, Lercher AOU (2003) A unification of mosaic structures in the human genome. Hum Mol Genet 12(19):2411–2415

Meaburn KJ, Misteli T, Soutoglou E (2007) Spatial genome organization in the formation of chromosomal translocations. Semin Cancer Biol 17(1):80–90

Mifsud B, Tavares-Cadete F, Young AN, Sugar R, Schoenfelder S, Ferreira L et al (2015) Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. Nat Genet 47(6):598–606

Miller MA, Cutter AD, Yamamoto I, Ward S, Greenstein D (2004) Clustered organization of reproductive genes in the *C. elegans* genome. Curr Biol 14(14):1284–1290

Misteli T (2007) Beyond the sequence: cellular organization of genome function. Cell 128(4):787–800

Mitra K, Carvunis A-R, Ramesh SK, Ideker T (2013) Integrative approaches for finding modular structure in biological networks. Nat Rev Genet 14(10):719–732

Mizuguchi T, Fudenberg G, Mehta S, Belton J-M, Taneja N, Folco HD et al (2014) Cohesin-dependent globules and heterochromatin shape 3D genome architecture in *S. pombe*. Nature 516(7531):432–435

Nagano T, Lubling Y, Stevens TJ, Schoenfelder S, Yaffe E, Dean W et al (2013) Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. Nature 502(7469):59–64

Nagano T, Várnai C, Schoenfelder S, Javierre B-M, Wingett SW, Fraser P (2015) Comparison of Hi-C results using in-solution versus in-nucleus ligation. Genome Biol 16:175

Nora EP, Lajoie BR, Schulz EG, Giorgetti L, Okamoto I, Servant N et al (2012) Spatial partitioning of the regulatory landscape of the X-inactivation centre. Nature 485(7398): 381–385

Oliveira TY, Resch W, Jankovic M, Casellas R, Nussenzweig MC, Klein IA (2012) Translocation capture sequencing: a method for high throughput mapping of chromosomal rearrangements. J Immunol Methods 375(1–2):176–181

Osborne CS, Chakalova L, Brown KE, Carter D, Horton A, Debrand E et al (2004) Active genes dynamically colocalize to shared sites of ongoing transcription. Nat Genet 36(10):1065–1071

Osbourn AE, Field B (2009) Operons. Cell Mol Life Sci 66(23):3755–3775

Pál C, Hurst LD (2003) Evidence for co-evolution of gene order and recombination rate. Nat Genet 33(3):392–395

Petkov PM, Graber JH, Churchill GA, DiPetrillo K, King BL, Paigen K (2007) Evidence of a large-scale functional organization of mammalian chromosomes. PLoS Biol 5(5):e127

Poyatos JF, Hurst LD (2007) The determinants of gene order conservation in yeasts. Genome Biol 8(11):R233

Pritykin Y, Singh M (2013) Simple topological features reflect dynamics and modularity in protein interaction networks. PLoS Comput Biol 9(10):e1003243

Rao SSP, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT et al (2014) A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell 159(7):1665–1680

Ravasz E, Somera AL, Mongru DA, Oltvai ZN, Barabási A-L (2002) Hierarchical organization of modularity in metabolic networks. Science 297(5586):1551–1555

Rousseau M, Fraser J, Ferraiuolo MA, Dostie J, Blanchette M (2011) Three-dimensional modeling of chromatin structure from interaction frequency data using Markov chain Monte Carlo sampling. BMC Bioinformatics 25(12):414

Roy PJ, Stuart JM, Lund J, Kim SK (2002) Chromosomal clustering of muscle-expressed genes in *Caenorhabditis elegans*. Nature 418(6901):975–979

Salgado H, Moreno-Hagelsieb G, Smith TF, Collado-Vides J (2000) Operons in *Escherichia coli*: genomic analyses and predictions. Proc Natl Acad Sci US A 97(12):6652–6657

Sanborn AL, Rao SSP, Huang S-C, Durand NC, Huntley MH, Jewett AI et al (2015) Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. PNAS 112(47):E6456–E6465

Schmidt T, Schwarzacher T, Heslop-Harrison JS (1994) Physical mapping of rRNA genes by fluorescent in-situ hybridization and structural analysis of 5S rRNA genes and intergenic spacer sequences in sugar beet (Beta vulgaris). Theoret Appl Genet 88(6–7):629–636

Schuster-Böckler B, Lehner B (2012) Chromatin organization is a major influence on regional mutation rates in human cancer cells. Nature 488(7412):504–507

Sémon M, Duret L (2006) Evolutionary origin and maintenance of coexpressed gene clusters in mammals. Mol Biol Evol 23(9):1715–1723

Sexton T, Yaffe E, Kenigsberg E, Bantignies F, Leblanc B, Hoichman M et al (2012) Three-dimensional folding and functional organization principles of the Drosophila genome. Cell 148(3):458–472

Simonis M, Klous P, Splinter E, Moshkin Y, Willemsen R, de Wit E et al (2006) Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). Nat Genet 38(11):1348–1354

Singer GAC, Lloyd AT, Huminiecki LB, Wolfe KH (2005) Clusters of co-expressed genes in mammalian genomes are conserved by natural selection. Mol Biol Evol 22(3):767–775

Slot JC, Rokas A (2010) Multiple GAL pathway gene clusters evolved independently and by different mechanisms in fungi. PNAS 107(22):10136–10141

Speicher MR, Ballard SG, Ward DC (1996) Karyotyping human chromosomes by combinatorial multi-fluor FISH. Nat Genet 12(4):368–375

Splinter E, de Wit E, Nora EP, Klous P, van de Werken HJG, Zhu Y et al (2011) The inactive X chromosome adopts a unique three-dimensional conformation that is dependent on Xist RNA. Genes Dev 25(13):1371–1383

Sproul D, Gilbert N, Bickmore WA (2005) The role of chromatin structure in regulating the expression of clustered genes. Nat Rev Genet 6(10):775–781

Stuart JM, Segal E, Koller D, Kim SK (2003) A gene-coexpression network for global discovery of conserved genetic modules. Science 302(5643):249–255

Tanizawa H, Iwasaki O, Tanaka A, Capizzi JR, Wickramasinghe P, Lee M et al (2010) Mapping of long-range associations throughout the fission yeast genome reveals global genome organization linked to transcriptional regulation. Nucl Acids Res 38(22):8164–8177

Teichmann SA, Veitia RA (2004) Genes encoding subunits of stable complexes are clustered on the yeast chromosomes. Genetics 167(4):2121–2125

Therizols P, Illingworth RS, Courilleau C, Boyle S, Wood AJ, Bickmore WA (2014) Chromatin decondensation is sufficient to alter nuclear organization in embryonic stem cells. Science 346 (6214):1238–1242

Thévenin A, Ein-Dor L, Ozery-Flato M, Shamir R (2014) Functional gene groups are concentrated within chromosomes, among chromosomes and in the nuclear space of the human genome. Nucl Acids Res 42(15):9854–9861

Tjong H, Gong K, Chen L, Alber F (2012) Physical tethering and volume exclusion determine higher-order genome organization in budding yeast. Genome Res 22(7):1295–1305

Tuller T, Rubinstein U, Bar D, Gurevitch M, Ruppin E, Kupiec M (2009) Higher-order genomic organization of cellular functions in yeast. J Comput Biol 16(2):303–316

van der Ploeg M (2000) Cytochemical nucleic acid research during the twentieth century. Euro J Histochem EJH 44(1):7–42

Valton A-L, Dekker J (2016) TAD disruption as oncogenic driver. Curr Opin Genet Dev 36:34–40

Vietri Rudan M, Barrington C, Henderson S, Ernst C, Odom DT, Tanay A et al (2015) Comparative Hi-C reveals that CTCF underlies evolution of chromosomal domain architecture. Cell Rep 10(8):1297–1309

Weber CC, Hurst LD (2011) Support for multiple classes of local expression clusters in *Drosophila melanogaster*, but no evidence for gene order conservation. Genome Biol 12 (3):1–15

Wijchers PJ, Krijger PHL, Geeven G, Zhu Y, Denker A, Verstegen MJAM et al (2016) Cause and consequence of tethering a SubTAD to different nuclear compartments. Mol cell 61(3): 461–473

Wong S, Wolfe KH (2005) Birth of a metabolic gene cluster in yeast by adaptive gene relocation. Nat Genet 37(7):777–782

Yaffe E, Tanay A (2011) Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. Nat Genet 43(11):1059–1065

Zhang Y, McCord RP, Ho Y-J, Lajoie BR, Hildebrand DG, Simon AC et al (2012) Spatial organization of the mouse genome and its role in recurrent chromosomal translocations. Cell 148(5):908–921

Part III
Concepts

Chapter 15
How Likely Are We? Evolution of Organismal Complexity

William Bains

Abstract All complex multicellular organisms are eukaryotes. How did the evolution of the highly complex architecture of the eukaryotic cell arise? I discuss the differences between bacteria and archaea (prokaryotes) and eukaryotes in terms of chemistry, cellular structure, energetic and genetics. Chemistry and cell structure are less diagnostic of eukaryote than they appear at first. I focus on two pivotal differences between eukaryotes and other forms of life: energetics and genetic control. Eukaryotes can generate substantially more energy per gene than prokaryotes, and this has been suggested as the key enabler of complex genetics. I suggest that a more basic difference is the genetic logic of eukaryotes (not the genetic chemistry, which is shared with all domains of life). Eukaryotic genes are by default 'off', prokaryotic ones by default 'on'. This difference makes growth in genome complexity easier, and growth in control complexity itself then drives a requirement for mitochondria and increased energy production. I conclude that, given 'default off' genetics, complex life is highly likely to evolve. The paths to the evolution of 'default off' genetics remain to be explored.

15.1 Introduction

Life on earth is divided into two great domains. One domain is inhabited by the animals, plants and fungi that we can see in the world around us. The other are the bacteria and archaea (which, despite the term being old-fashioned (Woese et al. 1990), I will call 'prokaryotes'). The prokaryotes dominate the earth's chemistry, inhabit more diverse habitats than eukaryotes [e.g. Cowan 2004; Williams and Hallsworth 2009; Javor 2012; Bains et al. 2015] and match the eukaryote's planetary mass (Whitman et al. 1998) and integrated planetary genome complexity (Landenmark et al. 2015). Only the largest are (just) visible to the naked eye (Angert 2012), and they have only developed the most rudimentary form of multicellular complexity.

W. Bains (✉)
EAPS MIT, 77 Mass. Ave, Cambridge, MA 02139, USA
e-mail: bains@mit.edu

P. Pontarotti (ed.), *Evolutionary Biology*, DOI 10.1007/978-3-319-41324-2_15

This chapter explores why this great divide has occurred, focusing on genome complexity as a key driver of organismal complexity. I have two motives to ask the question. Evolutionary conundrums are a challenge to any life scientist, and we hope that solving them will shed light on how modern living organisms function, with both theoretical and practical consequences. But we would also like to know how likely our own appearance is. If life is found on another world, what are the chances that it could evolve complex life? If some global catastrophe wiped out all but a few refuges of single-celled organisms on earth, would complex life emerge again to reforest the land, fill the sea with fish and write chapters on evolutionary biology?

The tentative conclusion of this chapter is that complex life is highly likely, primarily for reasons rooted in genetics.

15.2 What Is Unique About Multicellular Organisms

In general, large size and greater complexity go together. While this is a Pirate Rule (more what you would call guidelines than actual rules[1]), it is generally true across many phyla and size scales, as greater specialization enables larger size and larger size enables greater specialization (Bonner 2004). Greater complexity requires a more complex programme to construct the organism. This requires a complex genetic programme, cell adhesion and communication systems, as well as methods for bulk transport of nutrients between cells (Knoll 2011; Rokas 2008). Movement of material within cells is also important if the cells become too big for diffusion to allow metabolites and proteins to reach their targets. Eukaryotic cells have evolved multicellularity many times independently, prokaryotic cells never have (Knoll and Hewitt 2011; de Sousa António and Schulze-Makuch 2012; Knoll 2011; Rokas 2008). The origin of complex, multicellular organisms is therefore the origin of complex cells, i.e. of the eukaryotic cell, with its complex chemistry, structure, energetics and genetics.

The problem of identifying the key breakthrough which enabled the evolution of complex life therefore reduces to the problem of the key difference between eukaryotic and prokaryotic cells, and how that occurred.

15.3 What Is Unique About Eukaryote Cells

There are a wide variety of theories for the origin of eukaryotes (Martin et al. 2015), but it is now broadly accepted that they originated in the TACK group of archaea, and an early event was the acquisition of a proteobacterial symbiont that became the

[1] 'Pirates of the Caribbean: The legend of the Black Pearl' Disney Films.

mitochondrion (Spang et al. 2015) (also see Mariscal and Doolittle 2015) for a review of competing hypotheses). The origin of the eukaryotic genome is more complex, with substantial archaeal and eubacterial contributions (Pittis and Gabaldón 2016). Thus, the origin of the eukaryotic cell is not a simple linear series of steps from a prokaryotic precursor, but involves at least one, and maybe several, discontinuities.

There are many hypotheses about what enabled eukaryotic cells to develop their present complexity, which fall into the general categories of chemistry, cellular substructure, bioenergetics and genetics. None claim complete exclusivity; rather, they emphasize different aspects of the difference between prokaryotic and eukaryotic life as unique.

15.3.1 Chemistry

There are a number of aspects of eukaryotic chemistry that are unique. I will only summarize these briefly, because, while they are distinctive, I do not believe that their acquisition was the key step to allowing complex organisms.

In general, prokaryotes are at least as chemically diverse and adaptable than eukaryotes (Gunatilaka 2012; Sanchez et al. 2012; Fusetani 2012). Here genetic chemistry is discussed. The macromolecular chemistry of prokaryotes and eukaryotes (DNA and RNA and their modification, the 20 proteinaceous amino acids, etc.) is remarkably similar.

The cell surface carbohydrates of complex eukaryotes are central to cell: cell signalling and adhesion. Glycoproteins are actually more varied in prokaryotes than eukaryotes (Moens and Vanderleyden), but their deployment is different. For example, terminal sialic acids are common on eukaryotic, especially vertebrate, cell surface glycans and are the key for cell–cell surface signalling (Varki and Schauer 2009). However, sialic acids are also found in eubacterial and archaeal glycans, sometimes in terminal positions (Angata and Varki 2002). The difference is that bacterial cell surface glycans have functions in cell defence and intercellular conflict, not in cooperation (Moens and Vanderleyden 1997). Similarly, protein kinases are a prominent mechanism for cell–cell signalling and intracellular control in eukaryotes (McGuire 2015). However, signal transduction and intracellular signalling by membrane protein kinases are common in bacteria, although such signalling systems tend to interact directly with the transcription apparatus at the cell membrane, not with intermediate signalling systems (Karin and Hunter 1995). Some prokaryotes also contain analogues (whether they are homologues is still disputed) of cytoskeletal components, such as actin and tubulin (Williams et al. 2014; Roeben et al. 2006; Yutin and Koonin 2012; Erickson 1997): however, they are rarely used to construct a large intracellular network of fibres as in the eukaryotic cell. In these and other cases, the key difference between eukaryotes and other domains of life is the architecture of the chemistry, not its nature.

I therefore turn my attention to the architecture of the cell's chemistry.

15.3.2 Cell Structure

The eukaryotic cell is widely believed to have a much more complex structure than prokaryotic cells. The nucleus itself has a complex internal structure and is bounded by a complex membrane. The cytoskeleton, centrioles, mitochondria, plastids, Golgi bodies, complex system of vesicles and their trafficking proteins all are unique to eukaryotes. This allows very complex shape and, as noted above, the sophisticated selective modification and export of material.

A key difference between prokaryotes and eukaryotes is the presence of a nucleus with a complex, double-membrane boundary with sophisticated nuclear pores. It has been argued that the nucleus is essential to partition many linear chromosomes in eukaryotes into nucleus and cytoplasm (Lodé 2012). However, this ignores that some planctomycetes have a membrane-bound nuclear compartment without linear chromosomes (Fuerst 2005; Fuerst et al. 1998), and diverse prokaryotes have linear chromosomes without the need for a nucleus (Ferdows and Barbour 1989; Hinnebusch and Tilly 1993; Sherman et al. 2010; Kube et al. 2008). Other prokaryotes have a very large number of duplicates of circular chromosomes (polyploidy) (Soppa 2014; Zerulla and Soppa 2014; Griese et al. 2011; Mendell et al. 2008; Komaki and Ishikawa 2000) without the need for cytoplasmic–nuclear separation.

What the eukaryotic nucleus does is decouple transcription and translation (Lodé 2012). In prokaryotes, 'transertion' couples transcription, translation and insertion into membrane, which means actively transcribed DNA is near the cell membrane (Norris et al. 1996). In eukaryotes, this is decoupled, which might allow greater efficiency of translation (Lu et al. 2007). The endoplasmic reticulum and Golgi apparatus are central parts of this uncoupling.

The cytoskeleton is now understood to not be unique to eukaryotes either, as homologous proteins are shared with archaea (Williams et al. 2014; Roeben et al. 2006; Yutin and Koonin 2012; Erickson 1997). Some bacteria show intracellular compartmentalization of specific metabolic processes (Fuerst 2005), intracellular membrane compartments for secretion and processing (Stolz 2001), complex intracellular membrane structures such as the thylakoid stacks of cyanobacteria (Stolz 2001), and the capability of forming membrane vesicles for protein secretion (Kulp and Kuehn 2010). While they do not have anything as large and complex as the Golgi apparatus, they have the basics of making specialist intracellular compartments for specialist chemistry, moving these around the cell and secreting them as needed. Interestingly, the large bacterium *Epulopiscium fishelsoni* has developed internal structures analogous to the cytoskeleton. *E. fishelsoni* was originally classified as a protest because of its complex internal structures of vesicles and tubules analogous in some ways to the eukaryotic cell's cytoskeleton system (Montgomery and Pollak 1988) and only definitively identified as a huge prokaryote through rRNA gene sequencing (Angert et al. 1993). Thus, the difference between prokaryotic and eukaryotic internal organization may be as much a reflection of the need for overcoming diffusion limits in a cell as a basic limit on

biochemistry. [*Thiomargarita namibiensis*, which is slightly larger, does not contain these mechanisms—its cytoplasm is a thin layer between a large central vacuole and the cell wall (Schulz et al. 1999).]

However, prokaryotes do not have direct analogues of mitochondria or chloroplasts. The endosymbiotic origin of mitochondria and chloroplasts is now widely accepted (Martin et al. 2015; Zimorski et al. 2014). However, it is not clear that this origin was, of itself, the unique event that separated eukaryotes from prokaryotes, because all the steps leading from free-living commensal to embedded endosymbiont have occurred multiple times. Both endosymbiosis and prokaryote-to-eukaryote gene transfer are modern as well as ancient phenomena. Modern biology shows examples of endosymbiotic or endoparasitic bacteria that live inside eukaryotes (Dolan 2001), other bacteria (Thao et al. 2002; Wujek 1979) and bacteria that live inside modern mitochondria (Sassera et al. 2006). Modern endosymbiosis of photosynthetic bacteria is also known (Reyes-Prieto et al. 2010; Marin et al. 2005), as well as endosymbiotic capture of photosynthetic eukaryotes by other eukaryotes (Tirichine and Bowler 2011; Nagamune et al. 2008). Endosymbiosis is therefore possible, even common, and is likely to have occurred many times in the history of life. Transfer of genes from prokaryotic endoparasites and endosymbionts to eukaryotic hosts is also found to have occurred multiple times (Moran et al. 2008; McCutcheon and Moran 2012).

Thus, some other aspects of chloroplast and (more especially) mitochondrial origin may be the distinctive feature that distinguishes eukaryotic from prokaryotic cells rather than just the capture of one cell by another. This is the basis of the energetic hypothesis of eukaryotic origins, whose principle recent champions are Lane and Martin (2010).

15.3.3 Energetics

A leading hypothesis concerning the fundamental difference between prokaryotes and eukaryotes is that eukaryotic cells can capture energy in internal organelles—mitochondria and chloroplasts—whereas prokaryotes are limited to energy-capture mechanisms at the cell surface. Although the metabolic rate per unit mass is similar across all domains of life (Makarieva et al. 2008), the rate of energy production per gene is much higher in eukaryotes (Lane and Martin 2010, 2015). Lane and Martin (2010) argued that it is the acquisition of mitochondria, which enabled this greatly enhanced energy production per gene, which is the key to enable a complex genome.

Lane and Martin's hypothesis can be summarized as follows. 75 % of the energy used by a cell is spent on protein synthesis. A larger genome means a larger cell, both logically (as that DNA has to go somewhere) and by observation (Elliott and Gregory 2015). As the major constituent of the cell is protein, this means that doubling the size of the genome means doubling the amount of protein made, and hence doubling the energy required per cell (and per genome). However, doubling

the volume of the cell does not double its surface area, as volume is proportional to the cube of the radius, surface area to the square, given the same shape. If energy production is based on electron transport in the cell membrane, so that it only happens at the surface of the cell, then the genome cannot be expanded indefinitely before the cell runs out of energy.

This problem was solved by the evolution of internal organelles with a very large membrane area but very few genes, by forming an endosymbiotic relationship with a free-living eubacterium and subsequent transfer of most of the genome of the symbiote to the nucleus. This allows the eukaryote to greatly increase its energy production per gene and so greatly increase the number of genes per cell. The mitochondrion had to retain some genes to allow for the rapid, flexible control of oxidative phosphorylation, an inherently reactive and potentially damaging chemistry. The argument for retention of mitochondrial genes for control is called CoRR—Co-location for Redox Regulation (van der Giezen and Tovar 2005). Once started, gene transfer goes until only a minimum genome needed to control electron transfer is left in the mitochondrion, because the intermediate state, where a few genes have been transferred, is not stable to metabolic and genetic conflict between endosymbiont and host (Lane 2011).

Recent genomic analysis has suggested that the 'original' eukaryotic cell predated the acquisition of mitochondria. The eukaryotic nuclear genome has affinities to eubacterial and Lokiarchaeal genes that predate the earliest transferred mitochondrial genes (Pittis and Gabaldón 2016). However, this only states that the eukaryotic 'chassis' was in place before mitochondria were acquired, and does not invalidate the idea that the proto-eukaryotic genome could only expand after the mitochondrion was acquired.

It is clear that eukaryotes can flourish without mitochondria. Hydrogenosomes in a range of anaerobic eukaryotic parasites originated from mitochondria. They have no DNA and produce energy from substrate-level phosphorylation only [reviewed in Van Der Giezen 2009; Müller et al. 2012]. However, electron transport remains central to their generation of hydrogen to balance the redox of the cell. Other anaerobic parasitic eukaryotes lack hydrogenosomes and generate ATP in the cytosol (Lloyd et al. 2002). Lane and Martin's hypothesis does not require that the ancestral eukaryote be aerobic—indeed, it was probably anaerobic, and the mitochondrion was providing hydrogenic metabolism allowing much more efficient fermentation than substrate-level phosphorylation (Martin et al. 2015; Martin 2011). However, Cavalier-Smith (2013) has pointed out that there is no correlation of high-energy aerobic metabolism and cell size, with small aerobic cells and large anaerobic ones, and the real correlate of genome size is cell size (Elliott and Gregory 2015).

Other criticisms may be more fundamental. Booth and Doolittle state that these are far less of a 'major gap' between prokaryotes and eukaryotes than traditional microscopy suggests (as I have outlined above), and that energy per gene is not the primary difference as there is substantial overlap between cell sizes and gene numbers between prokaryotes and eukaryotes (Booth and Doolittle 2015). Larger prokaryotic genomes are larger than smaller eukaryotic ones, and such large

genomes are phylogenetically diverse and quite common, not an unusual specialization (Guieysse and Wuertz 2012; Bentkowski et al. 2015). Recently, Lynch et al. have commented that 'energy per gene' is not ecologically meaningful and in any case is based on the energy-per-base not energy-per-gene (Lynch and Marinov 2015, 2016) [but see Lane and Martin 2016].

So, Lane and Martin are quite correct that eukaryotes have evolved the ability to generate far more power per cell, and per base of DNA, than prokaryotes. However, it is disputed as to why this is so and so we might ask what other energy-requiring processes distinguish eukaryotes from prokaryotes. Given their greatly expanded genome, we look to the systems of gene control.

We know that synthesizing an mRNA in a eukaryote can take an average of up to 10 times as much ATP as in a prokaryote, depending on the species, because of the ubiquity of untranslated introns (Venter et al. 2001; Lander et al. 2001). Eukaryotes also transcribe an abundance of non-coding RNAs (Washietl et al. 2007; Kapranov et al. 2007; Nagano and Fraser 2011; Ørom et al. 2010; Geisler and Coller 2013). So maybe genetic mechanisms are more important than cell bulk and protein per se. Szathmáry (2015) also comments that Lane and Martin's focus on energy misses that life's major transitions are about information transfer (Maynard Smith and Szathmary 1995), not energy. If eukaryotes have more energy at their disposal, maybe this is in service of genetic control, not protein synthesis.

15.3.4 Eukaryotic Versus Prokaryotic Gene Control

Complex genetic controls are central to the complex developmental programmes that build multicellular organisms. Most tissues in multicellular organisms express only a small subset of the proteins coded in their genome, but the pattern of proteins expressed changes in a very complex pattern as the organism develops from a single cell to an adult, controlled by a huge range of influences within and outside the cell. Only when organisms developed, the ability to control gene activity in this sophisticated way was the evolution of multicellularity possible, and this ability is enabled not by the number of proteins but by the sophistication of the genetic programme that controls their synthesis. Once that is in place, development of complex multicellularity appears simple, as the multiple, independent origins of multicellularity attest (Antonio and Schulze-Makuch 2012; Knoll 2011; Rokas 2008). Indeed, as Wilkins (2002) says 'According to this point of view, the foundations of developmental evolution were laid long before there were multicellular eukaryotes. A crucial step was the evolution of molecular organizational modules involving both signal transduction and gene transcription systems. And, perhaps, with that property, the advent of multicellular complexity was a virtual inevitability.' (page 350).

The ability to build a complex developmental programme depends on the ability to build a complex genetic programme. Although the human genome is 1000 times the size of that of *E. coli*, it only codes for $\sim$10 times as many proteins (Venter

et al. 2001; Lander et al. 2001). Expansion of genome complexity in animals is almost all related to expansion of non-coding regions [reviewed in Mattick and Makunin 2006; Geisler and Coller 2013; Necsulea et al. 2014], not coding sequences. Many of these non-coding sequences are related to gene control.

Like cell structure, the chemistry of control of the genome was thought to be completely different between prokaryotes and eukaryotes. However, more detailed analysis has shown that all the features of eukaryotic genomes are present in prokaryotes (Elliott and Gregory 2015; Bains and Schulze-Makuch 2015), and the eukaryotic genome 'only' represents an enormous increase in complexity of gene control chemistry that exists in all forms of life. I just summarize this evidence here, before moving on to the key difference that I think does distinguish prokaryotes from eukaryotes. The reader is directed to Bains and Schulze-Makuch (2015) for a much more detailed analysis of all the chemical points below.

The genetic architecture of eukaryotes can be classified (purely for our convenience) into basic chromatin chemistry, the transcription apparatus, local and non-local control systems, and post-transcriptional mechanisms.

All domains of life condense DNA, so that it fits inside the cell, using specialist proteins, which together form nucleoprotein. The structure of nucleoprotein is also used to control gene activity in the DNA in all domains of life, through local or long-range interactions (Luijsterburg et al. 2008). ATP-driven remodelling of chromatin is essential for transcription in eukaryotes (Kireeva et al. 2002; Mizuguchi et al. 2004; Shen et al. 2000; Olave et al. 2002). Eukaryotic chromatin chemistry and higher-order structure have homologues in archaea (Sandman and Reeve 2001; Sandman and Reeve 2005; White and Bell 2002) and analogues in bacteria (Sandman and Reeve 2001; Drlica and Rouviere-Yaniv 1987). In eukaryotes, this machinery is targeted by local chromatin modification. There is analogous chemistry to DNA-binding-protein modification in prokaryotes, but its targeting is different (Wardleworth et al. 2002; Soppa 2010). The chemistry of the modification of the core components of gene control has evolved multiple times. DNA base modification is achieved in bacteria, archaea and eukaryotes by unrelated systems (Chan et al. 2004; Cao et al. 2003; Gaspin et al. 2000; Kumar et al. 1994) and has evolved independently. RNA modification is similarly diverse and universal.

Both eukaryotic and bacterial nucleoproteins are organized into larger-scale structure, the eukaryotic genome primarily by RNA and the bacterial genome by protein. Notably, the organization of the bacterial genome into 10-kb loops is related to the control of genes (Navarre et al. 2006, 2007; Zhang et al. 1996) and is rapidly assembled in a specific order on DNA replication (Viollier et al. 2004). So, gene control in prokaryotes and eukaryotes is a combination of direct targeting of DNA and indirectly targeting it through changing the structure of nucleoprotein to allow (or prevent) transcription.

In eukaryotes, RNA is used very extensively to control genes through interaction with nucleoprotein. Long RNAs with no obvious coding function (long non-coding RNAs—lncRNAs) are integral to the control machinery [reviewed in Ulitsky et al. 2011; Guttman et al. 2010; Ponting et al. 2009; Nagano and Fraser 2011]. However,

RNA involvement in gene control is not confined to eukaryotes. Long and short RNAs are involved in control of transcription, translation and mRNA breakdown in prokaryotes as well (Washietl et al. 2007; Livny et al. 2006; Vockenhuber et al. 2011; Rivas et al. 2001), sometimes in mechanisms that include proteins homologous to eukaryotic proteins.

It is now well known that RNA splicing occurs in all domains of life and probably occurred in the last common ancestor of all life [see Edgell et al. 2011; Martin and Koonin 2006; Roy and Gilbert 2006; Pyle 2012; Dumesic et al. 2013; William and Gilbert 2006]. The unrelated mechanism of translational skipping has a similar net effect to RNA splicing, the generation of a protein from non-adjacent regions of a transcript, and occurs in all domains. RNA is also central to priming DNA synthesis at specific sites in prokaryotes and eukaryotes.

All branches of life have a wealth of (unrelated) transcription factors to control the process of initiation of RNA synthesis (Bell and Jackson 2001; Baliga et al. 2000). The structure of the transcription control complex and its target promoter sequences in archaea are more similar to those in eukaryotes than those in bacteria, suggesting homology (Helmann and Chamberlin 1988; Rhee and Pugh 2012; Miller and Hahn 2006). Transcription elongation and termination is also controlled by a range of chemistries. Protein turnover is modulated by a range of systems in prokaryotes (Battesti and Gottesman 2013) and eukaryotes (Pickart 2001; Geng et al. 2012).

Figure 15.1 shows a very high level summary of these findings. Here, the interaction between molecules is classified (rather arbitrarily) into the targeting agent (e.g. a classic Jacob and Monod (1961) repressor protein) and a targeted moiety (e.g. the promoter DNA). DNA, short RNAs and proteins can be modified, for example, by methylation or acetylation, through the action of enzymes that are themselves targeted using DNA, RNA or protein. Figure 15.1 illustrates, and Bains and Schulze-Makuch (2015) summarize in much more detail, that almost any type of interaction chemistry found to be central to gene control in eukaryotes is also found in prokaryotes. Usually, the specific proteins and RNAs involved are not homologous, suggesting these mechanisms evolved separately. Bains and Schulze-Makuch (2015) suggest, based on their evolutionary model, that this means that these mechanisms are not uniquely unlikely evolutionary innovations, but are relatively easy to evolve.

This suggests that chemistry is not a key difference between the genetics of prokaryotes and eukaryotes. This leads to a problem. Endosymbiosis is common and has occurred frequently in evolutionary terms. The creation of mitochondria through endosymbiosis and subsequent specialization and gene transfer gave eukaryotic cells a greatly increased capacity to generate energy, which they spent on genetic control. This allowed more complex genetic circuitry. So why did endosymbiosis not allow prokaryotic cells to develop high-energy genetics? For that matter, why did not internal compartmentalization or the other mechanisms above allow this to happen?

A potential solution is in the genetic logic of eukaryotes. Unlike eukaryotic chromatin, prokaryotic nucleoprotein does not (in general) block transcription (Weinzierl 2013; Xie and Reeve 2004). Transcription of prokaryotic genes is enabled by recruitment of RNA polymerases to promoters and blocked by specific

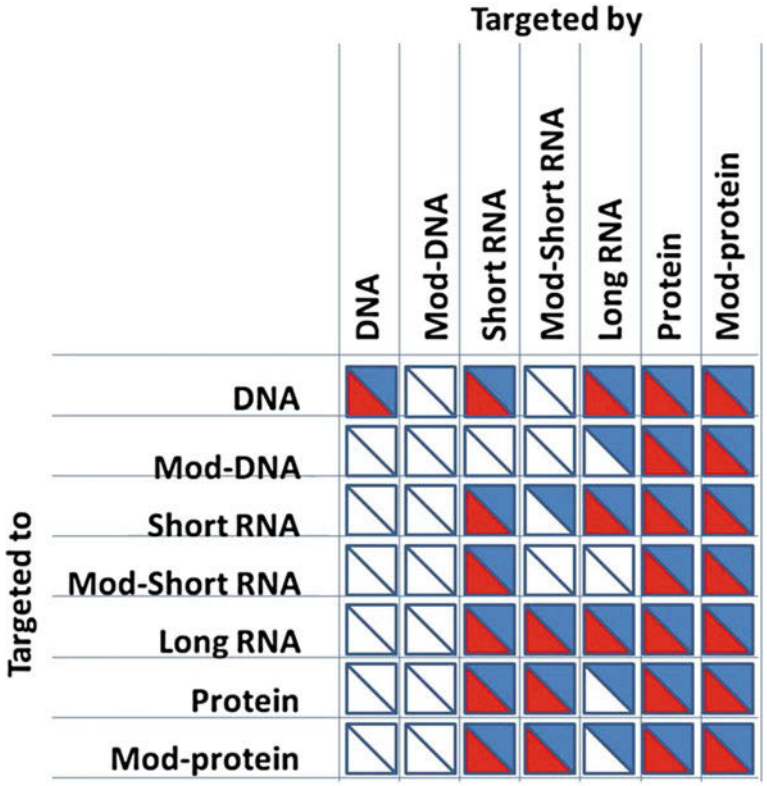


Fig. 15.1 Cartoon summary of chemical interactions in gene expression controls in eukaryotes and prokaryotes. Summarizing the analysis of (Bains and Schulze-Makuch 2015). Each cell represents the targeting of transcription or translation by a chemical type (*top*) to a chemical type (*left side*). *Red* found in prokaryotes and *blue* found in eukaryotes

promoter-binding proteins. By contrast, eukaryotic DNA is universally blocked from transcription by chromatin, which must be opened by energy-requiring reactions to be transcribed (Kireeva et al. 2002; Mizuguchi et al. 2004; Shen et al. 2000; Olave et al. 2002). Only when the chromatin around a gene is opened can the specific transcription and enhancer proteins bind to the start of a gene and initiate transcription. Continued, energy-expensive chromatin modelling is also needed to allow transcription to continue (Nechaev and Adelman 2011). Eukaryotic chromatin is inherently unable to be transcribed unless it is activated, while prokaryotic nucleoprotein is inherently able to be transcribed unless it is silenced. Eukaryotic chromatin is 'off' and prokaryotic nucleoprotein is 'on'.

This is central to the development of complex genetic controls, because genetic control is not structured like computer code, no matter that we describe it in those terms. In real organisms, many control systems touched on briefly above and described in more detail in Bains and Schulze-Makuch (2015) interact with *all* the

other mechanisms to control gene expression, and hence cell type and behaviour. In the animal examples that have been studied exhaustively, it is found that all of miRNA, piRNA, lncRNAs, protein transcription factors, specific DNA sequence elements, histone methylation and acetylation interact with each other in a genuinely chaotic network. The complexity of a genome is therefore not a linear function of the number of sequences, but rather a function of the number of ways those sequences and their products can interact—at least a polynomial function.

If a gene is duplicated, this does not add to that complexity. But if there is a selective advantage for the two copies to diverge, adopting different functions in space and time, then the relevant control mechanisms have not only to interact correctly with each other to make this happen, but also *not* to interact with all the other genetic control elements in the cell. If all the other genes in the nucleus are by default 'off', then the changed gene is only likely to be active in a few cell situations, and in those situations most of the other genes are silent. The adaptation of other genes within the cell is a less burdensome path. If genes are by default expressed, then allowing the new gene to be expressed is easier, avoiding unwanted changes in other genetic programmes is harder, requires changes across the whole genome. As the genome gets larger, this latter becomes harder to achieve.

Following this argument, Bains and Schulze-Makuch (2015) suggested that the original difference between eukaryotes and other domains of life was the adoption of a 'default off' genetic logic, which greatly facilitates the expansion of genome complexity. Enhanced genome complexity means more energy per gene is needed for control, at all levels from transcription to protein degradation. (I note that 'genetic control' is not just the control of transcription, but any control process driven by the genes.) This drives the emerging eukaryote to find better strategies to capture energy including symbiosis, and ultimately endosymbiosis, with other organisms.

15.4 Conclusions

I have examined a range of features that distinguish prokaryotes and eukaryotes. Two stand out as potentially fundamental—eukaryotes' ability to capture much more energy per gene, and their 'default off' genetic logic, which allows for complexification of genetic controls. Both are needed, as complex genetic controls consume a great deal of energy. Eukaryotes have acquired the means to capture high-energy fluxes by endosymbiosis and subsequent specialization of a eubacterium. However, the initial steps of this process—symbiosis, endosymbiosis, prokaryote-to-eukaryote gene transfer and metabolic interdependence—have occurred many times, and examples of all the stages from loose association to DNA-less 'degenerate' mitochondria can be seen today. Other mechanisms for capturing more energy are also feasible. What allowed the eukaryotes to develop complex genetics, and hence use the endosymbiotic proto-mitochondrion to such great effect, was their possession of a genetic logic that allowed them to add layers

of genetic complexity to their genome without disrupting the circuits that were already there. This in turn allowed the development of complex developmental programmes that could code for the assembly of complex organisms.

Was this a uniquely improbable event, or would complex organisms evolve on any world where life appeared, given time and a sufficiently stable environment? All the steps from a proto-eukaryote to man appear to be highly likely, including the capture of mitochondria. At least three lineages (animals, plants and fungi) have independently built on this proto-eukaryotic substructure to develop complex developmental programmes and multicellular organisms. So this question boils down to one of whether evolution of a 'default off' genetic logic is likely or not. At the moment, we cannot answer this, but it is possible that future research in synthetic biology or evolutionary simulation will give us some clues.

Acknowledgements I am very grateful to my colleague Dirk Schulze-Makuch (Washington State University, WA, USA, and Technical University, Berlin) for our work together on the major steps to complex life. I am also grateful to Janusz Petkowski (MIT, MA, USA) for many helpful comments, to Sara Seager (MIT, MA, USA) for her unfailing and generous support in this and other work, and to Pierre Pontarotti (Ai-Marseilles University, France) and the staff and attendees of the 19th Evolutionary Biology Meeting (Marseilles, France, September 2015) for encouraging me to order my thoughts on why any of us are here at all.

References

Angata T, Varki A (2002) Chemical diversity in the sialic acids and related α-keto acids: an evolutionary perspective. Chem Rev 102(2):439–470. doi:10.1021/cr000407m

Angert ER (2012) DNA replication and genomic architecture of very large bacteria. Ann Rev Microbiol 66(1):197–212. doi:10.1146/annurev-micro-090110-102827

Angert ER, Clements KD, Pace NR (1993) The largest bacterium. Nature 362(6417):239–241

Antonio M, Schulze-Makuch D (2012) Toward a new understanding of multicellularity 2, 1

Bains W, Schulze-Makuch D (2015) Mechanisms of evolutionary innovation point to genetic control logic as the key difference between prokaryotes and eukaryotes. J Mol Evol:1–20. doi:10.1007/s00239-00015-09688-00236). doi:10.1007/s00239-015-9688-6

Bains W, Xiao Y, Yu C (2015) Prediction of the maximum temperature for life based on the stability of metabolites to decomposition in water. Life 2:1054–1100

Baliga NS, Goo YA, Ng WV, Hood L, Daniels CJ, DasSarma S (2000) Is gene expression in Halobacterium NRC-1 regulated by multiple TBP and TFB transcription factors? Mol Microbiol 36(5):1184–1185. doi:10.1046/j.1365-2958.2000.01916.x

Battesti A, Gottesman S (2013) Roles of adaptor proteins in regulation of bacterial proteolysis. Curr Opin Microbiol 16(2):140–147. doi:10.1016/j.mib.2013.01.002

Bell SD, Jackson SP (2001) Mechanism and regulation of transcription in archaea. Curr Opin Microbiol 4(2):208–213. doi:10.1016/S1369-5274(00)00190-9

Bentkowski P, Van Oosterhout C, Mock T (2015) A model of genome size evolution for prokaryotes in stable and fluctuating environments. Genome Biol Evol 7(8):2344–2351. doi:10.1093/gbe/evv148

Bonner JT (2004) Perspective: the size-complexity rule. Evolution 58(9):1883–1890. doi:10.1111/j.0014-3820.2004.tb00476.x

Booth A, Doolittle WF (2015) Reply to Lane and Martin: being and becoming eukaryotes. Proc Natl Acad Sci 112(35):E4824. doi:10.1073/pnas.1513285112

Cao X, Aufsatz W, Zilberman D, Mette MF, Huang MS, Matzke M, Jacobsen SE (2003) Role of the DRM and CMT3 Methyltransferases in RNA-Directed DNA methylation. Curr Biol 13(24):2212–2217. doi:10.1016/j.cub.2003.11.052

Cavalier-Smith T (2013) Symbiogenesis: mechanisms, evolutionary consequences, and systematic implications. Ann Rev Ecol Evol Syst 44(1):145–172. doi:10.1146/annurev-ecolsys-110411-160320

Chan SW-L, Ziberman D, Xie Z, Johansen LK, Carrington JC, Jacobsen SE (2004) RNA silencing genes control de Novo DNA methylation. Science 303:1336

Cowan DA (2004) The upper temperature for life—where do we draw the line? Trends Microbiol 12(2):58–60

de Sousa R, António M, Schulze-Makuch D (2012) Toward a new understanding of multicellularity. Hypotheses Life Sci 2(1):4–14

Dolan MF (2001) Speciation of termite gut protists: the role of bacterial symbionts. Int Microbiol 4(4):203–208

Drlica K, Rouviere-Yaniv J (1987) Histonelike proteins of bacteria. Microbiol Rev 51(3):301–319

Dumesic Phillip A, Natarajan P, Chen C, Drinnenberg Ines A, Schiller Benjamin J, Thompson J, Moresco James J, Yates Iii John R, Bartel David P, Madhani Hiten D (2013) Stalled spliceosomes are a signal for RNAi-mediated genome defense. Cell 152(5):957–968. doi:10.1016/j.cell.2013.01.046

Edgell D, Chalamcharla V, Belfort M (2011) Learning to live together: mutualism between self-splicing introns and their hosts. BMC Biol 9(1):22

Elliott TA, Gregory TR (2015) What's in a genome? The C-value enigma and the evolution of eukaryotic genome content. Philos Trans R Soc Lon B: Biol Sci 370 (1678). doi:10.1098/rstb.2014.0331

Erickson HP (1997) FtsZ, a tubulin homologue in prokaryote cell division. Trends Cell Biol 7(9):362–367. doi:10.1016/S0962-8924(97)01108-2

Ferdows MS, Barbour AG (1989) Megabase-sized linear DNA in the bacterium Borrelia burgdorferi, the Lyme disease agent. Proc Natl Acad Sci 86(15):5969–5973

Fuerst JA (2005) Intracellular compartmentalization in Planctomycetes. Ann Rev Microbiol 59:299–328

Fuerst JA, Webb RI, Garson MJ, Hardy L, Reiswig HM (1998) Membrane-bounded nucleoids in microbial symbionts of marine sponges. FEMS Microbiol Lett 166(1):29–34. doi:10.1111/j.1574-6968.1998.tb13179.x

Fusetani N (2012) Marine natural products. In: Civjan N (ed) Natural products in chemical biology. Wiley, Hoboken, pp 31–64

Gaspin C, Cavaillé J, Erauso G, Bachellerie J-P (2000) Archaeal homologs of eukaryotic methylation guide small nucleolar RNAs: lessons from the Pyrococcus genomes. J Mol Biol 297(4):895–906. doi:10.1006/jmbi.2000.3593

Geisler S, Coller J (2013) RNA in unexpected places: long non-coding RNA functions in diverse cellular contexts. Nat Rev Mol Cell Biol 14(11):699–712. doi:10.1038/nrm3679

Geng F, Wenzel S, Tansey WP, Tansey WP (2012) Ubiquitin and proteasomes in transcription. Ann Rev Biochem 81:177–201

Griese M, Lange C, Soppa J (2011) Ploidy in cyanobacteria. FEMS Microbiol Lett 323(2):124–131

Guieysse B, Wuertz S (2012) Metabolically versatile large-genome prokaryotes. Curr Opin Biotechnol 23(3):467–473. doi:10.1016/j.copbio.2011.12.022

Gunatilaka AL (2012) Plant natural products. In: Civjan N (ed) Natural products in chemical biology. Wiley, Hoboken, pp 3–29

Guttman M, Garber M, Levin JZ, Donaghey J, Robinson J, Adiconis X, Fan L, Koziol MJ, Gnirke A, Nusbaum C, Rinn JL, Lander ES, Regev A (2010) Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. Nat Biotech 28 (5):503–510. doi:http://www.nature.com/nbt/journal/v28/n5/abs/nbt.1633.html#supplementary-information

Helmann JD, Chamberlin MJ (1988) Structure and function of bacterial sigma factors. Annu Rev Biochem 57(1):839–872. doi:10.1146/annurev.bi.57.070188.004203
Hinnebusch J, Tilly K (1993) Linear plasmids and chromosomes in bacteria. Mol Microbiol 10(5):917–922. doi:10.1111/j.1365-2958.1993.tb00963.x
Jacob F, Monod J (1961) On the regulation of gene activity. Cold Spring Harb Symp Quant Biol 26:193–211. doi:10.1101/sqb.1961.026.01.024
Javor B (2012) Hypersaline environments: microbiology and biogeochemistry. Springer, Berlin
Kapranov P, Cheng J, Dike S, Nix DA, Duttagupta R, Willingham AT, Stadler PF, Hertel J, Hackermüller J, Hofacker IL, Bell I, Cheung E, Drenkow J, Dumais E, Patel S, Helt G, Ganesh M, Ghosh S, Piccolboni A, Sementchenko V, Tammana H, Gingeras TR (2007) RNA maps reveal new RNA classes and a possible function for pervasive transcription. Science 316(5830):1484–1488. doi:10.1126/science.1138341
Karin M, Hunter T (1995) Transcriptional control by protein phosphorylation: signal transmission from the cell surface to the nucleus. Curr Biol 5(7):747–757. doi:10.1016/S0960-9822(95)00151-5
Kireeva ML, Walter W, Tchernajenko V, Bondarenko V, Kashlev M, Studitsky VM (2002) Nucleosome remodeling induced by RNA polymerase II: loss of the H2A/H2B dimer during transcription. Mol Cell 9(3):541–552. doi:10.1016/S1097-2765(02)00472-0
Knoll AH (2011) The multiple origins of complex multicellularity. Ann Rev Earth Planet Sci 39(1):217–239. doi:10.1146/annurev.earth.031208.100209
Knoll AH, Hewitt D (2011) Phylogenetic, functional and geological perspectives on complex multicellularity. In: Chalcott B, Sterelny K (eds) The major transitions in evolution revisited. MIT Press, Cambridge, pp 251–270
Komaki K, Ishikawa H (2000) Genomic copy number of intracellular bacterial symbionts of aphids varies in response to developmental stage and morph of their host. Insect Biochem Mol Biol 30(3):253–258. doi:10.1016/S0965-1748(99)00125-3
Kube M, Schneider B, Kuhl H, Dandekar T, Heitmann K, Migdoll A, Reinhardt R, Seemuller E (2008) The linear chromosome of the plant-pathogenic mycoplasma '*Candidatus Phytoplasma mali*'. BMC Genom 9(1):306
Kulp A, Kuehn MJ (2010) Biological functions and biogenesis of secreted bacterial outer membrane vesicles. Annu Rev Microbiol 64:163–184. doi:10.1146/annurev.micro.091208.073413
Kumar S, Cheng X, Klimasauskas S, Mi S, Posfai J, Roberts RJ, Wilson GG (1994) The DNA (cytosine-5) methyltransferases. Nucleic Acids Res 22(1):1–10
Landenmark HKE, Forgan DH, Cockell CS (2015) An estimate of the total DNA in the biosphere. PLoS Biol 13(6):e1002168
Lander et al (2001) Initial sequencing and analysis of the human genome. Nature 409(6822):860–921. doi:http://www.nature.com/nature/journal/v409/n6822/suppinfo/409860a0_S1.html
Lane N (2011) Energetics and genetics across the prokaryote-eukaryote divide. Biol Direct 6:35
Lane N, Martin W (2010) The energetics of genome complexity. Nature 467(7318):929–934
Lane N, Martin WF (2015) Eukaryotes really are special, and mitochondria are why. Proc Natl Acad Sci 112(35):E4823. doi:10.1073/pnas.1509237112
Lane N, Martin WF (2016) Mitochondria, complexity, and evolutionary deficit spending. Proc Natl Acad Sci 113(6):E666. doi:10.1073/pnas.1522213113
Livny J, Brencic A, Lory S, Waldor MK (2006) Identification of 17 *Pseudomonas aeruginosa* sRNAs and prediction of sRNA-encoding genes in 10 diverse pathogens using the bioinformatic tool sRNAPredict2. Nucleic Acids Res 34(12):3484–3493. doi:10.1093/nar/gkl453
Lloyd D, Ralphs JR, Harris JC (2002) Giardia intestinalis, a eukaryote without hydrogenosomes, produces hydrogen. Microbiology 148(3):727–733. doi:10.1099/00221287-148-3-727
Lodé T (2012) For quite a few chromosomes more: the origin of eukaryotes…. J Mol Biol 423(2):135–142. doi:10.1016/j.jmb.2012.07.005

Lu P, Vogel C, Wang R, Yao X, Marcotte EM (2007) Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. Nat Biotech 25(1):117–124. doi:10.1038/nbt1270

Luijsterburg MS, White MF, Van Driel R, Dame RT (2008) The major architects of chromatin: architectural proteins in bacteria, archaea and eukaryotes. Crit Rev Biochem Mol Biol 43:1–26

Lynch M, Marinov GK (2015) The bioenergetic costs of a gene. Proc Natl Acad Sci 112(51): 15690–15695. doi:10.1073/pnas.1514974112

Lynch M, Marinov GK (2016) Reply to Lane and Martin: mitochondria do not boost the bioenergetic capacity of eukaryotic cells. Proc Natl Acad Sci 113(6):E667–E668. doi:10.1073/pnas.1523394113

Makarieva AM, Gorshkov VG, Li B-L, Chown SL, Reich PB, Gavrilov VM (2008) Mean mass-specific metabolic rates are strikingly similar across life's major domains: evidence for life's metabolic optimum. Proc Natl Acad Sci 105(44):16994–16999. doi:10.1073/pnas.0802148105

Marin B, M. Nowack EC, Melkonian M (2005) A plastid in the making: evidence for a second primary endosymbiosis. Protist 156(4):425–432. doi:http://dx.doi.org/10.1016/j.protis.2005.09.001

Mariscal C, Doolittle WF (2015) Eukaryotes first: how could that be? Phil Trans Roy Soc B 370:20140322

Martin WF (2011) Early evolution without a tree of life. Biol Direct 6:36

Martin W, Koonin EV (2006) Introns and the origin of nucleus–cytosol compartmentalization. Nature 440:41–45

Martin WF, Garg S, Zimorski V (2015) Endosymbiotic theories for eukaryote origin. Philos Trans R Soc Lon B: Biol Sci 370 (1678). doi:10.1098/rstb.2014.0330

Mattick JS, Makunin IV (2006) Non-coding RNA. Hum Mol Genet 15(suppl 1):R17–R29. doi:10.1093/hmg/ddl046

Maynard Smith J, Szathmary E (1995) The major transitions in evolution. WH Freeman, Oxford

McCutcheon JP, Moran NA (2012) Extreme genome reduction in symbiotic bacteria. Nat Rev Microbiol 10:13–26

McGuire M (2015) Protein kinases. Current progress CALLISTO REFERENCENY, New York

Mendell JE, Clements KD, Choat JH, Angert ER (2008) Extreme polyploidy in a large bacterium. Proc Natl Acad Sci 105(18):6730–6734. doi:10.1073/pnas.0707522105

Miller G, Hahn S (2006) A DNA-tethered cleavage probe reveals the path for promoter DNA in the yeast preinitiation complex. Nat Struct Mol Biol 13(7):603–610. doi:http://www.nature.com/nsmb/journal/v13/n7/suppinfo/nsmb1117_S1.html

Mizuguchi G, Shen X, Landry J, Wu W-H, Sen S, Wu C (2004) ATP-driven exchange of histone H2AZ variant catalyzed by SWR1 chromatin remodeling complex. Science 303(5656): 343–348. doi:10.1126/science.1090701

Moens S, Vanderleyden J (1997) Glycoproteins in prokaryotes. Arch Microbiol 168(3):169–175. doi:10.1007/s002030050484

Montgomery WL, Pollak PE (1988) *Epulopiscium fishelsoni* N. G., N. Sp., a protist of uncertain taxonomic affinities from the gut of an herbivorous reef fish. Eukaryot Microbiol 35(4): 565–569

Moran NA, McCutcheon JP, Nakabachi A (2008) Genomics and evolution of heritable bacterial symbionts. Ann Rev Genet 42:165–190

Müller M, Mentel M, van Hellemond JJ, Henze K, Woehle C, Gould SB, Yu R-Y, van der Giezen M, Tielens AGM, Martin WF (2012) Biochemistry and evolution of anaerobic energy metabolism in eukaryotes. Microbiol Mol Biol Rev 76(2):444–495. doi:10.1128/mmbr.05024-11

Nagamune K, Xiong L, Chini E, Sibley LD (2008) Plants, endosymbionts and parasites. Communicative Integr Biol 1(1):62–65. doi:10.4161/cib.1.1.6106

Nagano T, Fraser P (2011) No-nonsense functions for long noncoding RNAs. Cell 145:178–181

Navarre WW, Porwollik S, Wang Y, McClelland M, Rosen H, Libby SJ, Fang FC (2006) Selective silencing of foreign DNA with low GC content by the H-NS protein in salmonella. Science 313 (5784):236–238. doi:10.1126/science.1128794

Navarre WW, McClelland M, Libby SJ, Fang FC (2007) Silencing of xenogeneic DNA by H-NS—facilitation of lateral gene transfer in bacteria by a defense system that recognizes foreign DNA. Genes Dev 21(12):1456–1471. doi:10.1101/gad.1543107

Nechaev S, Adelman K (2011) Pol II waiting in the starting gates: regulating the transition from transcription initiation into productive elongation. Biochim et Biophys Acta (BBA)—Gene Regul Mech 1809(1):34–45. doi:http://dx.doi.org/10.1016/j.bbagrm.2010.11.001

Necsulea A, Soumillon M, Warnefors M, Liechti A, Daish T, Zeller U, Baker JC, Grutzner F, Kaessmann H (2014) The evolution of lncRNA repertoires and expression patterns in tetrapods. Nature 505 (7485):635–640. doi:10.1038/nature12943. http://www.nature.com/nature/journal/v505/n7485/abs/nature12943.html#supplementary-information

Norris V, Turnock G, Sigee D (1996) The *escherichia coli* enzoskeleton. Mol Microbiol 19 (2):197–204. doi:10.1046/j.1365-2958.1996.373899.x

Olave IA, Peck-Peterson SI, Crabtree GR (2002) Nuclear actin and actin-related proteins in chromatin remodelling. Ann Rev Biochem 71:755–781

Ørom UA, Derrien T, Beringer M, Gumireddy K, Gardini A, Bussotti G, Lai F, Zytnicki M, Notredame C, Huang Q, Guigo R, Shiekhattar R (2010) Long noncoding RNAs with enhancer-like function in human cells. Cell 143(1):46–58. doi:10.1016/j.cell.2010.09.001

Pickart CM (2001) Mechanisms underlying ubiquitination. Ann Rev. Biochem 70:503–533

Pittis AA, Gabaldón T (2016) Late acquisition of mitochondria by a host with chimaeric prokaryotic ancestry. Nat Adv Online Publication. doi:10.1038/nature16941. http://www.nature.com/nature/journal/vaop/ncurrent/abs/nature16941.html#supplementary-information

Ponting CP, Oliver PL, Reik W (2009) Evolution and functions of long noncoding RNAs. Cell 136(4):629–641. doi:10.1016/j.cell.2009.02.006

Pyle AM (2012) Group II intron architecture and its implications for the development of eukaryotic splicing systems. FASEB J 26(217):213

Reyes-Prieto A, Yoon HS, Moustafa A, Yang EC, Andersen RA, Boo SM, Nakayama T, K-i Ishida, Bhattacharya D (2010) Differential gene retention in plastids of common recent origin. Mol Biol Evol 27(7):1530–1537. doi:10.1093/molbev/msq032

Rhee HS, Pugh F (2012) Genome-wide structure and organization of eukaryotic pre-initiation complexes. Nature 483:295–301

Rivas E, Klein RJ, Jones TA, Eddy SR (2001) Computational identification of noncoding RNAs in *E. coli* by comparative genomics. Curr Biol 11(17):1369–1373. doi:10.1016/S0960-9822(01)00401-8

Roeben A, Kofler C, Nagy I, Nickell S, Ulrich Hartl F, Bracher A (2006) Crystal structure of an archaeal actin homolog. J Mol Biol 358(1):145–156. doi:10.1016/j.jmb.2006.01.096

Rokas A (2008) The origins of multicellularity and the early history of the genetic toolkit for animal development. Ann Rev Genet 42(1):235–251. doi:10.1146/annurev.genet.42.110807.091513

Roy SW, Gilbert W (2006) The evolution of spliceosomal introns: patterns, puzzles and progress. Nat Rev Genet 7:211–221

Sanchez S, Guzman-Trampe S, Avalos M, Ruiz B, Rodriguez-Sanoja R, Jim´enez-Estrada M (2012) Bacterial natural products. In: Civjan N (ed) Natural products in chemical biology. Wiley, Hoboken, pp 65–108

Sandman K, Reeve JN (2001) Chromosome packaging by archael histones. In: Laskin AL, Bennett JW, Gadd gM (eds) Advances in applied microbiology, vol 50. Academic Press, San Diego, pp 73–100

Sandman K, Reeve JN (2005) Archaeal chromatin proteins: different structures but common function? Curr Opin Microbiol 8(6):656–661. doi:10.1016/j.mib.2005.10.007

Sassera D, Beninati T, Bandi C, Bouman EAP, Sacchi L, Fabbi M, Lo N (2006) 'Candidatus Midichloria mitochondrii', an endosymbiont of the tick Ixodes ricinus with a unique intramitochondrial lifestyle. Int J Syst Evol Microbiol 56(11):2535–2540. doi:10.1099/ijs.0.64386-0

Schulz HN, Brinkhoff T, Ferdelman TG, Mariné MH, Teske A, Jørgensen BB (1999) Dense populations of a giant sulfur bacterium in namibian shelf sediments. Science 284(5413): 493–495. doi:10.1126/science.284.5413.493

Shen X, Mizuguchi G, Hamiche A, Wu C (2000) A chromatin remodelling complex involved in transcription and DNA processing. Nature 406 (6795):541–544. doi:http://www.nature.com/nature/journal/v406/n6795/suppinfo/406541A0_S1.html

Sherman L, Min H, Toepel J, Pakrasi H (2010) Better living through cyanothece—unicellular diazotrophic cyanobacteria with highly versatile metabolic systems. In: Hallenbeck PC (ed) Recent advances in phototrophic prokaryotes, vol 675. Advances in experimental medicine and biology. Springer New York, pp 275–290. doi:10.1007/978-1-4419-1528-3_16

Soppa J (2010) Protein acetylation in archaea, bacteria, and eukaryotes. Archaea. doi:10.1155/2010/820681

Soppa J (2014) Polyploidy in archaea and bacteria: about desiccation resistance, giant cell size, long-term survival, enforcement by a eukaryotic host and additional aspects. J Mol Microbiol Biotechnol 24(5–6):409–419

Spang A, Saw JH, Jorgensen SL, Zaremba-Niedzwiedzka K, Martijn J, Lind AE, van Eijk R, Schleper C, Guy L, Ettema TJG (2015) Complex archaea that bridge the gap between prokaryotes and eukaryotes. Nature 521(7551):173–179. doi:10.1038/nature14447. http://www.nature.com/nature/journal/v521/n7551/abs/nature14447.html#supplementary-information

Stolz JF (2001) Bacterial intracellular membranes. In: eLS. Wiley. doi:10.1002/9780470015902.a0000303.pub2

Szathmáry E (2015) Toward major evolutionary transitions theory 2.0. Proc Natl Acad Sci 112 (33):10104–10111. doi:10.1073/pnas.1421398112

Thao ML, Gullan PJ, Baumann P (2002) Secondary (γ-Proteobacteria) endosymbionts infect the primary (β-Proteobacteria) endosymbionts of mealybugs multiple times and coevolve with their hosts. Appl Environ Microbiol 68(7):3190–3197. doi:10.1128/aem.68.7.3190-3197.2002

Tirichine L, Bowler C (2011) Decoding algal genomes: tracing back the history of photosynthetic life on Earth. Plant J 66:45–57

Ulitsky I, Shkumatava A, Jan Calvin H, Sive H, Bartel David P (2011) Conserved function of lincRNAs in vertebrate embryonic development despite rapid sequence evolution. Cell 147 (7):1537–1550. doi:10.1016/j.cell.2011.11.055

van Der Giezen M (2009) Hydrogenosomes and mitosomes: conservation and evolution of functions. J Eukaryot Microbiol 56(3):221–231

van der Giezen M, Tovar J (2005) Degenerate mitochondria. EMBO Rep 6(6):525–530. doi:10.1038/sj.embor.7400440

Varki A, Schauer R (2009) Sialic acids

Venter et al (2001) The sequence of the human genome. Science 291 (5507):1304–1351. doi:10.1126/science.1058040

Viollier PH, Thanbichler M, McGrath PT, West L, Meewan M, McAdams HH, Shapiro L (2004) Rapid and sequential movement of individual chromosomal loci to specific subcellular locations during bacterial DNA replication. Proc Nat Acad Sci 101 (9257–9262)

Vockenhuber M-P, Sharma CM, Statt MG, Schmidt D, Xu Z, Dietrich S, Liesegang H, Mathews DH, Suess B (2011) Deep sequencing-based identification of small non-coding RNAs in *Streptomyces coelicolor*. RNA Biol 8(3):468–477

Wardleworth BN, Russell RJM, Bell SD, Taylor GL, White MF (2002) Structure of Alba: an archaeal chromatin protein modulated by acetylation. EMBO J 21(17):4654–4662. doi:10.1093/emboj/cdf465

Washietl S, Pedersen JS, Korbel JO, Stocsits C, Gruber AR, Hackermüller J, Hertel J, Lindemeyer M, Reiche K, Tanzer A, Ucla C, Wyss C, Antonarakis SE, Denoeud F, Lagarde J, Drenkow J, Kapranov P, Gingeras TR, Guigó R, Snyder M, Gerstein MB, Reymond A, Hofacker IL, Stadler PF (2007) Structured RNAs in the ENCODE selected regions of the human genome. Genome Res 17(6):852–864. doi:10.1101/gr.5650707

Weinzierl ROJ (2013) The RNA polymerase factory and archaeal transcription. Chem Rev 113:8350–8376
White MF, Bell SD (2002) Holding it together: chromatin in the Archaea. Trends Genet 18(12):621–626. doi:10.1016/S0168-9525(02)02808-1
Whitman WB, Coleman DC, Wiebe WJ (1998) Prokaryotes: the unseen majority. Proc Natl Acad Sci 95(12):6578–6583
Wilkins A (2002) The evolution of developmental pathways. Sinauer Associates, Sunderland
William RS, Gilbert W (2006) The evolution of spliceosomal introns: patterns, puzzles and progress. Nat Rev Genet 7(3):211–221
Williams JP, Hallsworth JE (2009) Limits of life in hostile environments: no barriers to biosphere function? Environ Microbiol 11(12):3292–3308. doi:10.1111/j.1462-2920.2009.02079.x
Williams TA, Foster PG, Cox CJ, Embley TM (2014) An archeal origin of eukaryotes supports only two primary domains of life. Nature 504:231–236
Woese CR, Kandler O, Wheelis ML (1990) Towards a natural system of organisms: proposal for the domains archaea, bacteria, and eucarya. Proc Natl Acad Sci 87(12):4576–4579. doi:10.1073/pnas.87.12.4576
Wujek DE (1979) Intracellular bacteria in the blue-green alga *Pleurocapsa minor*. Trans Am Microsc Soc 98(1):143–145
Xie Y, Reeve JN (2004) Transcription by an Archaeal RNA polymerase is slowed but not blocked by an Archaeal nucleosome. J Bacteriol 186(11):3492–3498. doi:10.1128/jb.186.11.3492-3498.2004
Yutin N, Koonin EV (2012) Archaeal origin of tubulin. Biol Direct 7 (10)
Zerulla K, Soppa J (2014) Polyploidy in haloarchaea: advantages for growth and survival. Front Microbiol 5:274. doi:10.3389/fmicb.2014.00274
Zhang A, Rimsky S, Reaban ME, Buc H, Belfort M (1996) Escherichia coli protein analogs StpA and H-NS: regulatory loops, similar and disparate effects on nucleic acid dynamics. EMBO J 15(6):1340–1349
Zimorski V, Ku C, Martin WF, Gould SB (2014) Endosymbiotic theory for organelle origins. Curr Opin Microbiol 22:38–48. doi:10.1016/j.mib.2014.09.008

Chapter 16
Molecular Challenges to Adaptationism

Predrag Šustar and Zdenka Brzović

Abstract We examine how molecular challenges to adaptationism, as recently put into philosophical focus by Sarkar (2015), fare against different types of adaptationist theses that have been distinguished in the philosophical debate on adaptationism. Our aim is to defend a weak form of empirical adaptationism according to which a majority of phenotypic traits at the organismal level are fixed by natural selection. Sarkar (2015) indicates a possible puzzle for this view by claiming that the same arguments that challenge adaptationism at the level of the genome can potentially apply at higher levels. We argue that many of the disputes about the importance of neutrality of molecular evolution as a challenge to the adaptationist thesis stem from unprecise use of terms such as phenotypic traits or complex phenotypes. As an important step toward a solution of this problem, we propose a strategy that tracks how changes at the molecular level can cause phenotypic effects at the organismal level, and identifies as important those phenotypic changes where a new function at the organismal level was introduced. We argue that for such cases, it is justified to conclude that selection was responsible for the fixation of a new trait. While this does not yet vindicate empirical adaptationism until further empirical research is done, it does provide us with the strategy of testing it. It shows how the puzzle can be solved by pointing to cases where new and important functions at the organismal level arise and where it can be clearly shown that selection was strong enough to counteract the effects of drift.

P. Šustar · Z. Brzović
Faculty of Humanities and Social Sciences, Department of Philosophy, University of Rijeka, Rijeka, Croatia

Z. Brzović (✉)
Faculty of Humanities and Social Sciences, Department of Philosophy, University of Rijeka, Sveučilišna avenija 4, 51000 Rijeka, Croatia
e-mail: zdenka@uniri.hr

P. Pontarotti (ed.), *Evolutionary Biology*, DOI 10.1007/978-3-319-41324-2_16

16.1 Introduction

The debate on adaptationism, while prominent in biology from the late 19th century,[1] attracted philosophical interest after the famous *Spandrels* paper by Gould and Lewontin (1979). The philosophical debate on adaptationism focused primarily on conceptual and methodological issues such as how evolutionary explanations should look like; how exactly to formulate (and test) the adaptationist thesis; whether constraint hypotheses are genuine rivals to adaptive hypotheses, etc. (Orzack and Forber 2012). Also, it seems that many philosophers embraced Gould and Lewontin's view and rejected adaptationism as a problematic thesis. For instance, Dennet, in *Darwin's Dangerous Idea* (1995, p. 239), writes how the *Spandrels* paper was regarded by philosophers and other humanists as a refutation of adaptationism, while the same is not the case with the majority of biologists. Perhaps this was the reason why the challenge to adaptationism coming from molecular evolutionary biology, especially in the light of the neutral theory of molecular evolution (Kimura 1983), has not been widely addressed by philosophers.

Those philosophers of biology who did consider themselves adaptationist did not think that the data coming from evolutionary molecular biology go against their claims because they took adaptationism to be limited to non-molecular traits (Orzack and Sober 1994). However, recently Sarkar (2015) presented a challenge to adaptationism that can be extended to the non-molecular level as well. He argues that the same arguments that challenge adaptationism at the level of the genome can potentially apply at levels higher than that of genome architecture. Sarkar himself admits that this is a puzzle since there is no good reason to doubt that a significant number of phenotypic traits at the organismal level are results of selection.

In this paper, we will examine this so-called genomic challenge to adaptationism, and argue that Sarkar's argument regarding the threat to the adaptationist thesis about the phenotypical traits at the organismal level does not hold. In order to show this, we will consider different versions of adaptationist thesis that are being considered in the philosophical debate on adaptationism: empirical, explanatory, and methodological adaptationism (Godfrey-Smith 2001). On the backdrop of these views, we will defend a weak version of empirical adaptationism.

Adaptationists sometimes spell out their claim by stating that this concerns *important* phenotypic traits, or *complex* traits, thereby leaving it very vague what exactly is meant by these terms. For this reason, we propose a strategy of mapping out how the changes at the molecular level bring about changes in phenotypes at the organismal level. We argue that in those cases where the change at the molecular level brings about a new function at the organismal level, we are, prima facie, justified to conclude that selection will act (or has acted) to fix that trait (or eliminate it).

[1]Already Darwin in the later editions of the *Origin of species* discussed the relation between natural selection and other evolutionary factors.

As an informative case study, we use the phenomenon of neofunctionalization via gene duplication since it provides us with illustrative examples of how genetically redundant material fixed at the molecular level by neutral forces can serve as a basis for new functions at the organismal level that are then fixed by natural selection. In this respect, we examine three cases for which it can be established that such scenarios have occurred. These cases do not by any means provide evidence that such events do occur in majority of phenotypic traits at the organismal level, but they do provide us with a good starting point for empirically assessing adaptationist claims. In any case, we believe that they illustrate that Sarkar's (2015) claim that there is a puzzle involved in explaining how it is possible that a significant number of phenotypic traits were fixed by natural selection, is not an actual puzzle.

16.2 Challenges to Adaptationism: A Non-adaptationist Hypothesis Is a Better Explanation of Genomic and Organismal Complexity

According to the neutral theory of molecular evolution (Kimura 1983), most genetic variation at the molecular level is a consequence of mutation and random genetic drift and therefore cannot be explained by invoking natural selection. There are two basic reasons for this claim: first, the stochastic theory of population genetics which is mathematical in nature, and second, molecular genetics where molecular advances have brought more direct insight into molecular evolution (Kimura 1983). The theory asserts that only a minute fraction of DNA (or RNA) changes are adaptive. Also, most of the intraspecific variability at the molecular level, including protein and DNA polymorphism, are neutral, which means that the majority of polymorphic alleles are maintained by the balance between mutational input and random extinction (Kimura 1989). Most important evidence in support of the neutral theory of molecular evolution is the constancy of the rates of amino acid or nucleotide substitutions per site and the fact that functionally less important molecules evolve faster.

Another important set of data that support the claims of neutral theory are the data regarding the complexity of eukaryotic genomes. Eukaryotes on average have more genes, larger proteins, longer and more elaborate regulatory regions, and unique models of gene expression (Koonin 2004). Insights into eukaryotic genome architecture has, in Sarkar's (2015) words, created a puzzle since the emergence of their structural and behavioral complexity is hard to explain by invoking natural selection. Firstly, the amount of DNA in genomes of closely related eukaryotic species varies to a substantial degree. Secondly, there is no correlation between this amount and the morphological complexity of a species. Thirdly, eukaryotes contain much more DNA than seems necessary for the specification of their proteins (Gregory 2001).

Lynch (2007) argues that many aspects of genomic architecture, gene structure, and developmental pathways are difficult to explain without invoking genetic drift and mutation as the main factors. Complexity of eukaryotic genomes is hard to explain in adaptive terms because each addition to a gene increases its vulnerability to mutational inactivation which should lead to elimination under selective pressures. Furthermore, due to relatively small population sizes of multicellular species, they are expected to accumulate gratuitous gene-structural changes without any direct selection for them (Lynch 2007, pp. 8599–8600).

Koonin (2004) also talks about the puzzle that along with "useful" complex features, eukaryote genomes have accumulated many "selfish" elements which have no function for the organism containing them. He examines two potential interpretations of the evolution of biological complexity and entropy of the genomes: according to one, the increase in complexity is an adaptation, but the mechanisms leading to it are imperfect which led to entropy increase. According to the other, the increase in complexity is a by-product of entropy growth, which is a neutral process. He opts for the Lynch and Conery's (2003) theory that implies the second option by invoking the effective size of the evolving population. It is worth noting that both options might be perceived as a threat to adaptationism at the genomic level (even though the second option represents a more serious problem). From this, Koonin (2004) concludes that non-adaptationist explanation is the current null hypothesis on the origin of biological complexity.

Sarkar (2015) summarizes these conclusions in what he calls the core argument against adaptationism:

P1 The physical properties of DNA and its cellular environment lead to increased genome size and its baroque structure.
P2 Genome size is negatively correlated with population size.
P3 Selection acts against larger genomes.
P4 Small population sizes prevent the elimination of features selected against unless selection is very strong.
C Genomes increase in size, diversity, and so on and persist even though selection acts against these features. (Sarkar 2015, pp. 519–520).

Since the aim of this article is to examine neutralism as a threat to adaptationism, we will not examine alternative adaptationist explanations of genome complexity. We will take it as granted that genomic research has so far been successful in corroborating Kimura's neutral theory of molecular evolution. What remains to be seen is whether from this follows that adaptationism cannot hold even at the level of organismal phenotypes. It seems that Sarkar (2015, p. 529) would endorse this implication, since he says that the core argument can be applied beyond the level of genome architecture. However, it must be noted that according to Sarkar this possibility really presents a puzzle since he claims that there is no good reason to doubt that "a significant number of phenotypic features at the organismic level (and probably at higher levels of organization) are results of selection (…)" (Sarkar 2015, p. 529). Nevertheless, we think that the puzzle can be dispelled. In the next section, we begin to show how that could be done.

16.3 Does the Thesis of Neutrality of Molecular Evolution Threaten Adaptationism Regarding Phenotypic Traits?

Many adaptationists have discarded the neutrality of molecular evolution as a possible threat to adaptationism since adaptationism is concerned primarily with phenotypic, non-molecular traits (Orzack and Sober 1994). However, this view depends on what we take to constitute phenotypic traits. On the standard definition, phenotype comprises observable features of an organism that are based on the coding of the genotype (Rittner and McCabe 2004). On the basis of this definition, one might include molecules such as RNA and proteins into organism's phenotype. These molecules are not visible at the organismal level, but are observable and can be thus considered as parts of the phenotype. Since the neutral theory claims that most of the intraspecific variability at the molecular level (including protein and DNA polymorphism) is selectively neutral, it would follow that many phenotypic traits are neutral as well. However, this view would not capture the standard usage of the term phenotypic traits. Even Kimura (1983, pp. 55–97), the author of the neutral theory, does not consider molecular evolution to be concerned with phenotypic traits and actually marks a contrast between molecular evolutionary rates and phenotypic evolutionary rates.[2]

In order to make the adaptationist view more specific, we will follow Maeso et al. (2012) in distinguishing between different levels of biological organization. They note that phenotype should be considered as a continuum across different scales of biological complexity, but can be divided into three levels for practical reasons: organismal level (individual features such as anatomy, physiology, and behavior); cellular level (cell movements, secretory capacities, morphology, organellar composition, etc.); and molecular level (all observed traits below the cellular level: transcriptome, proteome, biochemical properties, chromatin structure, etc.).

We think that it is important to make these distinctions because in many debates the so-called adaptationists and anti-adaptationists (or pluralists) seem to argue past each other due to different uses of the term trait, phenotypic trait, complex trait, etc.[3] For instance, very often salient features such as wings, limbs, or eyes are taken as paradigmatic adaptive traits. Nevertheless, one can argue that we should consider them collections of traits, rather than as single traits. In addition, some adaptationists claim that complex traits must be a result of natural selection, but it is hard to determine how to test that claim if we take it that many factors must have been

[2]When discussing phenotypic evolution, Kimura focuses on fossil records, so phenotypic traits in this case refer to morphological characters. In addition, it can be noted that Kimura does not think that the neutral theory denies the role of natural selection in determining the course of adaptive evolution (Kimura 1989).

[3]For an informal but relevant debate on this topic, where the issues of clear definition of phenotypic traits are an important part of the discussion, see this blog: http://sandwalk.blogspot.hr/2011/02/dawkins-darwin-drift-and-neutral-theory.html.

involved in the evolution of some complex trait (which can also be taken as a set of simpler traits), some of which might have to do with purely physical laws (limitations such as Gould and Lewontin's (1979) example of body plans), some with neutral factors, and some with natural selection.

We will take it that adaptationists standardly have in mind the organismal phenotype when they claim that most phenotypic traits are adaptations (Orzack and Sober 1994; Maynard Smith 1978). This means that we will not be concerned with those authors who are adaptationist about the genomic level as well and offer adaptive stories that go against the neutral theory.[4]

We will argue that in order to identify important phenotypic traits that adaptationism refers to we need to trace how changes at the molecular level cause effects at the organismal level. In cases where this link can be clearly established and where a new function at the organismal level is introduced, it is justified to assume that selection will act to fix (or delete) such a trait. Also, there appears to be plenty of evidence that genomic complexity as a product of neutral evolution creates a precondition for strong selection pressures in evolution of phenotypic traits. In the next section, we provide grounding for this claim.

16.4 A Synergism Between Non-adaptive Evolution at the DNA Level and Adaptive Evolution at the Phenotypic Level

We follow Lynch (2007, p. 8601) in his view that there is a synergism between non-adaptive evolution at the molecular level and adaptive evolution at the phenotypic level. He argues that from the neutrality of molecular evolution does not follow that we need to abandon the view that many of the external morphological and behavioral manifestations in today's organisms are adaptive. In order to demonstrate how this synergism might work, we will examine gene duplication as one of the evolutionary neutral events that is important source of new functions that selection can act upon.

Gene duplication is a major source for genome evolution and the principal one for eukaryotes (Koonin 2011).[5] Genomic research has shown that the majority of genes in any genome belong to families of paralogs. Furthermore, it is the main

[4]For a survey and criticisms of attempts to offer adaptationist stories that would account for the architecture of the human genome and eukaryotic genomes in general see Sarkar (2015).

[5]However, it should be noted that there are many other mechanisms underlying the origins of new genes other than gene duplication. Novel genes can arise from messenger RNAs of ancestral genes, protein-coding genes metamorphosed into new RNA genes, and genomic parasites co-opted as new genes and new protein. Moreover, RNA genes can be composed from previously non-functional sequences (Kaessmann 2010). Here we take gene duplication as a good example because it is a well-studied source of new genes and potentially new functions at the level of organismal phenotype.

source of functional diversity on the level of the genotype (Lynch and Conery 2003; Ponting 2008; Conant and Wolfe 2008). This is due to the fact that after gene duplication each of the gene copies can evolve independently and acquire a functional novelty.

Ohno (1970) first proposed that the extra gene copies that are created by duplication are redundant; since the original gene already performs the necessary function, the extra copies are free from selective pressures and can acquire new mutations without being deleted by the purifying selection. This, in turn, allows them to acquire an additional function (whether it is a completely new one, or one of the functions performed by the original gene). This maintenance of duplicate genes can be a product of neutral evolution or of selection.

Several models of maintenance of gene duplicates have been proposed, most notably neofunctionalization, subfunctionalization, and increased gene-dosage advantage. In neofunctionalization model, one copy of the gene keeps the original function and is maintained by purifying selection, while the redundant copy is free to evolve and potentially acquire new function (Ohno 1970; Walsh 1995; Force et al. 1999). The copy is free to evolve because after duplication, the copy is freed from purifying selection which leads to the acceleration of evolutionary rate. In case, there are some adaptive changes in the new copy, positive selection can act to fix that changes. Thus, the neofunctionalization model invokes positive selection as a mechanism responsible for the fixation of a new function. In the subfunctionalization model, on the other hand, both duplicates accumulate mutations trough drift. Here, the original function can be divided between the two duplicates, each taking part of the original function (Force et al. 1999; Lynch and Force 2000). The two new functionally distinct copies are then preserved trough purifying selection. Another model is increased gene-dosage advantage according to which duplication is itself beneficial due to increased amount of gene product and that is the reason why both duplicates become rapidly fixed (Konrad et al. 2011)

There seems to be an agreement that certain amount of gene duplications have been fixed by positive selection (Romero and Palacios 1997; Kondrashov et al. 2002; Sebat et al. 2004; Hastings 2007). Nevertheless, the question remains whether selection plays an important role in fixation of a significant fraction of gene duplications. This is an interesting question in the debate on adaptationism at the molecular level, but it need not have a direct consequences for adaptationism about organismal phenotypic traits. While the majority of molecular changes can be a result of neutral evolution, there is still a possibility that changes that result in consequences at the level of organismal phenotype are under the strong influence of selection. We take the neofunctionalization model as possibly the most interesting case because it produces entirely new functions. What remains to be examined is whether the majority of new functions at the level of organismal phenotypes actually are fixed by selection.

We will argue that there are clear cases where we can assume an adaptationist approach to the problem of phenotypic evolution. These are the cases where it can be established that a change at the molecular level had as a consequence a change in the phenotypic level that brought about a new function. In order to demonstrate this, we will use three case studies that illustrate this kind of situation.

16.5 Neofunctionalization: Examples Where New Organismal Phenotypic Functions Were Fixed by Positive Selection

16.5.1 The Case of Color Vision

Opsins are light-sensitive proteins in the photoreceptor cells of the retina that mediate the conversion of a photon of light into electrochemical signal. Genes that contain coding sequences for opsins belong to the same gene family that accrued during our evolutionary past by the gene duplication processes and, then, by functional divergence, resulted in adaptive functional novelties. The trichromatic vision of Old World monkeys and primates represents a functional novelty conferring a high adaptive value on organisms bearing the trait of the gene family containing the corresponding gene duplicates (Golding and Dean 1998, p. 359). The adaptive value can be inferred by studying the ecological consequences of the color vision acquisition: individual organisms having the functional novelty in question could easily explore and construct completely new ecological niches, which, then, brought about a significant impact on the related ecosystem. The described case suggests an incremental construction of fit between organisms and their environment which leads us to conclude that a selective mechanism is at work (Sterelny 2006).

The described case of the evolution of trichromatic color vision represents a case where, unlike in most cases, the relationship between the genotypes and phenotypes is straightforward; a very simple change at the molecular level produces an important change at the phenotypic level for which we can be fairly certain that it brought advantage to the organisms bearing it.

Golding and Dean (1998) propose the same approach; in order to infer that an adaptive change took place, next to information on raw sequence and phylogeny, we need phenotypes. Thus, in cases where we can detect a direct link between a change at a molecular level and a change at the phenotypic level (Melin et al. 2014) and where the change at the phenotypic level brings an incremental construction of fit between organisms and their environment, we are justified to conclude that the positive Darwinian selection is at work. The combination of information on

phylogeny, structural information, and information on physiology and ecological conditions at the time of the fixation of this trait allows us to conclude that positive selection is responsible for the fixation of the trait.

16.5.2 The Case of Digestive RNASE1 Genes in Leaf-Eating Monkeys

Pancreatic RNASE genes are a particularly illustrative example of neofunctionalization in leaf-eating Colobine monkeys. Colobine monkeys have an important pancreatic enzyme RNASE1 that helps them digest bacterial ribonucleic acid (RNA), an important source of nitrogen. It was demonstrated that after the duplication of RNASE1 gene, the extra copy of the gene (called RNASE1B) mutated and acquired a new function of producing an enzyme that was more efficient in deriving nutrients from bacteria in the foregut (Zhang et al. 2002), thereby making monkeys more efficient in extracting energy from leaves. Using molecular analyses and functional assays, Zhang et al. have shown that the duplicated RNASE1 genes in two leaf-eating monkey species (Asian *Pygathrix nemaeus* and African *Colobus guereza*) evolved rapidly under positive selection for improved digestive efficiency, as a response to the increased demands for the enzyme for digesting bacterial RNA. In addition, it was shown that duplication occurred after the separation of African and Asian Colobines, which leads to conclusion that there were two separate duplication events that were followed by similar selection pressures (Zhang 2006). Thus, this is also a case where a small functional change at the molecular level following gene duplication brought about new, salient function that increased organisms' fitness.

16.5.3 Evolution of Antifreeze Protein

Due to the high salt content, the waters of Arctic and Antarctic can reach −2 °C. Fishes living in such an environment need to develop a mechanism that will enable their blood not to freeze in order to survive. Antifreeze proteins in the blood bind to ice crystals and deter the joining of additional water molecules which decreases the temperature of macroscopic ice expansion below the colligative freezing point.

It has been shown that antifreeze proteins (AFPs) arose independently in different polar marine teleost lineages due to strong selection pressures in the late Cenozoic sea-level glaciation (Deng et al. 2010). Several fish AFPs evolved by duplication from ancestral genes with different functions. For example, type III AFPs of polar

zoarcoid fishes are homologous with the small C-terminal domain of sialic acid synthase (SAS), a cytoplasmic enzyme that catalyzes intracellular synthesis of sialic acids from N-acetylmannosamine or Man-NAc-6-phosphate and phospho-enolpyruvate. Type III AFPs are secreted plasma proteins that bind to ice crystals and prevent ice growth. Deng et al. (2010) suggest that enzymatic and antifreeze functions within the same ancestral SAS molecule point to adaptive conflict that was resolved by gene duplication and neofunctionalization of the copied gene.

The three cases we considered illustrate the situation where a change at the molecular level brings forth a new, salient function for highly complex multicellular individual organisms.[6] We argue that in such cases an adaptationist approach to phenotypic evolution (at the organismal level) is, at least prima facie, justified. That is, they actually represent cases where, as required by Sarkar (2015), selection was strong enough to counteract the effects of drift that would follow as a consequence of the relatively small populations.

One might wonder whether these kinds of cases occur relatively rarely and whether one is justified in reaching an adaptationist conclusion from examining just a couple of examples. Nevertheless, we believe that this approach at least provides us with the strategy of spelling out the adaptationist thesis that can be put to empirical test. This strategy is sensitive to the hard problem (also acknowledged by Sarkar[7]) of producing precise estimates for population sizes and selection coefficients for historical populations. We agree that it is a problem to establish the action of selection with certainty, but we think that a careful examination of various factors, from the changes on the molecular level to the data on ecology, can provide us with the most reliable results.

In the philosophical debate on adaptationism, there has been many claims regarding the testability of adaptationist thesis and the specific ways of defining adaptationism. In the next section, we examine three theses of adaptationism as presented in the philosophical debate and analyze how they fare with respect to the so-called genomic challenge.

---

[6]We acknowledge the fact that in many cases where positive selection fixes traits at the molecular level, there are methods for detecting the act of selection, such as showing that nonsynonymous nucleotide substitutions exceed synonymous nucleotide substitutions during the early stages after duplication. However, we take it that this is not enough to conclusively establish adaptationism at the molecular level, and this is not the adaptationist thesis that we are concerned with. We limit adaptationism to the claim about the evolution of phenotypic traits. In addition, in order to reach the conclusion about the act of selection, further information about environmental conditions and the usefulness of the new evolved function should be taken into consideration.

[7]Sarkar (2015) addresses this issue and acknowledges that the lack of this information makes his argument "qualitative" instead of quantitative, but still not merely verbal as adaptationist "just so stories." We think that our strategy is a good starting point in avoiding the accusation of offering merely verbal accounts of evolution of phenotypic traits. This strategy is still not backed by enough quantitative data, but we take it that it represents a substantial step toward testing the adaptationist hypothesis regarding organismal phenotypic traits. In our case, this consists in tracing how the mechanisms at the molecular level cause organismal phenotypic changes.

16.6 Philosophical Debate on Adaptationism: Three Kinds of Adaptationist Views

Godfrey-Smith (2001) has identified three main types of adaptationism which became a standard reference point in the philosophical debates on the topic[8]: empirical, explanatory, and methodological adaptationism. Empirical adaptationism is the view according to which natural selection is a powerful and ubiquitous force that drives the evolutionary change. According to this view, in order to explain and predict evolutionary phenomena, it is enough to focus on the role played by natural selection and ignore other possible causal factors. This view is taken as an empirical claim about the biological world that we should be able to put to test.

Explanatory adaptationism appears to be a weaker claim that stresses the importance of selection for explaining the apparent design of organisms, which is taken to be the most important biological phenomenon. However, explanatory adaptationism leaves open the possibility that natural selection is not the most ubiquitous source of evolutionary change. On the other hand, methodological adaptationism is the claim about scientific methodology; it claims that the best approach for investigating evolutionary processes is to look for features of good adaptation and design. This methodological principle is also compatible with the claim that natural selection is not the most pervasive or even frequent force of evolutionary change.

The thesis of methodological adaptationism will not be of interest in this paper since we are concerned with the possibility of determining the role of selection in the evolution of phenotypic traits. However, methodological adaptationism is a claim that can be taken as valid regardless of the actual role that selection played in evolution. The thesis of adaptationism that is most interesting in the perspective of this paper is the claim that natural selection has been the only important cause of most of the phenotypic traits found in most species (Sober 1998, str. 72). This is standardly taken to be empirical adaptationism. In this respect, we want to make some clarifications concerning the type of empirical adaptationism that we wish to defend.

We will not enter into the debate on whether adaptationism needs to defend the optimality criterion, as proposed by Orzack and Sober's (1994) specification of adaptationist claim. According to the optimality criterion, for trait T of an individual in a given population, the claim (O) is true, where (O) states: Natural selection is a sufficient explanation of the evolution of T, and T is locally optimal. We will be concerned with the thesis regarding the ubiquity of natural selection, and not the specific thesis that relies on the claims of optimality. One of the reasons for this is the fact that the optimality criterion seems to be relative to the way we specify what it takes for something to be considered optimal. Also, the genomic challenge to

[8]There have also been proposals to divide the adaptationist views in more detail. For instance, Lewens (2009) distinguishes seven types of adaptationism (while acknowledging the main three types, which he then subdivides into distinct subtypes).

adaptationism as presented by Sarkar (2015) does not enter into the debate on optimality. Already this is enough to qualify our version of adaptationism as *weak*. However, we also further add the criterion (that is often taken for granted in the debate) that the claim about ubiquity of selection applies only to organismal phenotypic traits. Thus, weak adaptationism is a type of empirical adaptationism according to which natural selection is a ubiquitous force that operates at the level of organismal phenotypic traits.

Introduction of this middle position, we believe, clarifies some of the supposed ambiguity in views of the authors, such as Dawkins (1999) and Dennet (1995), who were identified by Godfrey-Smith (2001, pp. 339–340) as being in between empirical and explanatory adaptationism. For example, Godfrey-Smith identifies Dawkins as being explanatory adaptationist and holding a view according to which selection is rare, but occurs often enough to answer the big evolutionary questions (which is, according to Godfrey-Smith, sufficient for not including Dawkins among empirical adaptationists).

To support the claim that Dawkins is not empirical adaptationist, Godfrey-Smith cites the fact that Dawkins does not seem to have anything invested in the debate between neutralists and adaptationists. Dawkins, for instance, argues that biochemical controversy over neutralism is quite different from the adaptationist controversy since adaptationism is concerned with "(…) whether, given that we are dealing with a phenotypic effect big enough to see and ask questions about, we should assume that it is the product of natural selection." (Dawkins 1999, p. 32) He adds that biochemists' neutral mutations are not mutations at all for those who look at gross morphology, physiology, and behavior. He concludes that "If a whole-organism biologist sees a genetically determined difference among phenotypes, he already knows he cannot be dealing with neutrality in the sense of the modern controversy among biochemical geneticists." (Dawkins 1999, p. 32).

However, as suggested by Godfrey-Smith (2001, pp. 340–341), sometimes Dawkins seems to be making a more ambitious claim about the large amount of biological world that was shaped by natural selection. For instance he refers to the "sheer hugeness" of the phenomenon (Dawkins 1986, p. 15). This would appear to make his view ambiguous between explanatory and empirical adaptationism.

In our view, Dawkins can be taken to be a weak empirical adaptationist since his discussion is clearly limited to organismal phenotypic traits. We take it that this reading better captures Dawkins' claims about adaptations than the thesis of explanatory adaptationism as explained by Godfrey-Smith (2001). Due to the fact that Godfrey-Smith does not make a distinction regarding the levels at which phenotypic traits are considered, he cannot interpret Dawkins as an empirical adaptationist (since, for instance, it is clear that selection is not so ubiquitous at the genomic level). For this reason, he takes Dawkins' adaptationism to mainly consist in the claim that adaptation and apparent design are the most important biological problems worth considering. However, the problem with this view (explanatory adaptationism) is that it is not empirically testable and merely expresses a personal preference or inclination for one type of biological explanations. While it seems that Dawkins really does endorse this view on the importance of apparent design in

evolutionary biology, we believe that this view does not exhaust his adaptationism. Thus, we agree with Godfrey-Smith that explanatory adaptationism is not a testable claim, but disagree with him that Dawkins cannot be interpreted as a weak empirical adaptationist, which is a testable claim.

We take it that Dawkins' emphasis on the importance of phenotypic effects that are "big enough" and that are of interest to "whole-organism biologists" illustrates the need to introduce distinctions between different levels of phenotypic effects, and thereby to narrow the scope of the adaptationist thesis, as we have suggested. Moreover, if we consider Dawkins' view in the light of our strategy of trying to specify very closely what we mean by organismal phenotypic traits (that are products of selection), then we can see that the view does not fall prey to the accusation of untestability.

16.7 Conclusion

We have examined the so-called genomic challenge to adaptationism and argued that a weak version of empirical adaptationism is not challenged by the results from molecular evolutionary biology, as argued by Sarkar (2015). We defined weak empirical adaptationism as a view according to which majority of phenotypic changes at the organismal level were products of natural selection. Our proposal is that such a view can be tested by careful examination of molecular changes that bring about changes at the phenotypic level. For this purpose, we briefly examined three examples of cases where it can be concluded that selection fixed the phenotypic trait in question and was strong enough to counteract the effects of drift due to the small population size.

At this point, we do not wish to claim that weak empirical adaptationism is vindicated by the current data. What we contend, nevertheless, is that it is certainly a testable empirical claim. Thus, weak empirical adaptationism as a view about the power of selection in accounting for the evolution of phenotypic traits at the organismal level is not threatened by the data coming from molecular evolutionary biology.

Acknowledgments For helpful discussions of earlier drafts of the paper, we would like to thank Elliott Sober, Thomas Reydon, Marko Jurjako, Pierdaniele Giaretta, and audiences at the conferences "The 19th Evolutionary Biology Meeting in Marseille," "Model Selection: Ockham's Razor and Related Issues," Padova, and "Philosophy, Society, and the Sciences," Rijeka, which all were held in 2015. Many and special thanks also to Pierre Pontarotti. Work on this paper was partly supported by a grant from the European Social Fund for the project "Building a Support System for Young Researchers."

References

Conant GC, Wolfe KH (2008) Turning a hobby into a job: how duplicated genes find new functions. Nat Rev Genet 9:938–950
Dawkins R (1986) The blind watchmaker. Norton, New York
Dawkins R (1999) The extended phenotype. Oxford University Press, Oxford
Deng C, Deng CC, Ye H, He X, Chen L (2010) Evolution of an antifreeze protein by neofunctionalization under escape from adaptive conflict. Proc Natl Acad Sci U.S.A. 107 (50):21593–21598
Dennet D (1995) Darwin's dangerous idea. Penguin Books, London
Force A, Lynch M, Pickett F, Amores A, Yan Y, Postlethwait J (1999) Preservation of duplicate genes by complementary, degenerative mutations. Genetics 151:1531–1545
Godfrey-Smith P (2001) Three kinds of adaptationism. In: Orzack S, Sober E (eds) Adaptationism and optimality. Cambridge University Press, Cambridge, pp 335–357
Golding G, Dean A (1998) The structural basis of molecular adaptation. Mol Biol Evol 15 (4):355–369
Gould SJ, Lewontin RC (1979) The spandrels of san marco and the panglossian paradigm. Proc R Soc Lond B 205:581–598
Gregory T (2001) Coincidence, coevolution, or causation? DNA content, cell size, and the C-value enigma. Biol Rev Camb Philos Soc 76(1):65–101
Hastings JP (2007) Adaptive amplification. Crit Rev Biochem Mol Biol 42(4):271–283
Kaessmann H (2010) Origins, evolution, and phenotypic impact of new genes. Genome Res 1313–1326
Kimura M (1983) The neutral theory of molecular evolution. Cambridge University Press, Cambridge
Kimura M (1989) The neutral theory of molecular evolution and the world view of the neutralists. Genome 31(1):24–31
Kondrashov FA, Rogozin IB, Wolf YI, Koonin EV (2002) Selection in the evolution of gene duplications. Genome Biol 3
Konrad A, Teufel A, Grahnen J, Liberles D (2011) Toward a general model for the evolutionary dynamics of gene duplication. Genome Biol Evol 3:1197–1209
Koonin E (2004) A non-adaptationist perspective on evolution of genomic complexity or the continued dethroning of man. Cell Cycle 3(3):280–285
Koonin E (2011) The logic of chance: the nature and origin of biological evolution. FT Press, Upper Saddle River
Lewens T (2009) Seven types of adaptationism. Biol Philos 24:161–182
Lynch M (2007) The frailty of adaptive hypothesis for the origins of organismal complexity. Proc Natl Acad Sci U.S.A. 104:8597–8604
Lynch M, Conery J (2003) The origins of genome complexity. Science 302:1401–1404
Lynch M, Force A (2000) The probability of duplicate gene preservation by subfunctionalization. Genetics 459–473
Maeso I, Roy S, Irimia M (2012) Widespread recurrent evolution of genomic features. Genome Biol Evol 4(4):486–500
Maynard Smith J (1978) Optimization theory in evolution. Annu Rev Ecol Syst 9:31–56
Melin AD, Hiramatsu C, Parr NA, Matsushita Y, Kawamura S, Fedigan L (2014) The behavioral ecology of color vision: considering fruit conspicuity, detection distance and dietary importance. Int J Primatol 35(1):258–287
Ohno S (1970) Evolution by gene duplication. Springer, Berlin
Orzack S, Forber P (2012, February 3) Adaptationism. Retrieved from Stanford Encyclopedia of Philosophy: http://plato.stanford.edu/archives/win2012/entries/adaptationism/
Orzack S, Sober E (1994) Optimality models and the test of adaptationism. Am Nat 143 (3):361–380
Ponting C (2008) The functional repertoires of metazoan genomes. Nat Rev Genet 9:689–698

Rittner D, McCabe T (2004) Encyclopedia of biology. Facts on File Inc, New York
Romero D, Palacios R (1997) Gene amplification and genomic plasticity in prokaryotes. Annu Rev Genet 31:91–101
Sarkar S (2015) The genomic challenge to adaptationism. Br J Philos Sci 66(3):505–536
Sebat J, Lakshmi B, Troge J, Alexander J, Young J, Lundin P, Gilliam T (2004) Large-scale copy number polymorphism in the human genome. Science 305(5683):525–528
Sober E (1998) Six sayings about adaptationism. In: Hul D, Ruse M (eds) The philosophy of biology. Oxford University Press, Oxford, pp 72–86
Sterelny K (2006) Memes revisited. Br J Philos Sci 57:145–165
Walsh J (1995) How often do duplicate genes evolve new functions. Genetics 139:421–428
Zhang J (2006) Parallel adaptive origins of digestive RNases in Asian and African leaf monkeys. Nat Genet 38:819–823
Zhang J, Zhang YP, Rosenberg HF (2002) Adaptive evolution of a duplicated pancreatic ribonuclease gene in a leaf-eating monkey. Nat Genet 30:411–415

Chapter 17
Ontogeny, Oncogeny and Phylogeny: Deep Associations

Ramray Bhat and Dharma Pally

Abstract It is now commonplace to describe cancer as an evolutionary process involving the clonal expansion of malignant cells, which display heterogeneity in fitness based on inherited variation of their cellular phenotype. On the other hand, recent nuanced arguments refer to cancer cell populations as marginally evolving cell populations, not unlike their non-malignant somatic counterparts, with the resultant conclusion being that while selection might operate within a progressing cancerous mass, evolution of phenotypic complexity within them is unlikely. In this chapter, we review some past and current empirical observations, as well as the theoretical frameworks that have been constructed on their bases. We propose that a constantly changing dynamic between the invasive cancer cells and their microenvironment(s) is a relatively ignored, but imperative add-on to evolutionary developmental explanations of the disease.

17.1 Introduction

Cancer is an organism's disease that is brought about through emergent dysfunction at several levels: genes, proteins, cells, tissues and organs (Nelson and Bissell 2006; Bhat and Bissell 2014). The three overarching characteristics that can be identified among others within the primary cancer focus are alteration in tissue structure, an uncontrolled proliferation of cells, and mutations in genes. Indeed, one would be hard-pressed to find a malignant tumor that does not show evidence of all these three properties. Given the morbidity and mortality associated with the disease, clinicians and scientists have sought to investigate the etiopathology of cancer and to seek cure ever since ancient times. The interest in approaching cancer from an evolutionary point of view is in comparison with the study of the disease, a more recent project brought to the attention of the scientific community with two

R. Bhat (✉) · D. Pally
Department of Molecular Reproduction, Development and Genetics,
Indian Institute of Science, Bangalore 560012, India
e-mail: ramray@mrdg.iisc.ernet.in

P. Pontarotti (ed.), *Evolutionary Biology*, DOI 10.1007/978-3-319-41324-2_17

groundbreaking essays, one by John Cairns in Nature (Cairns 1975) and the other by Nowell in Science (1976), although one can trace an unambiguous attempt at synthesis of cellular and evolutionary thoughts in the context of cancer in a review by Theodore Hauschka titled "The chromosomes in ontogeny and oncogeny" (Hauschka 1961). Nowell's assertions on the selection of specific clonal cancer populations that have a growth advantage over other clones can be traced to extensive empirical studies conducted by Japanese cancer biologists on murine sarcomas that showed specific chromosome-level signatures are associated with cancer cells and primarily contributed to the growth of the tumor (Makino 1956). None of the mentioned workers were dyed-in-the-wool evolutionary biologists; however, their arguments were couched in a language that attracted evolutionary theorists to speculate on whether cancer cells could qualify as evolving Darwinian populations (Armitage and Doll 1957; Davies et al. 2011). We begin by asking whether neoplastic progressions could be understood from a more developmental or evolutionary perspective and whether it makes sense to distinguish between the two perspectives. We then review, in some detail, recent arguments that posit to different extents that cancer cells do evolve over time. We conclude by discussing what these arguments lack and how current theory and new empirical findings serve to advance them toward radically opposite conclusions.

17.2 A Developmental or an Evolutionary Metaphor?

The use of developmental metaphors to describe cellular phenomena in cancer has steadily gained ground with an increasing realization of the importance of the tissue microenvironment in both ontogeny and oncogeny. Bissell and Radisky posit that strong functional and interactive associations between cancer cells and their microenvironment make them a new "organ" that undergoes alterations as neoplastic progression ensues (Radisky et al. 2001). In the process, they define what organogenesis in a non-malignant developmental context is all about: the interactions between the parenchymal cells that will constitute and perform the function of the prospective organ, and their tissue microenvironment that consists of other cells such as incoming vascular and neuronal cells, fibroblasts, immune cells, pre- and – adipocytes, and the extracellular matrix (ECM), glycoproteins and proteoglycans that form the bulk of the extracellular space and constitute the basement membrane on which glandular organ epithelia are situated. Development is not merely the coming together of cells to form multicellular aggregations (although the latter is a necessary criterion for organogenesis): A series of inter-connected and inter-active processes have to occur spatiotemporally in this transformation. Below, we mention three specific processes that are exemplars of the complexity that is built into organogenesis.

(i) Pattern formation

The interaction between cells and their microenvironments during development ultimately leads to the former arranging themselves into specific deterministic arrangements. The formation of these arrangements is known as pattern formation, perhaps the most visually spectacular property of multicellular ontogeny. The mechanisms by which patterning is achieved frequently involve cellular movements [such as coordinated cellular revolutions observed in the elongation of the early Drosophila embryo (Bilder and Haigo 2012; Haigo and Bilder 2011), haptotactic movement of limb precartilage mesenchymal cells prior to chondrogenesis (Newman and Bhat 2007), spatially regulated cell movement during salivary branching (Harunaga et al. 2011) and coherent angular cell mechanics observed during acinar mammary morphogenesis in organomimetic culture contexts (Tanner et al. 2012)]. In fact, the ECM may play a passive [allowing for cells to deploy enzymes mediating cleavage of its proteolytic and glycan constituents to achieve displacement (Alcaraz et al. 2011; Gomes et al. 2015)] or active roles [the interaction between fibronectin secreted by cells and heparan sulfate on their cell surfaces allows for matrix-driven cellular translocation with collagen matrices (Newman et al. 1987), and myosin- and integrin-dependent interaction between basement membrane and outer epithelial cells drives their motility in developing salivary buds (Hsu et al. 2013)] in the process. Important mediators of cellular patterning are also morphogens, molecules which can diffuse across spaces to signal cells across spatiotemporal spectra to adopt different fates. Classical examples of morphogen networks (morphogens often interact with each other in order to mediate complex multicellular patterns) include the BMP–ADMP axis, which along with BMP signaling inhibitors, Chordin, Bambi and Sizzled, specifies the rostrocaudal axis in developing amphibian embryos (Reversade and De Robertis 2005), the BMP-Wnt-Sox9 (Sox9 is a transcription factor) substrate depletion network mediating digit patterning in developing murine embryonic limbs (Raspopovic et al. 2014), galectins and their cognate ligand-based reaction–diffusion–adhesion network that mediates early tetrapod element patterning in avian embryonic appendages (Bhat et al. 2011; Glimm et al. 2014), and FGFs secreted from the cardiac mesoderm specifying a hepatic or pulmonary fate to the ventral foregut endoderm through a threshold development mechanism (Serls et al. 2005).

A closer look at the pattern formation events occurring during the organismal development reveals several of the crucial interactions that lead to defined morphogenetic outcomes take place between molecules that have specific and non-overlapping effects on material (biochemical and biophysical) properties of biological matter (Newman and Bhat 2009; Newman et al. 2006; Forgács and Newman 2005). Their deployment individually or in interaction with each other leads to distinct patterns in accordance with the material properties they contribute to or deploy. Complex combinations of these molecules result in spatiotemporally heterogeneous phenotypes. For example, the interaction between Notch signaling (mediates periodic cell fate specification through cyclic gene expression) and FGF, a morphogen regulates the pace of somitogenesis (Gomez et al. 2008) and the

process by vertebrate segmentation takes place. Similarly, the interaction between the collagenolytic MT1-MMP and the matrix receptor β1-integrin mediates branching morphogenesis of mammary epithelia (Mori et al. 2013). Also, the stability of cell surface heparan sulfates as a function of the endoglycosidase heparanase modulates the signaling by the growth factor and morphogen FGF2 in the branching of submandibular glands (Patel et al. 2007).

This view of organismal development, which integrates the material properties of biological matter with the patterning molecules (genes and their products, proteins that are expressed within developing tissues), is complementary to the body of thought that studies organisms as "genomic computers" (Istrail et al. 2007) wherein elegant and intricately connected networks of transcription factors specify differentiation fates and development unfolds as a hierarchical cascade of progressive cell differentiations toward a terminally differentiated state (Davidson 2005). The latter view fails to capture the "vitalism" [sensu Kirschner (Kirschner et al. 2000)] of organized tissues and organs, whereas the former seeks to posit a framework by which physicochemical properties of deployed molecules (or a set of molecules) during the development seek to provide an understanding of why organization comes about, in addition to how they do so.

(ii) **Polarity**

The establishment of collective multicellular (tissue) polarity distinguishes an organ from a collection of cells. Tissue polarity has been well studied within glandular organs, but is not limited to them. Mesenchymal cells that prefigure the skeleta of vertebrate embryonic appendages show oriented polarity in parallel with the longitudinal axis of the developing skeletal elements. In fact, computational modeling implicates the requirement of polarity for limb bud outgrowth during the development of the appendage (Holmes and Trelstad 1980; Boehm et al. 2010). Within most secretory glands of mammals, both the ducts and acini exhibit radial polarity. In the center of the mammary gland is a lumen that results from the death of cells, which have lost contact with the basement membrane. The innermost cellular layer is the luminal epithelia, which are columnar in nature and show basoapical polarity with their apex pointing toward the lumen. Centrifugal to them lie the spindle-shaped myoepithelial cells that secrete the basement membrane, a superstructure consisting of non-fibrillar type IV collagen, laminins, nidogen and tenascin among other ECM glycoproteins (Engvall 1995; LeBleu et al. 2007). It is relevant to ask whether polarity is the cause or the effect of pattern formation and hence morphogenesis? Whereas a multicellular full-blown polar architecture can be considered an endpoint of extracellular patterning and intracellular signaling events, examples from across species show that tissue polarity and tissue morphogenesis could dynamically and reciprocally regulate each other.

One hotly studied but as yet poorly understood example is planar cell polarity, the alignment of cell polarity in a spatially synchronized way along the axis of the tissue plane (Newman and Bhat 2008). The mechanism of the establishment of polarity within each cell, in a manner identical to its neighbors, is beginning to be

understood and most likely involves an antipolar trafficking of two multiprotein complexes: Fz–Dsh–Fmi and Vang–Pk–Fmi (Devenport 2014). However, we still know very little as to how this event is coordinated across cells. Chemical cues such as extracellular gradients of Wnts are strong candidates and probably can act as both instructive and permissive signals, depending on the context. In addition, mechanical cue such as interstitial tension within growing organ primordia can (re) synchronize planar polarity as demonstrated by Aigouy and workers in pupal wings of Drosophila (Aigouy et al. 2010). In case of radial polarity as is seen in glandular epithelia, Tanner and coworkers used an organomimetic three-dimensional culture system to interrogate the link between radially coordinated polarity and associated morphogenetic movement. They were able to demonstrate that the coordination of polarity correlated with the coordinated angular motion of mammary epithelia and disruption of the former led to incoherent motility among cells (Tanner et al. 2012).

(iii) **Quiescence**

With the exception of tissues that have adapted to recycle at rapid rates, cells that have reached a differentiated state wherein they are involved in the physiological functions of their organ undergo reversible growth arrest for large periods of their life (Spencer et al. 2011; Lian et al. 2004). Cellular quiescence is both the consequence of, and a marker for, tissue homoeostasis. Although seen in unicellular organization where quiescence is a means to preserve genetic material and phenotype when tiding over periods of nutrient deprivation, quiescence in multicellular organisms occurs even in nutrient availability and depends on microenvironmental signals (Valcourt et al. 2012).

The mechanisms by which cellular quiescence is achieved depend on the tissue context. In an adult healthy liver, for example, 98–99 % of cells are in a quiescent phase at any given point of time (Grisham 1962). Injury and chemical insult trigger the cells to enter the cell cycle and replenish lost populations of hepatocytes. In fact, the replenishment of hepatocytes occurs to maintain an equilibrium ratio between the body mass and liver size, also known as hepatostat (Michalopoulos 2010), through mechanisms that are still not known (Berasain and Avila 2015). Recent investigations implicate Hippo pathway in regulation of the hepatocyte quiescence and liver homoeostasis (Avruch et al. 2011; Yimlamai et al. 2014). Hippo pathway regulates through serine/threonine kinases, the phosphorylation, and nuclear import of Yap1, a co-transcription factor that can induce the expression of genes encoding mitogenic signals. Unsurprisingly, the regulation of Hippo pathway can be mediated through cell–cell contact and cues from the ECM (Urtasun et al. 2011).

During mammary branching morphogenesis, cells that are part of distally extending terminal end buds are proliferative, whereas epithelia that constitute the duct that has already formed proximally with respect to the end buds are growth-arrested and quiescent (Spencer et al. 2011). The mechanism by which quiescence is maintained was shown to be dependent downstream on the levels of nuclear actin, which co-localizes with transcription foci, and upstream with the duct-specific laying down of the basement membrane that is centrifugal to the

mammary epithelia. The presence of laminin-111, a principal constituent of the basement membrane surrounding the ductal epithelia and absent around the end buds, was experimentally demonstrated to deplete actin levels within nuclei, destabilize the interactions between RNA polymerases and nuclear substructures and bring about a coordinated decrease in gene expression levels through condensation of chromatin by histone deacetylation (Le Beyec et al. 2007) and increase DNA methylation though the expression of methyl CpG-binding protein (Plachot and Lelievre 2004).

17.3 Developing Organ: Specific Yet Plastic?

Our brief description of the three salient characteristics of developing organs serves to highlight an important proposition: The progression from mere cellular populations to tissues and organs is a process that involves a gradual restriction of what could be termed "spatiotemporal randomness." This is achieved through coordination across tissue mesoscale of information using chemical and mechanical cues. The restriction is by no means passive or temporally irreversible: Indeed, the adult mammary gland having achieved a mammary arbor of quiescent epithelial cells will proliferate in response to endocrine changes during pregnancy and lactation (Chuong et al. 2014). There is emerging, although by no means conclusive evidence that the hepatocytes that replenish liver tissue loss due to injury or steatosis are derived from the existing, predominantly quiescent hepatocyte populations (Tarlow et al. 2014).

Developing tissues and developed organs are therefore, specific yet plastic. Specificity allows organs to acquire their innate pattern, architecture and hence distinct function(s). Organs as part of established patterns also become more insulated to the organismal macroenvironment. The integration of the microenvironment with the organ, however, ensures its continued involvement in their homeostasis and implies that organs remain sensitive to agencies within the organism, such as endocrine and exocrine cues. We feel obliged to point out that this plasticity is distinct from cell fate specification which lies on a spectrum of cellular differentiation (Balazsi et al. 2011). The plasticity we refer to, as well as specificity, is mesoscale properties of tissues and organs and pertains to overall switches in tissue and organ phenotypes and therefore cannot be reduced to changes in differentiation fates of individual cells.

17.4 Cancers: Plastic Yet Specific?

The conventional understanding of cancers as a clonal expansion of cells with genetic mutations can reconcile with the fact that cancer cells have heterogeneous phenotypes. In fact, tumor heterogeneity is an intensely studied field in itself,

although the heterogeneity is attributed solely to the oncogenic genetic mutations which accumulate within, but differ between, cells within a cancerous mass. This attribution by itself is inadequate and can be challenged by two independent lines of empirically grounded arguments: First, perturbation in levels of matricellular modulators such as the matrix metallorproteinase-3 (MMP-3) in normal mammary epithelial cells can lead to their malignant transformation through a ROS-dependent upregulation of genomic instability (Radisky et al. 2005). Secondly, recent work using ultradeep sequencing shows a heavy burden of cancer-causing mutations in about a quarter of eyelid epidermal epithelia exposed to ultraviolet radiation in light, but with no resultant sign of transformation or tumorigenesis (Martincorena et al. 2015). Therefore, the microenvironment can act both to suppress mutations and to even drive them in order to give rise to cancer (Bissell and Hines 2011). In addition, the occurrence of mutations does not ensure tumorigenesis. Given that cancer cells are also responsive and interactive with their microenvironment, tumors have been considered to be "organs" in their own right. This consideration is not new and is found in Hauschka's treatise on chromosomal changes in cancer (Hauschka 1961).

We can put the assertion of tumors as organs to test by probing whether there is experimental evidence for the three characteristics that we elaborated above were central to organogenesis. In other words, can we evince signatures of pattern formation, multicellularly synchronized polarity and/or context-dependent quiescence?

We begin with the first feature: pattern formation. Whereas one would be hard-pressed to witness the visually spectacular multicellular patterns one sees during ontogeny, in cancers probably due to a temporally florid dynamics in the life history of tumors (Nik-Zainal et al. 2012b), a basic minimum requisite for cellular patterns to emerge is for cells to come together and to act as a stable collective through an upregulation in cell–cell and cell–matrix adhesion (Forgács and Newman 2005). In fact, the mere emergence of thresholds in adhesivity can lead to sorting of cell populations, a phenomena studied under the aegis of differential adhesion hypothesis by Steinberg and coworkers (Foty and Steinberg 2013; Duguay et al. 2003), although it had been demonstrated much earlier by Holtfreter (Steinberg and Gilbert 2004). Cell adhesion is therefore essential for subsequent division of labor and spatiotemporally heterogeneous adoption of differentiation fates. There is strong evidence for very specific cell aggregative events that occur and are vital for the progression of malignant tumors. We will discuss one specific example here.

Our example relates to the formation of multicellular clusters of cancer cells that are disseminated into circulative environments such as peritoneal cavity in case of cancers of epithelial ovarian cancers, colon and hepatocellular carcinoma (Sangisetty and Miner 2012), and vasculature in case of cancers of breast, lung and pancreas (Ting et al. 2014). Elegant experiments by Aceto and coworkers show that circulating tumor clusters from mice injected with human breast cancer cells are oligoclonal clusters that are not the product of intravascular aggregative or proliferative events but are likely disintegrated chunks of the primary focal cancer (Aceto et al. 2014). Cheung and coworkers show that an important determinant of

metastasis is polyclonal mammary cancer cell clusters with upregulated desmosomal and hemidesmosomal adhesion complex genes implying the role of cell–cell and cell–matrix adhesion in metastatic progression of cancer disease (Cheung et al. 2016). On the other hand, "spheroids" multicellular clusters of cells composed predominantly of cancer epithelia that are found within malignant ascites (edematous fluid that accumulates within the peritoneal cavity of individuals afflicted with cancer of ovaries, colon and liver) have been shown to form through coalescence of nonadherent cells in situ (Latifi et al. 2012). Both breast CTC clusters and ovarian spheroids have been shown to be more prone to metastasis than their single-cell counterparts. Little is known whether there are specific cell arrangements inside the tumor cell clusters seen either in the vasculature or in malignant ascites. Detailed cytopathological and morphological studies conducted in the future will reveal whether multiple populations of cells with distinct adhesivities coexist within such aggregations.

To the best of our knowledge, there is no published information as to whether cancer cells regain polarity within multicellular clusters in vivo. Examination of the morphologies of a wide panel of breast cancer cell lines cultivated within laminin-rich ECMs reveals the absence of a polar organization of cells within multicellular clusters (Kenny et al. 2007). The spheroids generated by culturing ovarian cancer epithelial cell lines are geometrically more uniform than their primary counterparts, probably because the latter subsume other cells such as those from mesothelial or hematopoietic lineages. There is however no evidence of any polarity within these clusters. Single-cell transcriptomic screen studies show high expression of ECM protein-encoding genes by pancreatic CTCs, although the proteins they code such as osteonectin, decorin, TIMP2 and IGFBP5 are more localized to the stroma rather than basement membranes (Ting et al. 2014). Culture studies show a partial restoration of basoapical polarity in breast cancer cells cultured within laminin-rich microenvironments upon correction of aberrant signaling within them (Weaver et al. 1997). Not just that, the abovementioned intervention leads to arrest of proliferation and context-dependent quiescence in breast cancer cells, which can again be reverted upon putting them on 2D plastic substrata.

Contextual quiescence of cancer cells is relevant to the time interval between the diagnosis and treatment of cancer and its recurrence. The phenomena by which cancer cells disseminate to distant sites of prospective colonization but remain quiescent for extended periods of time is known as dormancy, the mechanisms of which are yet to be completely understood, although interactions between urokinase plasminogen activator receptor (uPAR) and α5β1 have been shown to play a role (Aguirre Ghiso et al. 1999; Aguirre-Ghiso 2007). Using an elegant coculture system mimicking a pulmonary microvascular colonization niche, Ghajar and workers have shown quiescence in cancer epithelia can be induced by thombospondin-1 from endothelia is sensitive to its contextual location with respect to vascular geometries. Cancer cells close to sprouting microvascular tips recovered their ability to grow in an unregulated manner, whereas those proximal to stable vasculatures remained quiescent (Ghajar et al. 2013).

Our arguments here reinforce the point that cancers, like organs, retain a modicum of organization while remaining sensitive to cues from their microenvironment. They are therefore just like organs, plastic and specific. In the following section, we change track and move from a developmental to an evolutionary narrative to examine how cancer cells are (dis)similar to non-malignant populations of cells in the ability of their phenotype to evolve.

17.5 Evolvability of Cancer Cells

In 2013, Pierre-Luc St Germain wrote an intensive critique of the oft-asserted statement by cell biologists that cancer cells are a population of cells undergoing the process of somatic evolution: In other words, the alteration in behavior and phenotype of cancer cells is the culmination of a temporal series of genetic changes with each change serving to marginally improve the fitness of the phenotype (Germain 2012; Merlo et al. 2006; Pepper et al. 2009).

In the following section, we will summarize the arguments of St Germain that pertain to the evolution (or lack of it) of cancer cells. The bulwark of St Germain's thesis is a framework by Peter Godfrey-Smith for phenotypic evolution and particularly the acquisition of a complex adaptive phenotype (Godfrey-Smith 2009). His framework consists of a set of relatively autonomous criteria for evolution of phenotypic complexity within evolving populations of entities using the principles encapsulated within, or compatible with, the modern synthesis. Godfrey-Smith distinguishes between a *minimal Darwinian population*, "a collection of causally connected individual things in which there is variation in character, which leads to differences in reproductive output (differences in how much or how quickly individuals reproduce), and which is inherited to some extent," and a *paradigm population*, which shows all the characteristics of minimal populations with the added feature that significant novelty emerges within such populations. They are in a broad sense evolvable [sensu Kirschner and Gerhart (Kirschner and Gerhart 1998)]. What are these criteria that according to Godfrey-Smith transform minimal Darwinian populations into paradigm ones? Godfrey-Smith describes six of these, although by his own admission, the list is by no means an exhaustive one:

The first three criteria are **fidelity of heredity**, **smoothness of the fitness landscape and dependence of the phenotype on intrinsic properties**. By fidelity of heredity, Godfrey-Smith means the overall ability of a cellular or tissue phenotype to be inherited by its progenies. Fidelity of heredity is one of the prerequisites for phenotypic evolution. Cancer cells accumulate not just mutations but are intrinsically genomically unstable. However, the instability of their genotype does not automatically correspond to phenotypic instability. The investigations on normal skin cells accumulating mutations due to ultraviolet radiation while remaining non-malignant confirm the inadequacy of this mechanism to automatically translate into cellular phenotypic instability (Martincorena et al. 2015). In fact, tracing the early life history of breast tumors, Nik-Zainal and coworkers observe a punctuated

rate of tumor progression wherein breast cancer cells acquire driver mutations that presumably give rise to genomic instability. This is followed by accumulation of a large number of closely localized mutations or more catastrophic chromosome-level rearrangements and truncations, known as chromothripsis (Roberts et al. 2012). However, a full-blown alteration in the tumor phenotype occurs not before a rate-limiting step, which is separated from the hypermutations that accumulate within "relatively long-lived quiescent populations" (Setlur and Lee 2012). Therefore, there is little evidence to prove that the fidelity of heredity is lost in cancer cells. In fact, Huang and coworkers have likened cancerous states of cells to specific attractors within their own right, metaphorically underscoring the fact that cancer phenotypes possess a certain degree of robustness (Huang et al. 2009).

Godfrey-Smith likens the fitness space to landscapes, positing that if "mountains" correspond to phase sub-spaces representing high fitness, whereas "valleys" correspond to phase sub-spaces of low fitness. A ''smooth'' landscape is when change in phenotypic traits does not result in a traversal on the landscape with sudden altitudinal shifts. In other words, a small change in phenotypic traits leads to small changes in fitness, and the fitness landscape can be considered to be smooth. Smooth landscapes therefore allow for explorations of higher peaks without falling into valleys of low fitness, whereas rugged landscapes discourage exploration. In our opinion, this feature is related to the third property, i.e., dependence of the differences in fitness on intrinsic characters. The closer the integration between intrinsic character and their insulation from "extrinsic characters," the smoother should be the landscape. Examined in the context of cellular populations, this property implies a highly specific and non-plastic state. Godfrey-Smith opines that a certain degree of insensitivity to the externa is necessary for inheritance of fitness enhancing variation which is genetic in origin. Using the example of location, Godfrey-Smith posits that physical space can be "inherited" in the sense an offspring inherits its niche from its parent, but "if extrinsic features are most of what matters to realized fitness—if intrinsic character is not very important—then other than this physical wandering, not much can happen." Rather "mutation and recombination enable a population to 'search' (more metaphysically) a space of possible genetic and phenotypic properties." This search through mutation and recombination occurs on a smooth landscape and ensures that fitness differences occur in a smooth fashion.

Here, St Germain identifies the first limitation of cancer cells with this requisite for phenotypic evolution. St Germain bases his narrative on the empirical observations of Nik-Zainal et al. (2012a, b) to mark the starting point of his thesis: that it is indubitable that there is heterogeneity in genotypic characteristics within cancer cell populations leading to coexistence of different cellular clonal subpopulations, differential persistence among them leading to the extinction of some clones and survival and dominance of one clonal population.

Although indicative of natural selection occurring, St Germain posits that the narrative falls short by the framework of Godfrey-Smith in demarcating the necessary and sufficient conditions for the evolution of complexity. He invokes the systems' theoretic elucidation of cancer by Huang that proposes cancer cell states as

attractors, similar to differentiation fates of cells in development (Huang et al. 2009). However, cancer cell attractors have smaller basins of attraction since they are not as fixed as long-evolved differentiation states and may represent exapted states within the cancer cell phenospace (Huang 2011; Vrba et al. 2005). Kaneko also talks of the latter as metastable attractors in his analysis of cancer cell progenitors (Kaneko 2011).

According to Godfrey-Smith, the dependence of normal cells on their microenvironment is responsible for their "deDarwinization," a process by which evolution is prevented within a physiologically functioning organismal soma. This is relaxed in case of cancer cells, whose behavior is guided by the mutations in their genes. St Germain departs from Godfrey-Smith's thesis by showcasing the plastic nature of cancer cells in a manner similar to what we have described above. Their dependence on, and interaction with, their microenvironment and the rugged nature of their fitness landscape make cancer populations marginally evolving populations in the opinion of St Germain.

Three other features that are argued to be essential for evolutionary innovations of novelty are the presence of **bottlenecks**, **integration** and **specialization**. Bottlenecks are contractions in population size between two generations. Specialization refers to the germ/soma binary, the facts that most entities within their population could be dead ends. Taken together, both these relate to the introduction of a stochastic agency to phenotypic evolution: Bottlenecks represent temporal stochasticity, and specializations represent spatial stochasticity and stratification. St Germain does not discuss bottlenecks and confines his arguments on specialization to the discussion of the cancer stem cells. Despite the fact that cancer stem cells could represent a specialized cellular niche that can be deployed for long-term reproduction of the cancer population, St Germain states that they still remain an "idealization" with indefinite estimates as to their frequency within cancer populations.

Integration is according to Godfrey-Smith an abstraction that sums up "the extent of division of labor, the mutual dependence (loss of autonomy) of parts and the maintenance of a boundary between a collective and what is outside it" (Anderson and McShea 2001). While Godfrey-Smith requires integration to be a necessary step for phenotypic evolution, St Germain once again departs from this opinion citing opinions from widely divergent schools of cancer theory to make the point that cancer cells are integrated better than what is generally thought and that an inordinately high level of integration could impair evolution (Radisky et al. 2001; Hanahan and Weinberg 2011, 2000; Sonnenschein and Soto 2008).

17.6 Taking Apart and Putting Together

We believe that a crucial agency that is missing from St Germain's critique and gets only a passing mention in Godfrey-Smith's narrative of the mechanism(s) of acquisition of phenotypic novelty is phenotypic plasticity. Godfrey-Smith considers it in his expansion of the relative importance of intrinsic and extrinsic properties in

determining phenotypes and their fitness distinguishing between "obvious" (ecological) and "unobvious" ways through which environment can induce change in phenotype by altering the reaction norm of trait expression.

Without mentioning it in explicit terms, St Germain appreciates the role of specificity of cancer "tissues" in his analysis of Godfrey-Smith's criteria of integration and dependence on intrinsic character, but what is overlooked is the significant role of plasticity in the evolution of phenotype. The reason perhaps closely related is the idea that if plasticity was to play a role in the acquisition of novelty, the fitness landscape need not be smooth, since a plastic phenotype automatically implies complex, dynamic and multilevel relationships between different phenotypic traits, such as seen in the examples of phenotypic accommodation, a process in which the change in expression of a phenotypic trait is accompanied within the same generation by changes in phenotypes of several other traits in accommodating the former (West-Eberhard 1989, 2003).

The idea of "intrinsicality" of cellular properties that do not depend on extrinsic factors is in our opinion a problematic one. No two spatial points surrounding cells are exactly alike. It is also not unreasonable to suppose that even two neighboring isogenic cells will not be intrinsically alike. Indeed, in order to elaborate on the importance of intrinsic properties, Godfrey-Smith uses the example of the germ–soma divide. Boulanger and coworkers extracted murine testicular cells from seminiferous tubules of adult mice and grafted them with limiting dilutions of mammary epithelial cells into cleared fat pads of transgenic mice that were engineered to express LacZ under the control of the promoter of the gene encoding whey acidic protein (WAP). They found that the seminiferous tubule cells destined for spermatogenic fate had differentiated into mammary epithelial progenitor cells with the capacity to differentiate into functional mammary epithelia (Boulanger et al. 2007). In other words, plasticity subsumes any contribution to the probability of phenotypic evolution, by the dependence of the fitness of cellular populations on their intrinsic character.

Whereas we do agree with Godfrey-Smith that bottlenecks and specialization support the evolution of novelty, we disagree with St Germain's critique that at least the latter may not be relevant to cancer populations. Indeed, the propensity of high periplakin expressing cells to be preferentially incorporated into circulating tumor cell clusters that depart from the primary foci and show a higher propensity for metastasis is an example of 'bottleneckishness' (Aceto et al. 2014). Marusyk and coworkers have shown altruistic behaviors in subpopulations within growing tumor masses exemplifying specialization essential for phenotypic evolution of cancer (Marusyk et al. 2014).

17.7 Plasticity and Evolution

Plasticity, the sensitivity of the phenotype to environmental fluctuations, or developmental reaction norm, the variance in phenotype by a single genotype under environmental fluctuations, is an incontrovertible concept in development (Newman

et al. 2009; Nanjundiah and Newman 2009). However, its role and/or relevance in phenotypic evolution has been argued especially with respect to the temporality of its deployment: whether plasticity serves to alter the behavior of an organism in a beneficial way after being selected for through genetic variation, or whether it can give rise to novelty as a result of an environmental input and prefigure the genetic change that fixes it in the organisms evolutionary history (Müller et al. 2010; Muller 2007; Laland et al. 2014; Wada and Sewall 2014; Moczek et al. 2011). West-Eberhard advocates a strong version of the effect of environmental contribution to novelty through the process of genetic accommodation, wherein "adaptive evolution is cross-generational change in phenotype frequency, accompanied by change in frequency of expressed phenotypically influential genetic alleles under selection that maximizes the positive fitness effects of the phenotypic traits whose development is influenced by these alleles" (West-Eberhard 1989, 2003; Crispo 2007). Although not known how frequently it occurs in nature, accommodation of a plastic response sensitive to environmental induction has been unambiguously shown to occur (Suzuki and Nijhout 2006; Braendle and Flatt 2006). Our intention of emphasizing the viability of genetic accommodation is to point out that there exist mechanisms by which extrinsic cues can be integrated into the "intrinsic" genetic repertoire of cells. In fact, West-Eberhard argues that environmental induction is superior to mutational induction in its ability to innovate novelty, since it affects a greater proportion of individuals in a population and it samples a greater spectrum of genetic backgrounds. This would presuppose that the population be integrated (sensu Godfrey-Smith) or specific and at the same time plastic. The environmental effect on cellular populations can also be non-adaptive, directly affecting changes in morphologies through alteration in the material properties of excitable and soft biological matter (Mikhailov 1990; Ermakova et al. 1986; Forgács and Newman 2005).

Kaneko has gone further to propose a quantitative relationship between phenotypic fluctuations due to developmental dynamics (plasticity), phenotypic fluctuations due to genetic variation (evolvability) and speed of evolution (Kaneko 2009). Analyzing results from in vivo experiments (Sato et al. 2003), and simulations of catalytic reaction networks (Kaneko and Furusawa 2006) and gene regulatory networks (Kaneko 2007, 2008), Kaneko's group observes a positive correlation between phenotypic plasticity and evolvability as well as speed of evolution. In other words, susceptibility to developmental dynamics (plasticity in phenotype as a result of changes in environment or stochasticity) promotes phenotypic sensitivity to mutations as well as the rate at which the phenotype evolves.

How do these findings bear upon somatic evolution of cancer cells? It is conceivable, if the predictions of Kaneko and coworkers hold true for cancer cells that a correlation exists between their evolvability, evolution and phenotypic plasticity. This could explain how tissue architecture serves to damp down gene expression through epigenetic means, impairing developmental noise in the process, and its loss induces genomic instability and tumorigenesis. The minimal specificity of a cancer population potentially ensures that the plasticity induced by shifts in the environment or loss of tissue architecture can be genetically accommodated.

17.8 Conclusion

Revisiting their earlier review on the hallmarks of neoplastic progression in 2011, Hanahan and Weinberg pay particular attention to the role of the tumor microenvironment: collection of non-malignant stromal cells that are recruited into a signaling dialectic with the malignant cells in order for the cancer cells to survive and propagate (Hanahan and Weinberg 2011).

We wish to develop further on this thesis by proposing that a tumor is a "neo-organ" that acquires a greater deal of plasticity and a limited cache of specificity. The latter ensures that it is evolvable even within the relatively short period of time it remains within the host's body and the specificity at best ensures and, at the least, permits that environmentally induced plastic changes in phenotype spread through the population of cells. Loss of the original tissue architecture, stratification into subpopulations with distinct phenotypic traits, fidelity of heredity and movement to new microenvironments are events through which the cancer evolves to acquire and fix new phenotypes in its evolutionary developmental trajectory. Further studies that target the cause of phenotypic plasticity within cancer cells are likely to reveal further clues to how they can be therapeutically targeted.

Acknowledgements We thank the Indian Institute of Science seed grant for support. R.B. would like to thank Pierre Pontarotti for inviting him to the 19th Evolutionary Biology Meeting at Marseille where this work was first conceived as a result of a conversation with Predrag Šustar.

References

Aceto N, Bardia A, Miyamoto DT, Donaldson MC, Wittner BS, Spencer JA, Yu M, Pely A, Engstrom A, Zhu H, Brannigan BW, Kapur R, Stott SL, Shioda T, Ramaswamy S, Ting DT, Lin CP, Toner M, Haber DA, Maheswaran S (2014) Circulating tumor cell clusters are oligoclonal precursors of breast cancer metastasis. Cell 158(5):1110–1122. doi:10.1016/j.cell.2014.07.013

Aguirre Ghiso JA, Kovalski K, Ossowski L (1999) Tumor dormancy induced by downregulation of urokinase receptor in human carcinoma involves integrin and MAPK signaling. J cell Biol 147(1):89–104

Aguirre-Ghiso JA (2007) Models, mechanisms and clinical evidence for cancer dormancy. Nat Rev Cancer 7(11):834–846. doi:10.1038/nrc2256

Aigouy B, Farhadifar R, Staple DB, Sagner A, Roper JC, Julicher F, Eaton S (2010) Cell flow reorients the axis of planar polarity in the wing epithelium of Drosophila. Cell 142 (5):773–786. doi: 10.1016/j.cell.2010.07.042

Alcaraz J, Mori H, Ghajar CM, Brownfield D, Galgoczy R, Bissell MJ (2011) Collective epithelial cell invasion overcomes mechanical barriers of collagenous extracellular matrix by a narrow tube-like geometry and MMP14-dependent local softening. Integr Biol Quant Biosci Nano Macro 3(12):1153–1166. doi:10.1039/c1ib00073j

Anderson C, McShea DW (2001) Individual versus social complexity, with particular reference to ant colonies. Biol Rev Camb Philos Soc 76(2):211–237

Armitage P, Doll R (1957) A two-stage theory of carcinogenesis in relation to the age distribution of human cancer. Br J Cancer 11(2):161–169

Avruch J, Zhou D, Fitamant J, Bardeesy N (2011) Mst1/2 signalling to yap: gatekeeper for liver size and tumour development. Br J Cancer 104(1):24–32. doi:10.1038/sj.bjc.6606011

Balazsi G, van Oudenaarden A, Collins JJ (2011) Cellular decision making and biological noise: from microbes to mammals. Cell 144(6):910–925. doi:10.1016/j.cell.2011.01.030

Berasain C, Avila MA (2015) Regulation of hepatocyte identity and quiescence. Cellular and molecular life sciences: CMLS 72(20):3831–3851. doi:10.1007/s00018-015-1970-7

Bhat R, Bissell MJ (2014) Of plasticity and specificity: dialectics of the microenvironment and macroenvironment and the organ phenotype. Wiley Interdisc Rev Dev Biol 3(2):147–163

Bhat R, Lerea KM, Peng H, Kaltner H, Gabius HJ, Newman SA (2011) A regulatory network of two galectins mediates the earliest steps of avian limb skeletal morphogenesis. BMC Dev Biol 11:6. doi:10.1186/1471-213X-11-6

Bilder D, Haigo SL (2012) Expanding the morphogenetic repertoire: perspectives from the Drosophila egg. Dev Cell 22(1):12–23. doi:10.1016/j.devcel.2011.12.003

Bissell MJ, Hines WC (2011) Why don't we get more cancer? a proposed role of the microenvironment in restraining cancer progression. Nat Med 17(3):320–329. doi:10.1038/nm.2328

Boehm B, Westerberg H, Lesnicar-Pucko G, Raja S, Rautschka M, Cotterell J, Swoger J, Sharpe J (2010) The role of spatially controlled cell proliferation in limb bud morphogenesis. PLoS Biol 8(7):e1000420. doi:10.1371/journal.pbio.1000420

Boulanger CA, Mack DL, Booth BW, Smith GH (2007) Interaction with the mammary microenvironment redirects spermatogenic cell fate in vivo. Proc Natl Acad Sci U.S.A. 104(10):3871–3876. doi:10.1073/pnas.0611637104

Braendle C, Flatt T (2006) A role for genetic accommodation in evolution? BioEssays News Rev Mol Cell Dev Biol 28(9):868–873. doi:10.1002/bies.20456

Cairns J (1975) Mutation selection and the natural history of cancer. Nature 255(5505):197–200

Cheung KJ, Padmanaban V, Silvestri V, Schipper K, Cohen JD, Fairchild AN, Gorin MA, Verdone JE, Pienta KJ, Bader JS, Ewald AJ (2016) Polyclonal breast cancer metastases arise from collective dissemination of keratin 14-expressing tumor cell clusters. Proc Natl Acad Sci U.S.A. 113(7):E854–863. doi:10.1073/pnas.1508541113

Chuong CM, Bhat R, Widelitz RB, Bissell MJ (2014) SnapShot: branching morphogenesis. Cell 158(5):1212–1212. doi:10.1016/j.cell.2014.08.019

Crispo E (2007) The Baldwin effect and genetic assimilation: revisiting two mechanisms of evolutionary change mediated by phenotypic plasticity. Evol Int J Org Evol 61(11):2469–2479. doi:10.1111/j.1558-5646.2007.00203.x

Davidson EH (2005) The regulatory genome: gene regulatory networks in development and evolution, New edn. Academic, Oxford

Davies PC, Demetrius L, Tuszynski JA (2011) Cancer as a dynamical phase transition. Theor Biol Med Model 8:30. doi:10.1186/1742-4682-8-30

Devenport D (2014) The cell biology of planar cell polarity. J cell Biol 207(2):171–179. doi:10.1083/jcb.201408039

Duguay D, Foty RA, Steinberg MS (2003) Cadherin-mediated cell adhesion and tissue segregation: qualitative and quantitative determinants. Dev Biol 253(2):309–323

Engvall E (1995) Structure and function of basement membranes. Int J Dev Biol 39(5):781–787

Ermakova EA, Krinsky VI, Panfilov AV, Pertsov AM (1986) Interaction between spiral and flat periodic autowaves in an active medium. Biofizika 31(2):318–323

Forgács G, Newman S (2005) Biological physics of the developing embryo. Cambridge University Press, Cambridge, New York

Foty RA, Steinberg MS (2013) Differential adhesion in model systems. Wiley Interdisc Rev Dev Biol 2(5):631–645. doi:10.1002/wdev.104

Germain PL (2012) Cancer cells and adaptive explanations. Biol Philos 27(6):785–810. doi:10.1007/s10539-012-9334-2

Ghajar CM, Peinado H, Mori H, Matei IR, Evason KJ, Brazier H, Almeida D, Koller A, Hajjar KA, Stainier DY, Chen EI, Lyden D, Bissell MJ (2013) The perivascular niche regulates breast tumour dormancy. Nat Cell Biol 15(7):807–817. doi:10.1038/ncb2767

Glimm T, Bhat R, Newman SA (2014) Modeling the morphodynamic galectin patterning network of the developing avian limb skeleton. J Theor Biol 346:86–108. doi:10.1016/j.jtbi.2013.12.004

Godfrey-Smith P (2009) Darwinian populations and natural selection. OUP Oxford

Gomes AM, Bhat R, Correia AL, Mott JD, Ilan N, Vlodavsky I, Pavao MS, Bissell M (2015) Mammary branching morphogenesis requires reciprocal signaling by heparanase and MMP-14. J Cell Biochem 116(8):1668–1679. doi:10.1002/jcb.25127

Gomez C, Ozbudak EM, Wunderlich J, Baumann D, Lewis J, Pourquie O (2008) Control of segment number in vertebrate embryos. Nature 454(7202):335–339. doi:10.1038/nature07020

Grisham JW (1962) A morphologic study of deoxyribonucleic acid synthesis and cell proliferation in regenerating rat liver; autoradiography with thymidine-H3. Cancer Res 22:842–849

Haigo SL, Bilder D (2011) Global tissue revolutions in a morphogenetic movement controlling elongation. Science 331(6020):1071–1074. doi:10.1126/science.1199424

Hanahan D, Weinberg RA (2000) The hallmarks of cancer. Cell 100(1):57–70

Hanahan D, Weinberg RA (2011) Hallmarks of cancer: the next generation. Cell 144(5):646–674. doi:10.1016/j.cell.2011.02.013

Harunaga J, Hsu JC, Yamada KM (2011) Dynamics of salivary gland morphogenesis. J Dent Res 90(9):1070–1077. doi:10.1177/0022034511405330

Hauschka TS (1961) The chromosomes in ontogeny and oncogeny. Cancer Res 21:957–974

Holmes LB, Trelstad RL (1980) Cell polarity in precartilage mouse limb mesenchyme cells. Dev Biol 78(2):511–520

Hsu JC, Koo H, Harunaga JS, Matsumoto K, Doyle AD, Yamada KM (2013) Region-specific epithelial cell dynamics during branching morphogenesis. Dev Dyn Official Publ Am Assoc Anatomists 242(9):1066–1077. doi:10.1002/dvdy.24000

Huang S (2011) On the intrinsic inevitability of cancer: from foetal to fatal attraction. Semin Cancer Biol 21(3):183–199. doi:10.1016/j.semcancer.2011.05.003

Huang S, Ernberg I, Kauffman S (2009) Cancer attractors: a systems view of tumors from a gene network dynamics and developmental perspective. Semin cell Dev Biol 20(7):869–876. doi:10.1016/j.semcdb.2009.07.003 (S1084-9521(09)00149-9 [pii])

Istrail S, De-Leon SB, Davidson EH (2007) The regulatory genome and the computer. Dev Biol 310(2):187–195. doi:10.1016/j.ydbio.2007.08.009

Kaneko K (2007) Evolution of robustness to noise and mutation in gene expression dynamics. PLoS ONE 2(5):e434. doi:10.1371/journal.pone.0000434

Kaneko K (2008) Shaping robust system through evolution. Chaos 18(2):026112. doi:10.1063/1.2912458

Kaneko K (2009) Relationship among phenotypic plasticity, phenotypic fluctuations, robustness, and evolvability; Waddington's legacy revisited under the spirit of Einstein. J Biosci 34(4):529–542

Kaneko K (2011) Characterization of stem cells and cancer cells on the basis of gene expression profile stability, plasticity, and robustness: dynamical systems theory of gene expressions under cell-cell interaction explains mutational robustness of differentiated cells and suggests how cancer cells emerge. BioEssays News Rev Mol Cell Dev Biol 33(6):403–413. doi:10.1002/bies.201000153

Kaneko K, Furusawa C (2006) An evolutionary relationship between genetic variation and phenotypic fluctuation. J Theor Biol 240(1):78–86. doi:10.1016/j.jtbi.2005.08.029

Kenny PA, Lee GY, Myers CA, Neve RM, Semeiks JR, Spellman PT, Lorenz K, Lee EH, Barcellos-Hoff MH, Petersen OW, Gray JW, Bissell MJ (2007) The morphologies of breast cancer cell lines in three-dimensional assays correlate with their profiles of gene expression. Mol Oncol 1(1):84–96. doi:10.1016/j.molonc.2007.02.004

Kirschner M, Gerhart J (1998) Evolvability. Proc Natl Acad Sci U.S.A. 95(15):8420–8427

Kirschner M, Gerhart J, Mitchison T (2000) Molecular "vitalism". Cell 100(1):79–88

Laland K, Uller T, Feldman M, Sterelny K, Muller GB, Moczek A, Jablonka E, Odling-Smee J, Wray GA, Hoekstra HE, Futuyma DJ, Lenski RE, Mackay TF, Schluter D, Strassmann JE

(2014) Does evolutionary theory need a rethink? Nature 514(7521):161–164. doi:10.1038/514161a

Latifi A, Luwor RB, Bilandzic M, Nazaretian S, Stenvers K, Pyman J, Zhu H, Thompson EW, Quinn MA, Findlay JK, Ahmed N (2012) Isolation and characterization of tumor cells from the ascites of ovarian cancer patients: molecular phenotype of chemoresistant ovarian tumors. PLoS ONE 7(10):e46858. doi:10.1371/journal.pone.0046858

Le Beyec J, Xu R, Lee SY, Nelson CM, Rizki A, Alcaraz J, Bissell MJ (2007) Cell shape regulates global histone acetylation in human mammary epithelial cells. Exp Cell Res 313(14):3066–3075. doi:10.1016/j.yexcr.2007.04.022

LeBleu VS, Macdonald B, Kalluri R (2007) Structure and function of basement membranes. Exp Biol Med 232(9):1121–1129. doi:10.3181/0703-MR-72

Lian JB, Javed A, Zaidi SK, Lengner C, Montecino M, van Wijnen AJ, Stein JL, Stein GS (2004) Regulatory controls for osteoblast growth and differentiation: role of Runx/Cbfa/AML factors. Crit Rev Eukaryot Gene Expr 14(1–2):1–41

Makino S (1956) Further evidence favoring the concept of the stem cell in ascites tumors of rats. Ann N Y Acad Sci 63(5):818–830

Martincorena I, Roshan A, Gerstung M, Ellis P, Van Loo P, McLaren S, Wedge DC, Fullam A, Alexandrov LB, Tubio JM, Stebbings L, Menzies A, Widaa S, Stratton MR, Jones PH, Campbell PJ (2015) Tumor evolution. High burden and pervasive positive selection of somatic mutations in normal human skin. Science 348(6237):880–886. doi:10.1126/science.aaa6806

Marusyk A, Tabassum DP, Altrock PM, Almendro V, Michor F, Polyak K (2014) Non-cell-autonomous driving of tumour growth supports sub-clonal heterogeneity. Nature 514(7520):54–58. doi:10.1038/nature13556

Merlo LM, Pepper JW, Reid BJ, Maley CC (2006) Cancer as an evolutionary and ecological process. Nat Rev Cancer 6(12):924–935. doi:10.1038/nrc2013

Michalopoulos GK (2010) Liver regeneration after partial hepatectomy: critical analysis of mechanistic dilemmas. Am J Pathol 176(1):2–13. doi:10.2353/ajpath.2010.090675

Mikhailov AS (1990) Foundations of synergetics. Springer series in synergetics, vol 51–52. Springer, Berlin

Moczek AP, Sultan S, Foster S, Ledon-Rettig C, Dworkin I, Nijhout HF, Abouheif E, Pfennig DW (2011) The role of developmental plasticity in evolutionary innovation. Proc Biol Sci/R Soc 278(1719):2705–2713. doi:10.1098/rspb.2011.0971

Mori H, Lo AT, Inman JL, Alcaraz J, Ghajar CM, Mott JD, Nelson CM, Chen CS, Zhang H, Bascom JL, Seiki M, Bissell MJ (2013) Transmembrane/cytoplasmic, rather than catalytic, domains of Mmp14 signal to MAPK activation and mammary branching morphogenesis via binding to integrin beta1. Development 140(2):343–352. doi:10.1242/dev.084236

Muller GB (2007) Evo-devo: extending the evolutionary synthesis. Nat Rev Genet 8(12):943–949. doi:10.1038/nrg2219

Müller G, Pigliucci M, Konrad Lorenz Institute for Evolution and Cognition Research, Project Muse (2010) Evolution, the extended synthesis

Nanjundiah V, Newman SA (2009) Phenotypic and developmental plasticity. J Biosci 34(4):493–494

Nelson CM, Bissell MJ (2006) Of extracellular matrix, scaffolds, and signaling: tissue architecture regulates development, homeostasis, and cancer. Annu Rev Cell Dev Biol 22:287–309. doi:10.1146/annurev.cellbio.22.010305.104315

Newman SA, Bhat R (2007) Activator-inhibitor dynamics of vertebrate limb pattern formation. Birth Defects Res C Embryo Today 81(4):305–319

Newman SA, Bhat R (2008) Dynamical patterning modules: physico-genetic determinants of morphological development and evolution. Phys Biol 5(1):015008

Newman SA, Bhat R (2009) Dynamical patterning modules: a “pattern language” for development and evolution of multicellular form. Int J Dev Biol 53(5–6):693–705

Newman SA, Frenz DA, Hasegawa E, Akiyama SK (1987) Matrix-driven translocation: dependence on interaction of amino-terminal domain of fibronectin with heparin-like surface components of cells or particles. Proc Natl Acad Sci U.S.A. 84(14):4791–4795

Newman SA, Forgacs G, Muller GB (2006) Before programs: the physical origination of multicellular forms. Int J Dev Biol 50(2–3):289–299

Newman SA, Bhat R, Mezentseva NV (2009) Cell state switching factors and dynamical patterning modules: complementary mediators of plasticity in development and evolution. J Biosci 34(4):553–572

Nik-Zainal S, Alexandrov LB, Wedge DC, Van Loo P, Greenman CD, Raine K, Jones D, Hinton J, Marshall J, Stebbings LA, Menzies A, Martin S, Leung K, Chen L, Leroy C, Ramakrishna M, Rance R, Lau KW, Mudie LJ, Varela I, McBride DJ, Bignell GR, Cooke SL, Shlien A, Gamble J, Whitmore I, Maddison M, Tarpey PS, Davies HR, Papaemmanuil E, Stephens PJ, McLaren S, Butler AP, Teague JW, Jonsson G, Garber JE, Silver D, Miron P, Fatima A, Boyault S, Langerod A, Tutt A, Martens JW, Aparicio SA, Borg A, Salomon AV, Thomas G, Borresen-Dale AL, Richardson AL, Neuberger MS, Futreal PA, Campbell PJ, Stratton MR, Breast Cancer Working Group of the International Cancer Genome C (2012a) Mutational processes molding the genomes of 21 breast cancers. Cell 149(5):979–993. doi:10.1016/j.cell.2012.04.024

Nik-Zainal S, Van Loo P, Wedge DC, Alexandrov LB, Greenman CD, Lau KW, Raine K, Jones D, Marshall J, Ramakrishna M, Shlien A, Cooke SL, Hinton J, Menzies A, Stebbings LA, Leroy C, Jia M, Rance R, Mudie LJ, Gamble SJ, Stephens PJ, McLaren S, Tarpey PS, Papaemmanuil E, Davies HR, Varela I, McBride DJ, Bignell GR, Leung K, Butler AP, Teague JW, Martin S, Jonsson G, Mariani O, Boyault S, Miron P, Fatima A, Langerod A, Aparicio SA, Tutt A, Sieuwerts AM, Borg A, Thomas G, Salomon AV, Richardson AL, Borresen-Dale AL, Futreal PA, Stratton MR, Campbell PJ, Breast Cancer Working Group of the International Cancer Genome C (2012b) The life history of 21 breast cancers. Cell 149(5):994–1007. doi:10.1016/j.cell.2012.04.023

Nowell PC (1976) The clonal evolution of tumor cell populations. Science 194(4260):23–28

Patel VN, Knox SM, Likar KM, Lathrop CA, Hossain R, Eftekhari S, Whitelock JM, Elkin M, Vlodavsky I, Hoffman MP (2007) Heparanase cleavage of perlecan heparan sulfate modulates FGF10 activity during ex vivo submandibular gland branching morphogenesis. Development 134(23):4177–4186. doi:10.1242/dev.011171

Pepper JW, Scott Findlay C, Kassen R, Spencer SL, Maley CC (2009) Cancer research meets evolutionary biology. Evol Appl 2(1):62–70. doi:10.1111/j.1752-4571.2008.00063.x

Plachot C, Lelievre SA (2004) DNA methylation control of tissue polarity and cellular differentiation in the mammary epithelium. Exp Cell Res 298(1):122–132. doi:10.1016/j.yexcr.2004.04.024S0014482704002393 [pii]

Radisky D, Hagios C, Bissell MJ (2001) Tumors are unique organs defined by abnormal signaling and context. Semin Cancer Biol 11(2):87–95. doi:10.1006/scbi.2000.0360

Radisky DC, Levy DD, Littlepage LE, Liu H, Nelson CM, Fata JE, Leake D, Godden EL, Albertson DG, Nieto MA, Werb Z, Bissell MJ (2005) Rac1b and reactive oxygen species mediate MMP-3-induced EMT and genomic instability. Nature 436(7047):123–127. doi:10.1038/nature03688

Raspopovic J, Marcon L, Russo L, Sharpe J (2014) Modeling digits. Digit patterning is controlled by a Bmp-Sox9-Wnt Turing network modulated by morphogen gradients. Science 345 (6196):566–570. doi:10.1126/science.1252960

Reversade B, De Robertis EM (2005) Regulation of ADMP and BMP2/4/7 at opposite embryonic poles generates a self-regulating morphogenetic field. Cell 123(6):1147–1160. doi:10.1016/j.cell.2005.08.047

Roberts SA, Sterling J, Thompson C, Harris S, Mav D, Shah R, Klimczak LJ, Kryukov GV, Malc E, Mieczkowski PA, Resnick MA, Gordenin DA (2012) Clustered mutations in yeast and in human cancers can arise from damaged long single-strand DNA regions. Mol Cell 46 (4):424–435. doi:10.1016/j.molcel.2012.03.030

Sangisetty SL, Miner TJ (2012) Malignant ascites: a review of prognostic factors, pathophysiology and therapeutic measures. World J Gastrointest Surg 4(4):87–95. doi:10.4240/wjgs.v4.i4.87

Sato K, Ito Y, Yomo T, Kaneko K (2003) On the relation between fluctuation and response in biological systems. Proc Natl Acad Sci U.S.A. 100(24):14086–14090. doi:10.1073/pnas.2334996100

Serls AE, Doherty S, Parvatiyar P, Wells JM, Deutsch GH (2005) Different thresholds of fibroblast growth factors pattern the ventral foregut into liver and lung. Development 132(1):35–47. doi:10.1242/dev.01570

Setlur SR, Lee C (2012) Tumor archaeology reveals that mutations love company. Cell 149 (5):959–961. doi:10.1016/j.cell.2012.05.010

Sonnenschein C, Soto AM (2008) Theories of carcinogenesis: an emerging perspective. Semin Cancer Biol 18(5):372–377

Spencer VA, Costes S, Inman JL, Xu R, Chen J, Hendzel MJ, Bissell MJ (2011) Depletion of nuclear actin is a key mediator of quiescence in epithelial cells. J Cell Sci 124(Pt 1):123–132. doi:10.1242/jcs.073197

Steinberg MS, Gilbert SF (2004) Townes and Holtfreter (1955): directed movements and selective adhesion of embryonic amphibian cells. J Exp Zool Part A Comp Exp Biol 301(9):701–706. doi:10.1002/jez.a.114

Suzuki Y, Nijhout HF (2006) Evolution of a polyphenism by genetic accommodation. Science 311 (5761):650–652. doi:10.1126/science.1118888

Tanner K, Mori H, Mroue R, Bruni-Cardoso A, Bissell MJ (2012) Coherent angular motion in the establishment of multicellular architecture of glandular tissues. Proc Natl Acad Sci U.S.A. 109 (6):1973–1978. doi:10.1073/pnas.1119578109

Tarlow BD, Pelz C, Naugler WE, Wakefield L, Wilson EM, Finegold MJ, Grompe M (2014) Bipotential adult liver progenitors are derived from chronically injured mature hepatocytes. Cell Stem Cell 15(5):605–618. doi:10.1016/j.stem.2014.09.008

Ting DT, Wittner BS, Ligorio M, Vincent Jordan N, Shah AM, Miyamoto DT, Aceto N, Bersani F, Brannigan BW, Xega K, Ciciliano JC, Zhu H, MacKenzie OC, Trautwein J, Arora KS, Shahid M, Ellis HL, Qu N, Bardeesy N, Rivera MN, Deshpande V, Ferrone CR, Kapur R, Ramaswamy S, Shioda T, Toner M, Maheswaran S, Haber DA (2014) Single-cell RNA sequencing identifies extracellular matrix gene expression by pancreatic circulating tumor cells. Cell Rep 8(6):1905–1918. doi:10.1016/j.celrep.2014.08.029

Urtasun R, Latasa MU, Demartis MI, Balzani S, Goni S, Garcia-Irigoyen O, Elizalde M, Azcona M, Pascale RM, Feo F, Bioulac-Sage P, Balabaud C, Muntane J, Prieto J, Berasain C, Avila MA (2011) Connective tissue growth factor autocriny in human hepatocellular carcinoma: oncogenic role and regulation by epidermal growth factor receptor/yes-associated protein-mediated activation. Hepatology 54(6):2149–2158. doi:10.1002/hep.24587

Valcourt JR, Lemons JM, Haley EM, Kojima M, Demuren OO, Coller HA (2012) Staying alive: metabolic adaptations to quiescence. Cell Cycle 11(9):1680–1696. doi:10.4161/cc.19879

Vrba ES, Eldredge N, Gould SJ, Paleontological Society (2005) Macroevolution: diversity, disparity, contingency: essays in honor of Stephen Jay Gould. Published by the Paleontological Society, Lawrence, KS

Wada H, Sewall KB (2014) Introduction to the symposium-uniting evolutionary and physiological approaches to understanding phenotypic plasticity. Integr Comp Biol 54(5):774–782. doi:10.1093/icb/icu097

Weaver VM, Petersen OW, Wang F, Larabell CA, Briand P, Damsky C, Bissell MJ (1997) Reversion of the malignant phenotype of human breast cells in three-dimensional culture and in vivo by integrin blocking antibodies. J Cell Biol 137(1):231–245

West-Eberhard MJ (1989) Phenotypic plasticity and the origins of diversity. Annu Rev Ecol Syst 20:249–278

West-Eberhard MJ (2003) Developmental plasticity and evolution. Oxford University Press, Oxford, New York

Yimlamai D, Christodoulou C, Galli GG, Yanger K, Pepe-Mooney B, Gurung B, Shrestha K, Cahan P, Stanger BZ, Camargo FD (2014) Hippo pathway activity influences liver cell fate. Cell 157(6):1324–1338. doi:10.1016/j.cell.2014.03.060

Chapter 18
Separating Spandrels from Phenotypic Targets of Selection in Adaptive Molecular Evolution

Stevan A. Springer, Michael Manhart and Alexandre V. Morozov

Abstract There are many examples of adaptive molecular evolution in natural populations, but there is no existing method to verify which phenotypic changes were directly targeted by selection. The problem is that correlations between traits make it difficult to distinguish between direct and indirect selection. A phenotype is a direct target of selection when that trait in particular was shaped by selection to better perform a function. An indirect target of selection, also known as an evolutionary spandrel, is a phenotype that changes only because it is correlated with another trait under direct selection. Studies that mutate genes and examine the phenotypic consequences are increasingly common, and these experiments could estimate the mutational accessibility of the phenotypic changes that arise during an instance of adaptive molecular evolution. Under indirect selection, we expect phenotypes to evolve toward states that are more accessible by mutation (i.e., states with high mutational entropy). Deviation from this null expectation (evolution toward a phenotypic state rarely produced by mutation) would be compelling evidence of adaptation and could be used to distinguish direct selection from indirect selection on correlated traits. To be practical, this molecular test of adaptation requires phenotypic differences that are caused by changes in a small number of genes. These kinds of genetically-simple traits have been observed in many empirical studies of adaptive evolution. Here, we describe how to use mutational accessibility to separate spandrels from direct targets of selection and thus verify adaptive hypotheses for phenotypes that evolve by molecular changes at one or a few genes.

S.A. Springer
Department of Cellular and Molecular Medicine,
University of California San Diego, La Jolla, CA 92109, USA

M. Manhart
Department of Chemistry and Chemical Biology,
Harvard University, Cambridge, MA 02138, USA

A.V. Morozov (✉)
Department of Physics and Astronomy, Rutgers University,
Piscataway, NJ 08854, USA
e-mail: morozov@physics.rutgers.edu

P. Pontarotti (ed.), *Evolutionary Biology*, DOI 10.1007/978-3-319-41324-2_18

18.1 Inferring the Phenotypic Target of Selection

18.1.1 The Problems of Pleiotropy, Correlated Traits, and Indirect Selection

Adaptations are phenotypes shaped by selection to perform a function. But we cannot assume that a trait evolved for the function that we happen to assign it (Williams 1966; Gould and Lewontin 1979; Nielsen 2009). To understand the reason a trait evolved, we must formulate and test adaptive hypotheses—scenarios that specify exactly which phenotypic differences created the fitness differences that drove evolution (Williams 1966; Mayr 1983). Tests of adaptive hypotheses are confounded by two problems. The first is that individual mutations typically affect many traits simultaneously, a phenomenon known as pleiotropy. The second is that, as a result, mutational effects on different traits can be correlated, and if so, a trait can change toward a particular state even when it has no direct consequence for fitness. Traits can evolve solely because they are correlated with some other trait that is important to fitness. We refer to this apparent selection resulting from correlations between traits as indirect selection, and to traits that evolve in this manner as evolutionary spandrels (Gould and Lewontin 1979). To verify an adaptive hypothesis, one must be able to distinguish between a spandrel and a trait that was truly a direct target of selection (Pearson 1903; Lande and Arnold 1983).

If we could empirically determine the spectrum of phenotypes available by mutation, we could make two straightforward predictions: (i) Traits that are not themselves under direct selection should tend toward phenotypic states that are reached frequently by mutation (Stadler et al. 2001). In other words, in the absence of direct selection, traits should evolve toward phenotypes that are mutationally accessible (i.e., characterized by high mutational entropy). If the trait in question is weakly coupled to another trait under direct selection, evolution is expected to move its phenotypes toward more accessible states (although, depending on the strength of correlation with the trait under selection, maximally accessible states may never be reached). (ii) On the other hand, direct selection pushes traits along evolutionary paths that increase fitness and can fix beneficial mutations even if they are not easily accessible by mutation alone. Similarly, if a trait persists after a series of mutations, even though most potential mutations change it, then it must be under stabilizing selection. These predictions form the basis of the accessibility test described here—a method that uses mutational accessibility to verify the phenotypic targets of adaptive molecular evolution.

As an example, consider a protein that binds to a ligand but must be correctly folded to do so. The protein has two relevant traits: folding stability (quantified by the fraction of proteins in the cell that are properly folded) and binding strength (quantified by the fraction of proteins in the cell that are bound to the ligand). These two traits are inevitably correlated because a protein can only bind when properly folded, so that the fraction of folded proteins is always greater than or equal to the bound fraction (Fig. 18.1). In principle, selection could act directly on one of these

traits but not the other—solely to improve either the binding interaction, or the stability of folding. But even if selection directly acts on only one of these traits, indirect selection can drive change in the other trait because the effects of mutations on the two traits are correlated (Manhart and Morozov 2015). For example, even if there is no direct selection for the binding interaction under consideration—in which case there is direct selection for folding only (Fig. 18.1, blue horizontal arrow) due to the loss of other protein functions or the toxicity of misfolded proteins—the fraction of bound proteins will tend to increase simply because protein sequences that bind strongly are more abundant among protein sequences that also fold stably. In other words, protein sequences that bind well become more accessible by mutation as folding improves. Improved binding could thus be a spandrel that evolves in the absence of direct selection.

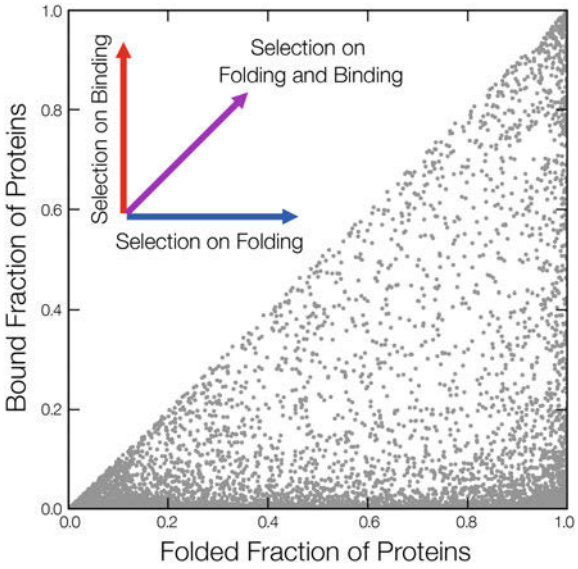


Fig. 18.1 Mutational correlations and indirect selection on protein traits. Distribution of two possible protein traits—folding (quantified by the fraction of proteins that are folded) and binding (quantified by the fraction of proteins that are bound to a ligand)—in a simple thermodynamic model of protein kinetics (Manhart and Morozov 2015). Each residue makes an independent, additive contribution to the free energies of folding and binding (Wells 1990), which in turn determine the folding and binding probabilities via the Boltzmann distribution. Here, we consider protein sequences varying only at 6 sites (e.g., at the binding interface), with a reduced alphabet of 5 amino acid types. The free energy contribution of each amino acid type at each site is randomly sampled from a Gaussian distribution with mean and standard deviation of 1 kcal/mol, consistent with observed distributions of mutational effects (Thorn and Bogan 2001; Tokuriki et al. 2007); total free energies are offset such that mean (over all possible sequences at the binding interface) free energy of folding is 2 kcal/mol and the mean free energy of binding is 5 kcal/mol. Each *gray* point represents the folding and binding traits of a different protein sequence. These two traits are correlated because the fraction of folded proteins is always greater than, or equal to, the fraction of bound proteins (since the protein can only bind when properly folded). Arrows indicate different direct selection scenarios: direct selection on binding only (*red arrow*), direct selection on folding only (*blue arrow*), and direct selection on both binding and folding (*magenta arrow*)

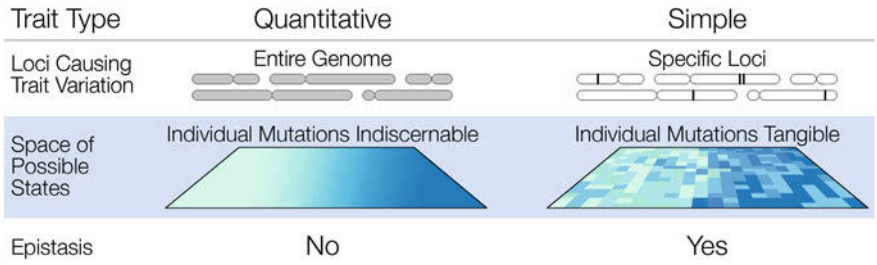


Fig. 18.2 The phenotype spaces of quantitative and simple traits. Finding the phenotypic targets of selection in quantitative traits and genetically-simple traits requires different assumptions about phenotype space. Classical optimality models reduce the complexity of phenotype space by assuming that any small phenotypic change can occur, that there is no epistasis, and that the structure of the phenotype space does not change from one mutation to the next. This is justified for quantitative traits determined by a large number of genomic loci. In contrast, the mutational accessibility test considered here focuses on simple traits (traits dependent on just a few genomic loci). In simple traits, the size of the phenotype space is reduced, allowing us to study variants of just those loci that cause an observed phenotypic difference. The effects of individual mutations may be epistatic for simple traits, which can significantly alter the space of mutational effects available to evolution. Therefore, these traits will benefit most from an explicit genetic investigation

Molecular tests of phenotypic accessibility are only useful if we can quantify the space of possible phenotypes available by mutation for a particular trait. For quantitative traits determined by a large number of genomic loci, this difficult problem can be avoided by assuming that any small phenotypic change can be achieved by mutation (Maynard Smith 1978; Grafen 1984) (Fig. 18.2). Measures of the variances and covariances of quantitative traits can help to identify the phenotypic targets of single selective events and find Pareto-optimal compromises between different selective regimes (Lande and Arnold 1983; Crespi 1990; Shoval et al. 2012). However, extending these kinds of analyses from quantitative traits to traits with a simple genetic basis (phenotypic differences caused by mutations at a few specific genomic loci) requires information or assumptions about mutational constraints on phenotypic change.

18.1.2 Natural Adaptations Can Have Simple Genetics

A striking empirical finding is that some instances of natural adaptive evolution have a simple genetic basis (Orr and Coyne 1992; Bell 2009; Conte et al. 2012; Martin and Orgogozo 2013; Gallant et al. 2014; Rosenblum and Parent 2014). Natural phenotypic changes sometimes occur by mutations in one or a few genes (Nachman et al. 2003; Bradshaw and Schemske 2003; Hoekstra et al. 2006; Storz et al. 2007), and the same genes or even the same mutations can evolve in parallel (Wichman et al. 1999; Holder and Bull 2001; Cresko et al. 2004; Colosimo et al. 2005; Zhang 2006; Musset et al. 2007; McDonald et al. 2009; Jiang et al. 2012;

Shen et al. 2012; Frankel et al. 2012; Springer et al. 2014; Wessinger and Rausher 2015). Artificial mutants of laboratory organisms can have phenotypes that resemble related species, sometimes due to the same genes or pathways that cause the natural difference (Koufopanou and Bell 1991; Parichy and Johnson 2001; Shapiro et al. 2004; Owen and Bradshaw 2011). Mimics are known to use the same genes as the organisms they model to achieve similar forms, and even regulatory elements can evolve in parallel (Reed et al. 2011; Gallant et al. 2014). Coevolving proteins maintain their partners over long-time intervals (Clark et al. 2009; Hellberg et al. 2012), and similar protein adaptations can arise independently in response to similar ecological interactions (Feldman et al. 2012; Dobler et al. 2012; Zhen et al. 2012). Evolution sometimes uses the same genetic elements repeatedly, so for some adaptations there must be only a few loci whose mutations can create an appropriate phenotypic change (Fig. 18.2). In such cases, it may be possible to move beyond just identifying the genes that cause phenotypic differences, by quantitatively estimating which phenotypic changes are common and which are rare when these causal loci are experimentally mutated. For natural phenotypic differences with simple genetic causes, we can thus attempt to empirically evaluate the accessibility of the derived phenotype by mutagenesis.

18.1.3 Distinguishing Traits

There are fundamental limitations on our ability to resolve direct and indirect selection. It will not be possible to distinguish the phenotypic targets of selection from their pleiotropic effects when these components of phenotypic variation are strongly correlated (i.e., with the linear correlation coefficient close to ± 1.0). But selection is also powerless to make this distinction—in terms of evolutionary response, perfectly-correlated phenotypes can in fact be considered one trait (Lewontin 1978; Stadler et al. 2001; Wagner and Zhang 2011). All tests of adaptation (whether quantitative-genetic or molecular) thus rest on the fact that we can only separate direct and indirect selection when mutational effects on the measured traits are not too strongly correlated. In other words, for a set of mutations with the same effect on the trait under selection, each must have an independent suite of effects on other traits (Lewontin 1978; Stadler et al. 2001). Fortunately, the extent of correlations between traits can be verified experimentally. For example, directed protein evolution experiments commonly observe weak or no correlations between various biophysical traits: mutations that modify protein–ligand affinity or catalytic activity typically have varying effects on stability, and non-pleiotropic mutations which affect one trait and not another are nearly always available (Bloom and Arnold 2009). Lack of correlation between traits is also observed in natural contexts. For example, mutations of the Agouti locus have a suite of effects on light and dark coloration in mice, but selection can separate one trait from another (Linnen et al. 2013). In this article, we will assume that when we refer to a trait, we mean the feature being measured, together with all the features that are strongly correlated

with it. Therefore, strictly speaking, a trait is a set of phenotypic features with correlations that cannot be weakened by mutation.

18.2 Practical Exploration of Phenotype Space

For traits with a simple genetic basis, it is possible to recreate the set of mutations that caused a natural phenotypic difference and explore their phenotypic effects one-by-one (Dean and Thornton 2007; Harms and Thornton 2013). In one of the first examples, Weinreich et al. recreated mutational paths during the evolution of the TEM antibiotic resistance gene (Weinreich et al. 2006). Each path is a different way of ordering the set of mutations that separate the ancestral and derived alleles. The researchers measured each mutation's phenotypic effect on cefotaxime resistance and estimated its fixation probability by assuming that selection acted only on this measured trait. Exploring phenotype space in this way is ambitious even for the simplest traits. As the number of mutations between the ancestral and derived alleles increases, the number of possible intermediate alleles goes up exponentially, and the number of possible mutational paths increases factorially (Weinreich et al. 2005). To estimate accessibility, we need methods that generate vast molecular diversity and measure its phenotypic consequences. Equally importantly, we need evolutionary models that describe tractable representations of phenotype space.

18.2.1 Creating and Phenotyping Molecular Diversity with Combinatorial Molecular Biology

Modern molecular biology methods can be used to measure the phenotypic effects of many protein variants and link each phenotype to its DNA sequence (Scott and Smith 1990). Combinatorial mutagenesis can generate mutational diversity exceeding 10^{13} variants (Weiss et al. 2000; Overstreet et al. 2012). Together, these methods have been used to explore huge spaces of molecular variation, e.g., seeking new drug-binding partners, or mapping functional residues in protein–protein interactions. For example, the binding energetics of human growth hormone (hGH) and its receptor (hGHR) have been completely mapped by quantitative saturation mutagenesis: Every possible single amino acid mutation of the hGH binding site has been created and tested for receptor affinity (Pál et al. 2006). High-throughput combinatorial methods can recreate ancestral proteins (Zhu et al. 2005; Lunzer et al. 2005; Szendro et al. 2013), find mutations that interconvert the activity of related but functionally divergent proteins (O'Maille et al. 2008), and explore the function of intermediates between the ancestral and derived state (Weinreich et al. 2006; Bridgham et al. 2009; Field and Matz 2010). Libraries of cis-regulatory regions can measure the effect of regulatory sequence variation in transcription factor binding and protein expression (Gertz et al. 2009). Cell sorters

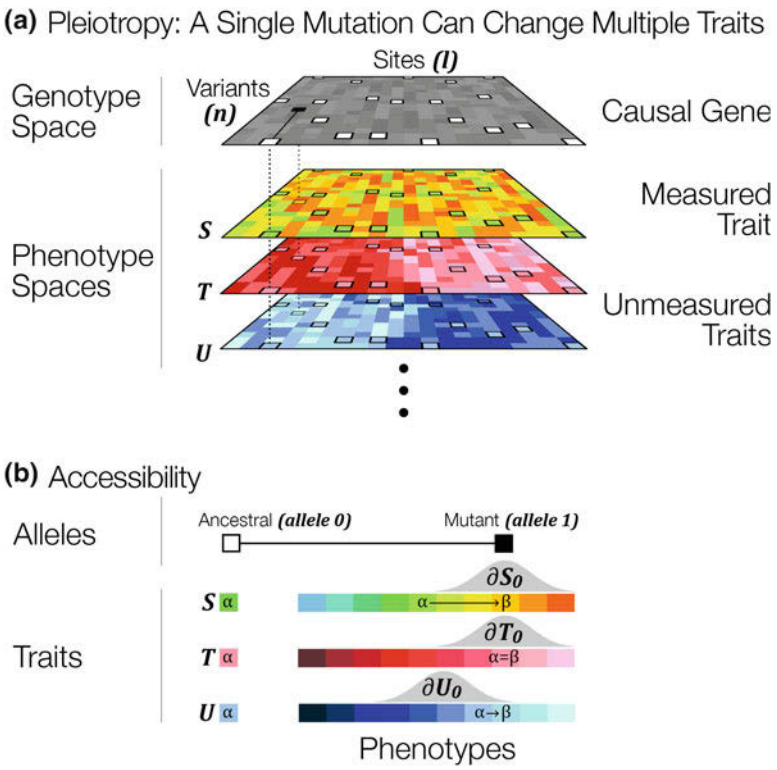


Fig. 18.3 Pleiotropy, accessibility, and evolution by indirect selection. **a** A genotype space and its corresponding phenotype spaces. Each of the $\boldsymbol{l}$ sites has $\boldsymbol{n}$ mutational variants. Thus, the mutational neighborhood size (i.e., the number of single-point mutations available to a sequence) is $\boldsymbol{l} \times \boldsymbol{n}$, and the total number of sequences is $\boldsymbol{n}^{\boldsymbol{l}}$. For simplicity, we have not shown strongly deleterious mutations, which would result in null phenotypes. *White boxes* show the wild-type sequence; the *black box* shows a mutation, which will change the measured trait **S**, as well as unmeasured traits $\boldsymbol{T}$ and $\boldsymbol{U}$. It is difficult to verify the target of selection because a single mutation can have pleiotropic effects on multiple traits. **b** Mutational neighborhoods of individual alleles are denoted as $\partial\, \boldsymbol{Trait}_{\boldsymbol{allele}}$. *Gray* frequency distributions represent mutational accessibility: They show how often a particular phenotypic change occurs by a single mutation of an ancestral allele (***allele 0***) for different traits ($\boldsymbol{S}$, $\boldsymbol{T}$, $\boldsymbol{U}$). Trait $\boldsymbol{S}$: most mutations shift the ancestral phenotype *green* (α) toward *orange* (β); the observed phenotypic change (from *green* to *orange*, $\alpha \rightarrow \beta$) would occur frequently even in the absence of direct selection. Trait $\boldsymbol{T}$ is neutral with respect to this mutation; its phenotype did not change, nor was it expected to, given that mutations with phenotype $\alpha = \beta$ are common in $\partial\, \boldsymbol{T_0}$. Trait $\boldsymbol{U}$ evolved in an unexpected direction, away from mutationally accessible states. Selection on this trait may have caused the change observed in trait $\boldsymbol{S}$ because of their shared genetic basis

can be employed to quickly measure many thousands of cells for any phenotype that can be coupled to a fluorescent marker. Thus, practical tools for creating and phenotyping molecular diversity on a large scale exist. To verify which phenotypes were targets of selection, we must use these tools to explore phenotype space in ways that inform us about the accessibility of a phenotypic change (Fig. 18.3).

18.2.2 Tractable Representations of Phenotype Space

To create the most general model of phenotypic evolution, we would need to study every phenotypic change created by every mutation of the genome, and then consider evolutionary trajectories in this vast phenotypic space (Maynard Smith 1978). This is clearly impossible and will remain so—the number of possible phenotypes and evolutionary trajectories is simply too large to be ever explored completely. If we hope to gain an empirical measure of accessibility, we must reduce the combinatorial complexity and the size of phenotype space. To reduce the number of evolutionary trajectories to be considered, we often assume that selection eliminates or fixes mutations more quickly than the waiting time between mutations (the strong selection, weak mutation, or SSWM regime). Therefore, mutations fix one at a time, and evolution works on the set of phenotypes made available by single mutations of each successively-fixed allele (Gillespie 1984; Weinreich et al. 2005). Thus we can explore the neighborhood of single mutations to estimate the null probability of a particular phenotypic change (note that although this null is shaped by correlations between traits, we do not need to explicitly specify the correlated traits). The size of this neighborhood depends on the number of mutations in the genome that influence our trait of interest. Are mutations of small phenotypic effect spread out evenly, as is typically assumed in quantitative-trait genetics, or are there a few loci that influence a trait much more than others (Fig. 18.2)? The approach described in this article is only needed when genetic constraints significantly influence the course of phenotypic evolution, which is more likely when traits have simple genetic causes (Grafen 1984; Springer et al. 2011). In cases with simple genetics, we can focus on coding and regulatory sequences of the genes known (or suspected) to form the genetic basis of the trait of interest. In the opposite limit of numerous mutations with small phenotypic effects, existing methods from quantitative genetics are capable of distinguishing direct and indirect selection (Lande and Arnold 1983; Shoval et al. 2012).

18.3 The Mutational Accessibility Test

To illustrate the accessibility test, we will use one of the many examples of positive selection whose phenotypic effects can be determined but whose exact evolutionary cause is unknown (Stolz et al. 2003; Nielsen 2009). The ventral and dorsal light organs of Jamaican click beetles (*Pyrophorus plagiophthalamus*) fluoresce in a variety of colors because of variation in their luciferase proteins (Wood et al. 1989). The inferred ancestral luciferase emits green light. Alleles encoding yellow and orange fluorescence have evolved by positive selection, as confirmed by both McDonald-Kreitman and dN/dS tests (Stolz et al. 2003). But what was the phenotypic target that caused this adaptive molecular evolution? Is emitting orange an adaptation, in the sense that alleles producing more orange light conferred higher

fitness? Or is the color shift inevitable, an indirect consequence of selection on an unmeasured trait that evolved because most mutations of the luciferase protein change the phenotype toward orange?

We can answer this question by determining the probability of evolving toward orange in the absence of direct selection. If most mutations in the neighborhood of the ancestral click beetle luciferase cause it to glow orange, we are not forced to invoke direct selection on color to explain the switch to orange, as it could have evolved with or without direct selection. However, if these mutations are in fact non-neutral, high mutational accessibility of orange implies that this phenotypic change could have been driven by indirect selection. In other words, selection may have targeted another trait encoded by luciferase—for example, antioxidant activity or protease sensitivity (Thompson et al. 1997; Barros and Bechara 2000). Alternatively, the switch to orange could be both beneficial as well as highly accessible. In contrast to the above scenario, observation of inaccessible mutations is an indicator of direct selection on the color trait.

18.3.1 Single-Mutation Scenario

If there were only a single mutation separating ancestral and derived alleles, the test would be conceptually simple. Specifically, we could sample a neighborhood of possible mutations of the ancestral sequence by mutagenesis. When expressed in *E. coli*, luciferase proteins fluoresce at the same wavelength as their natural equivalents (Stolz et al. 2003). We can thus measure the effects of individual mutations by cloning a library of artificial mutations *en masse* into individual *E. coli* cells and phenotyping each variant with a fluorescence activated cell sorter (FACS). The aim is to compare the phenotype of the natural mutation with the distribution of phenotypes of possible single mutations to determine whether the natural mutation is atypical. Specifically, we could use the neighborhood to assign a p-value to the probability of a random mutation occurring with a phenotypic change at least as extreme as the one observed. A low p-value would thus imply that a mutation with the observed effect occurs rarely in the absence of direct selection. Note however that, as discussed above, a high p-value does not rule out direct selection on the trait. Empirical maps of mutational neighborhoods allow us to estimate accessibility (i.e., the probability of moving from one phenotypic state to another) in our null model of evolution by indirect selection, and therefore to assess whether or not we need to invoke direct selection to explain an observed phenotypic difference (Fig. 18.4).

18.3.2 Multiple-Mutation Scenario

Now let us consider a more common case. Typically, we have a set of several substitutions that cause a phenotypic difference between the ancestral and derived

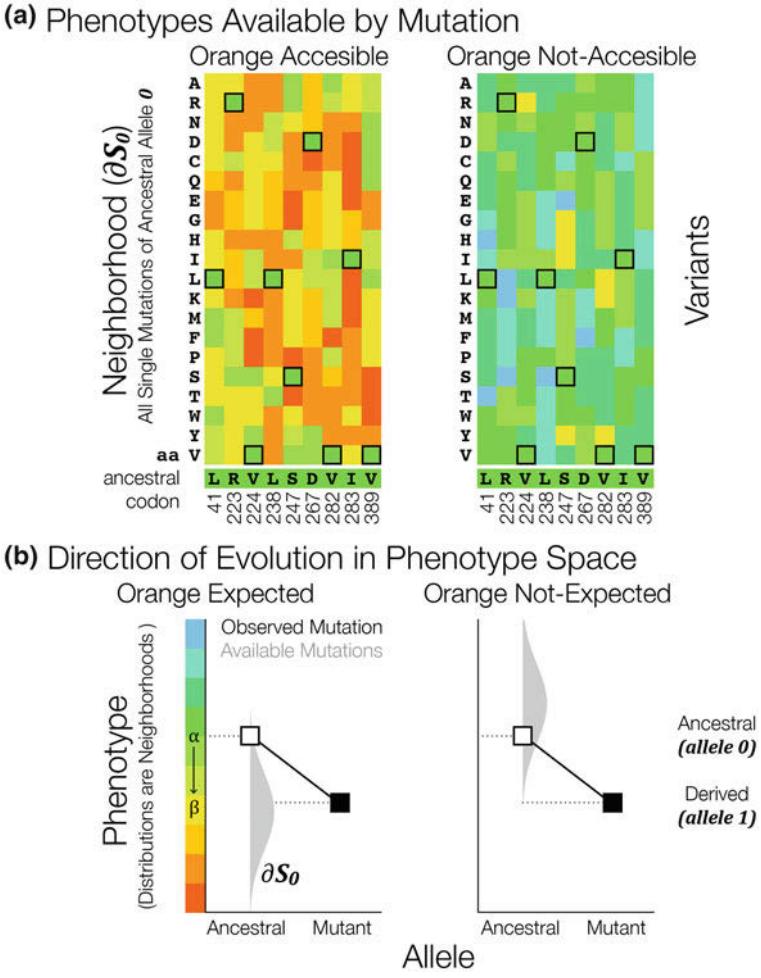


Fig. 18.4 Distribution of phenotypic effects due to a single mutation. **a** Two possible mutational neighborhoods of a green luciferase allele. Amino acids are arranged alphabetically by their full names. The amino acid states of the ancestral luciferase allele are outlined in boxes, and their identity and position in the sequence are shown below. All amino acid substitutions at the positions that are known to be variable in nature (Stolz et al. 2003) are shown. For simplicity, we assume that none of the mutations result in non-viable, "no color" phenotypes. The distribution of phenotypic changes caused by all single amino acid substitutions of the ancestral allele ***0*** is denoted as $\partial\, \boldsymbol{S_0}$. In the *left panel*, the *orange color* is highly accessible. A change to this luciferase for any reason is likely to shift its emission toward *orange*. In the *right panel*, most mutations move the phenotype toward *blue*. A shift toward *orange* would therefore not be expected by chance or by indirect selection on a trait that is *not* strongly correlated with the color trait under consideration. **b** Comparing the effect of an observed mutation with a neighborhood of possible mutations. *White* and *black boxes* indicate the ancestral and derived (mutant) alleles, respectively; their corresponding phenotypes are on the Y-axis. *Gray* distributions show frequencies of phenotypic changes for each mutational neighborhood. These distributions can be used to estimate the probabilities that a random mutation of the ancestral allele will move the phenotype toward or away from the derived state for this particular trait

phenotypes (the positively-selected mutations of the luciferase protein in this example). We have inferred (or know) the ancestral state, but we do not know the temporal order of these substitutions or the evolutionary paths taken by the population. In particular, we do not know whether the number of substitutions observed between the ancestral and derived sequences is typical, given the amount of evolutionary time that elapsed in the course of the observed shift from green to orange. Given that the actual path taken from the ancestral to the derived allele is difficult or impossible to reconstruct, we propose the following extension of the single-mutation accessibility test. Consider phenotypic measurements for a random subset of genotypes separated by the same number of mutations from the ancestral genotype as the derived genotype. For simplicity, only mutations at the sites found to be different between the ancestral and derived alleles will be taken into account; mutations at all other sites are assumed not to change the phenotype (this assumption can be tested experimentally, and if it proves incorrect, mutations at additional sites can be considered as well). These measurements of phenotypic states can be used to estimate a phenotype probability distribution, which in turn enables a *p*-value assignment to the derived state's phenotype in the absence of direct selection.

This type of reasoning, based on collecting information about likely and unlikely changes in the value of the observed trait, can also be extended to the situations when information about evolutionary paths is available. This may be the case in laboratory evolution experiments, where evolving populations can be monitored at regular intervals (Szendro et al. 2013; Jiang et al. 2013; González et al. 2015). In this scenario, the likelihood of each phenotypic change along the path (a member of the ensemble of paths connecting the ancestral and derived alleles) can be assessed using probability distributions constructed from observations of phenotypic states of mutational neighborhoods for each state along the observed path (Fig. 18.5). Even a single low-likelihood move anywhere along the evolutionary path will constitute strong evidence for the direct selection scenario on the trait under observation. Furthermore, weaker evidence from several steps along the path can be combined by considering their joint probability, which should increase the overall statistical power of the null (indirect selection only) hypothesis test.

18.4 Challenges Facing Molecular Tests of Adaptation

Phenotypic measurements. Any test of phenotypic evolution rests on our ability to measure phenotypic effects accurately. Phenotypes measured in the laboratory must resemble their natural counterparts and must be independent of historical, biological, or experimental context that might that obscure their natural effects.

Size of phenotype space. The approach we have described is only useful for genetically-simple traits, whose variation depends on a small number of genomic loci (i.e., mutations elsewhere do not change the phenotype under study). Many natural adaptations have simple genetic underpinnings, and many examples of

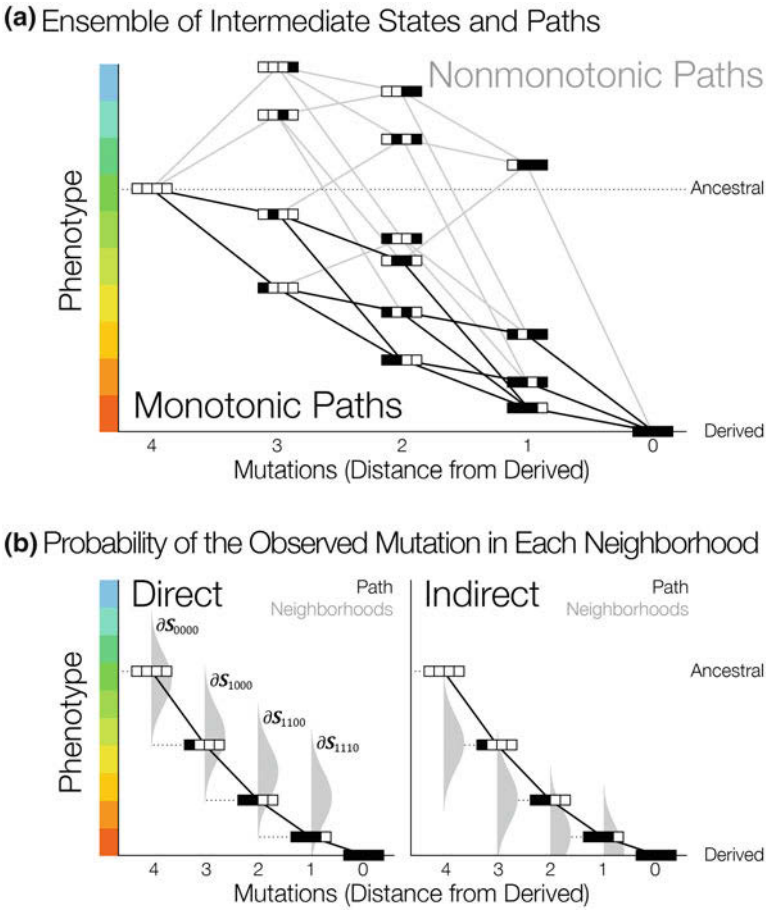


Fig. 18.5 Using mutational accessibility along evolutionary paths to infer the phenotypic target of selection. **a** The ensemble of paths showing all possible intermediate states connecting the ancestral allele and the derived allele. The value of the trait changes monotonically along some of the paths ("monotonic paths"), while on others ("nonmonotonic paths"), there is at least one reversal of the trait—a potential indication of indirect rather than direct selection if this type of path is commonly observed. Note that here we assume for simplicity that there are no strongly deleterious mutations, which would not produce any phenotype. **b** Estimating the probability of making the observed mutational substitution at each step along the observed path. For each intermediate allele, we calculate the probability of making the observed substitution using the current allele's mutational neighborhood. In the *left panel*, all successive substitutions are unlikely, making direct selection the most probable scenario. In contrast, in the *right panel,* each mutation is likely (according to the distribution of phenotypic changes in the mutational neighborhoods), meaning that the accessibility test gives no reason to reject the indirect selection scenario

molecular, regulatory, and network evolution can be studied mutation-by-mutation (Stern and Orgogozo 2008; Bell 2009; Conte et al. 2012). Indeed, it is simple traits that benefit the most from explicit genetic analysis, as their evolution may be

constrained by their genetics. In these traits, evolutionary processes can be subject to explicit molecular investigations of phenotype space, provided that one can devise appropriate large-scale phenotype assays.

Reconstruction of ancestral states and paths. Accurate inference of ancestral states and corresponding evolutionary trajectories is a serious challenge for molecular tests of adaptation. Outside of lineages whose evolution can be observed in real time, the ancestral sequence and the temporal order of mutations could be difficult or impossible to reconstruct (Maynard Smith and Haigh 1974; Gillespie 2000). Ancestral alleles can sometimes remain in a population, but we will only find these cases circumstantially; they cannot underpin a general approach (Springer and Crespi 2007; Rebeiz et al. 2009). Tests of direct and indirect selection in molecular adaptation will, therefore, have to allow for multiple ancestral sequence candidates and multiple evolutionary paths.

Population genetics assumptions. Our approach is based on the SSWM assumption—mutations arise one at a time and either fix or disappear, keeping the population monomorphic at all times. Depending on the strength of selection, either the first beneficial mutation discovered will always fix or, somewhat more realistically, the fixation probabilities of beneficial mutations will be proportional to their selection coefficients (Gillespie 1984). In reality, populations may be polymorphic, which could significantly complicate the analysis (although a certain degree of polymorphism can be tolerated by considering the most common sequence variants found in the population). Furthermore, in small populations, substitutions may occur due to genetic drift rather than selective advantage, although standard tests to distinguish selection from neutrality (such as McDonald-Kreitman and dN/dS) should help to address this problem. Finally, the effects of fluctuating selection and, more generally, past environmental changes may be difficult to capture in artificial conditions.

18.5 Conclusions: Testing Adaptation Itself

When we call a trait an adaptation, we imagine selection on a particular aspect of phenotypic variation ("hummingbird-pollinated flowers are red because…" or "cave fish lose their eyes because…"). Adaptive explanations invoke specific historical scenarios—both a particular form of selection and, as importantly, assumptions about evolutionary constraints or lack thereof. Adaptation is the center of evolutionary biology, and determining which phenotypes were direct targets of selection is the key to testing adaptation. We cannot verify an adaptive hypothesis without eliminating the alternative: evolution as a response to selection on a correlated trait (Williams 1966; Gould and Lewontin 1979; Nielsen 2009). A growing number of studies are using artificial mutagenesis to uncover the range of functional phenotypic variation available by mutation. These methods could in principle also be used to verify direct selection, if we could identify unambiguous hallmarks of adaptation. The evolution of a phenotype away from mutationally accessible states is a feature of adaptation by direct selection that could potentially be estimated.

Measuring mutational accessibility is not trivial, but it can be achieved using existing molecular biology tools. Conveniently, the only traits that truly need genotype–phenotype analysis are genetically the simplest and most tractable (Grafen 1984; Springer et al. 2011). Thus, molecular genetics can overcome the problems posed by pleiotropy and phenotypic correlations and satisfy the deepest goal of evolutionary research—verifying that a phenotypic difference was a direct target of selection and can therefore be considered an adaptation.

Competing Interests

The authors declare no competing interests.

Acknowledgements SAS is grateful to WJ Swanson, J Tyerman, F Breden, and HD Bradshaw for helpful discussions. An NSERC PGSD International Graduate Fellowship to SAS supported this work. MM was supported by NIH grant F32 GM116217. AVM and SAS are grateful to Pierre Pontarotti for his extraordinary hospitality, which enabled the exchange of ideas that ultimately made this work possible.

References

Barros MP, Bechara E (2000) Luciferase and urate may act as antioxidant defenses in larval Pyrearinus termitilluminans (Elateridae: Coleoptera) during natural development and upon 20-hydroxyecdysone treatment. Photochem Photobiol 71:648–654

Bell G (2009) The oligogenic view of adaptation. Cold Spring Harb Symp Quant Biol 74: 139–144. doi:10.1101/sqb.2009.74.003

Bloom JD, Arnold FH (2009) In the light of directed evolution: pathways of adaptive protein evolution. Proc Natl Acad Sci USA 106(Suppl 1):9995–10000. doi:10.1073/pnas.0901522106

Bradshaw HD, Schemske DW (2003) Allele substitution at a flower colour locus produces a pollinator shift in monkeyflowers. Nature 426:176–178. doi:10.1038/nature02106

Bridgham JT, Ortlund EA, Thornton JW (2009) An epistatic ratchet constrains the direction of glucocorticoid receptor evolution. Nature 461:515–519. doi:10.1038/nature08249

Clark NL, Gasper J, Sekino M et al (2009) Coevolution of interacting fertilization proteins. Plos Genet 5:e1000570. doi:10.1371/journal.pgen.1000570

Colosimo PF, Hosemann KE, Balabhadra S et al (2005) Widespread parallel evolution in sticklebacks by repeated fixation of Ectodysplasin alleles. Science 307:1928–1933. doi:10.1126/science.1107239

Conte GL, Arnegard ME, Peichel CL, Schluter D (2012) The probability of genetic parallelism and convergence in natural populations. Proc Roy Soc Lond B 279:5039–5047. doi:10.1098/rspb.2012.2146

Cresko WA, Amores A, Wilson C et al (2004) Parallel genetic basis for repeated evolution of armor loss in Alaskan threespine stickleback populations. Proc Natl Acad Sci USA 101:6050–6055. doi:10.1073/pnas.0308479101

Crespi BJ (1990) Measuring the effect of natural-selection on phenotypic interaction systems. Am Nat 135:32–47

Dean AM, Thornton JW (2007) Mechanistic approaches to the study of evolution: the functional synthesis. Nat Rev Genet 8:675–688. doi:10.1038/nrg2160

Dobler S, Dalla S, Wagschal V, Agrawal AA (2012) Community-wide convergent evolution in insect adaptation to toxic cardenolides by substitutions in the Na, K-ATPase. Proc Natl Acad Sci USA 109:13040–13045. doi:10.2307/41685674?ref=no-x-route:9c57945a3d654bf0c27836b49e38d101

Feldman CR, Brodie ED, Pfrender ME (2012) Constraint shapes convergence in tetrodotoxin-resistant sodium channels of snakes. Proc Natl Acad Sci 109:4556–4561. doi:10.1073/pnas.1113468109
Field SF, Matz MV (2010) Retracing evolution of red fluorescence in GFP-like proteins from faviina corals. Mol Biol Evol 27:225–233. doi:10.1093/molbev/msp230
Frankel N, Wang S, Stern DL (2012) Conserved regulatory architecture underlies parallel genetic changes and convergent phenotypic evolution. Proc Natl Acad Sci USA 109:20975–20979. doi:10.1073/pnas.1207715109
Gallant JR, Imhoff VE, Martin A et al (2014) Ancient homology underlies adaptive mimetic diversity across butterflies. Nat Commun 5:4817. doi:10.1038/ncomms5817
Gertz J, Siggia ED, Cohen BA (2009) Analysis of combinatorial cis-regulation in synthetic and genomic promoters. Nature 457:215–218. doi:10.1038/nature07521
Gillespie J (1984) Molecular evolution over the mutational landscape. Evolution 38:1116–1129
Gillespie J (2000) Genetic drift in an infinite population. The pseudohitchhiking model. Genetics 155:909–919
González C, Ray JCJ, Manhart M et al (2015) Stress-response balance drives the evolution of a network module and its host genome. Mol Syst Biol 11:827–827. doi:10.15252/msb.20156185
Gould SJ, Lewontin R (1979) The spandrels of San Marco and the Panglossian paradigm: a critique of the adaptationist programme. Proc Roy Soc Lond B 205:581–598
Grafen A (1984) Natural selection, kin selection and group selection. In: Behavioural ecology: an evolutionary approach, pp 62–84
Harms MJ, Thornton JW (2013) Evolutionary biochemistry: revealing the historical and physical causes of protein properties. Nat Rev Genet 14:559–571. doi:10.1038/nrg3540
Hellberg ME, Dennis AB, Arbour-Reily P et al (2012) The tegula tango: a coevolutionary dance of interacting, positively selected sperm and egg proteins. Evolution 66:1681–1694. doi:10.1111/j.1558-5646.2011.01530.x
Hoekstra HE, Hirschmann RJ, Bundey RA et al (2006) A single amino acid mutation contributes to adaptive beach mouse color pattern. Science 313:101–104. doi:10.1126/science.1126121
Holder KK, Bull JJ (2001) Profiles of adaptation in two similar viruses. Genetics 159:1393–1404
Jiang P, Josue J, Li X et al (2012) Major taste loss in carnivorous mammals. Proc Natl Acad Sci USA 109:4956–4961. doi:10.1073/pnas.1118360109
Jiang P-P, Corbett-Detig RB, Hartl DL, Lozovsky ER (2013) Accessible mutational trajectories for the evolution of pyrimethamine resistance in the malaria parasite *Plasmodium vivax*. J Mol Evol 77:81–91. doi:10.1007/s00239-013-9582-z
Koufopanou V, Bell G (1991) Developmental mutants of volvox: does mutation recreate the patterns of phylogenetic diversity? Evolution 45:1806–1822
Lande R, Arnold SJ (1983) The measurement of selection on correlated characters. Evolution, pp 1210–1226
Lewontin RC (1978) Adaptation. Sci Am 239:212-8–220-222 passim
Linnen CR, Poh Y-P, Peterson BK et al (2013) Adaptive evolution of multiple traits through multiple mutations at a single gene. Science 339:1312–1316. doi:10.1126/science.1233213
Lunzer M, Miller SP, Felsheim R, Dean AM (2005) The biochemical architecture of an ancient adaptive landscape. Science 310:499–501. doi:10.1126/science.1115649
Manhart M, Morozov AV (2015) Protein folding and binding can emerge as evolutionary spandrels through structural coupling. Proc Natl Acad Sci 112:1797–1802. doi:10.1073/pnas.1415895112
Martin A, Orgogozo V (2013) The Loci of repeated evolution: a catalog of genetic hotspots of phenotypic variation. Evolution 67:1235–1250. doi:10.1111/evo.12081
Maynard Smith J (1978) Optimization theory in evolution. Ann Rev Ecol Syst 9:31–56
Maynard Smith J, Haigh J (1974) The hitch-hiking effect of a favourable gene. Genet Res 23:23–35
Mayr E (1983) How to carry out the adaptationist program? Am Nat 121:324–334
McDonald MJ, Gehrig SM, Meintjes PL et al (2009) Adaptive divergence in experimental populations of Pseudomonas fluorescens. IV. Genetic constraints guide evolutionary

trajectories in a parallel adaptive radiation. Genetics 183:1041–1053. doi:10.1534/genetics.109.107110
Musset L, Le Bras J, Clain J (2007) Parallel evolution of adaptive mutations in Plasmodium falciparum mitochondrial DNA during atovaquone-proguanil treatment. Mol Biol Evol 24:1582–1585. doi:10.1093/molbev/msm087
Nachman MW, Hoekstra HE, D'Agostino SL (2003) The genetic basis of adaptive melanism in pocket mice. Proc Natl Acad Sci USA 100:5268–5273. doi:10.1073/pnas.0431157100
Nielsen R (2009) Adaptionism-30 years after Gould and Lewontin. Evolution 63:2487–2490. doi:10.1111/j.1558-5646.2009.00799.x
O'Maille PE, Malone A, Dellas N et al (2008) Quantitative exploration of the catalytic landscape separating divergent plant sesquiterpene synthases. Nat Chem Biol 4:617–623. doi:10.1038/nchembio.113
Orr HA, Coyne JA (1992) The genetics of adaptation: a reassessment. Am Nat 140:725–742. doi:10.1086/285437
Overstreet CM, Yuan TZ, Levin AM et al (2012) Self-made phage libraries with heterologous inserts in the Mtd of Bordetella bronchiseptica. Prot Eng Des Sel 25:145–151. doi:10.1093/protein/gzr068
Owen CR, Bradshaw H (2011) Induced mutations affecting pollinator choice in Mimulus lewisii (Phrymaceae). Arthropod-Plant Interactions, pp 1–10
Parichy DM, Johnson SL (2001) Zebrafish hybrids suggest genetic mechanisms for pigment pattern diversification in Danio. Dev Genes Evol 211:319–328
Pál G, Kouadio J-LK, Artis DR et al (2006) Comprehensive and quantitative mapping of energy landscapes for protein-protein interactions by rapid combinatorial scanning. J Biol Chem 281:22378–22385. doi:10.1074/jbc.M603826200
Pearson K (1903) Mathematical contributions to the theory of evolution. XI. On the influence of natural selection on the variability and correlation of organs. Philosophical Transactions of the Royal Society of London Series A, Containing Papers of a Mathematical or Physical Character 200:1–66
Rebeiz M, Pool JE, Kassner VA et al (2009) Stepwise modification of a modular enhancer underlies adaptation in a Drosophila population. Science 326:1663–1667. doi:10.1126/science.1178357
Reed RD, Papa R, Martin A et al (2011) optix drives the repeated convergent evolution of butterfly wing pattern mimicry. Science 333:1137–1141. doi:10.1126/science.1208227
Rosenblum EB, Parent CE (2014) The molecular basis of phenotypic convergence. Ann Rev Ecol Evol Syst 45:203–226. doi:10.1146/annurev-ecolsys-120213-091851
Scott JK, Smith GP (1990) Searching for peptide ligands with an epitope library. Science 249:386–390
Shapiro MD, Marks ME, Peichel CL et al (2004) Genetic and developmental basis of evolutionary pelvic reduction in threespine sticklebacks. Nature 428:717–723. doi:10.1038/nature02415
Shen Y-Y, Liang L, Li G-S et al (2012) Parallel evolution of auditory genes for echolocation in bats and toothed whales. Plos Genet 8:e1002788. doi:10.1371/journal.pgen.1002788.t001
Shoval O, Sheftel H, Shinar G et al (2012) Evolutionary trade-offs, Pareto optimality, and the geometry of phenotype space. Science 336:1157–1160. doi:10.1126/science.1217405
Springer S, Crespi BJ (2007) Adaptive gamete-recognition divergence in a hybridizing Mytilus population. Evolution 61:772–783. doi:10.1111/j.1558-5646.2007.00073.x
Springer S, Crespi BJ, Swanson WJ (2011) Beyond the phenotypic gambit: molecular behavioural ecology and the evolution of genetic architecture. Mol Ecol 20:2240–2257. doi:10.1111/j.1365-294X.2011.05116.x
Springer S, Diaz SL, Gagneux P (2014) Parallel evolution of a self-signal: humans and new world monkeys independently lost the cell surface sugar Neu5Gc. Immunogenetics 66:671–674. doi:10.1007/s00251-014-0795-0
Stadler BM, Stadler PF, Wagner GP, Fontana W (2001) The topology of the possible: formal spaces underlying patterns of evolutionary change. J Theor Biol 213:241–274. doi:10.1006/jtbi.2001.2423

Stern DL, Orgogozo V (2008) The loci of evolution: how predictable is genetic evolution? Evolution 62:2155–2177. doi:10.1111/j.1558-5646.2008.00450.x

Stolz U, Velez S, Wood KV et al (2003) Darwinian natural selection for orange bioluminescent color in a Jamaican click beetle. Proc Natl Acad Sci USA 100:14955–14959. doi:10.1073/pnas.2432563100

Storz JF, Sabatino SJ, Hoffmann FG et al (2007) The molecular basis of high-altitude adaptation in deer mice. Plos Genet 3:e45. doi:10.1371/journal.pgen.0030045

Szendro IG, Franke J, de Visser JAGM, Krug J (2013) Predictability of evolution depends nonmonotonically on population size. Proc Natl Acad Sci USA 110:571–576. doi:10.1073/pnas.1213613110

Thompson JF, Geoghegan KF, Lloyd DB et al (1997) Mutation of a protease-sensitive region in firefly luciferase alters light emission properties. J Biol Chem 272:18766–18771. doi:10.1074/jbc.272.30.18766

Thorn KS, Bogan AA (2001) ASEdb: a database of alanine mutations and their effects on the free energy of binding in protein interactions. Bioinformatics 17:284–285

Tokuriki N, Stricher F, Schymkowitz J et al (2007) The stability effects of protein mutations appear to be universally distributed. J Mol Biol 369:1318–1332. doi:10.1016/j.jmb.2007.03.069

Wagner GP, Zhang J (2011) The pleiotropic structure of the genotype-phenotype map: the evolvability of complex organisms. Nat Rev Genet 12:204–213. doi:10.1038/nrg2949

Weinreich DM, Delaney NF, Depristo MA, Hartl DL (2006) Darwinian evolution can follow only very few mutational paths to fitter proteins. Science 312:111–114. doi:10.1126/science.1123539

Weinreich DM, Watson RA, Chao L (2005) Perspective: sign epistasis and genetic constraint on evolutionary trajectories. Evolution 59:1165–1174

Weiss GA, Watanabe CK, Zhong A et al (2000) Rapid mapping of protein functional epitopes by combinatorial alanine scanning. Proc Natl Acad Sci USA 97:8950–8954. doi:10.1073/pnas.160252097

Wells JA (1990) Additivity of mutational effects in proteins. Biochemistry 29:8509–8517

Wessinger CA, Rausher MD (2015) Ecological transition predictably associated with gene degeneration. Mol Biol Evol 32:347–354. doi:10.1093/molbev/msu298

Wichman HA, Badgett MR, Scott LA et al (1999) Different trajectories of parallel evolution during viral adaptation. Science 285:422–424

Williams GC (1966) Adaptation and natural selection. Princeton University Press

Wood KV, Lam YA, McElroy WD, Seliger HH (1989) Bioluminescent click beetles revisited. J Biolumin Chemilumin 4:31–39. doi:10.1002/bio.1170040110

Zhang J (2006) Parallel adaptive origins of digestive RNases in Asian and African leaf monkeys. Nat Genet 38:819–823. doi:10.1038/ng1812

Zhen Y, Aardema ML, Medina EM et al (2012) Parallel molecular evolution in an herbivore community. Science 337:1634–1637. doi:10.1126/science.1226630

Zhu G, Golding GB, Dean AM (2005) The selective cause of an ancient adaptation. Science 307:1279–1282. doi:10.1126/science.1106974

Chapter 19
From Compositional Chemical Ecologies to Self-replicating Ribosomes and on to Functional Trait Ecological Networks

Robert Root-Bernstein and Meredith Root-Bernstein

Abstract In contrast to theories arguing that cellular life has evolved to transmit genes, we propose instead that cellular life evolved to facilitate the full potential of self-replicating ribosomes. Our theory explicitly rejects "master molecule" theories such as Dawkins's "selfish gene" in favor of the emergence of life by means of systems of increasingly networked interactions that carried out metabolic and genetic functions concurrently within a complex chemical ecology. The critical role of networking chemical interactions within this ecology was (as it still is) mediated by all possible forms of molecular complementarity, of which base-pairing in RNA and DNA is just one. Selection for molecular complementarity functional and structural modules vastly increased the probability that networked systems would evolve, eventually resulting in the first self-replicating entity, which we believe was the ribosome. We make six predictions from our ribosome-first theory of cellular evolution that may seem, at first glance, heretical: (1) Ribosomal RNA (rRNA) contains genetic information encoding its own proteins, meaning that it also encodes messenger RNA (mRNA); (2) these proteins bind to the rRNA to form the functional ribosomal structure, but since the rRNA is also functioning as mRNA, the ribosomal proteins must bind to their own mRNA as well; (3) rRNA encodes all of the transfer RNAs (tRNA) required for the translation of its genetic information; (4) thus, tRNAs may be the precursor modules that gave rise to rRNA; (5) rRNA is pleiofunctional, integrating genetic, protein, translational, and structural information often in the same or overlapping sequences and in all reading frames; and (6) since the ribosome gave rise to cellular life, tRNA- and rRNA-like genetic

R. Root-Bernstein (✉)
Department of Physiology, Michigan State University, East Lansing, MI 48824, USA
e-mail: rootbern@msu.edu

M. Root-Bernstein
Department of Bioscience, Aarhus University, Aarhus, Denmark
e-mail: mrootbernstein@gmail.com

M. Root-Bernstein
Institute of Ecology and Biodiversity, Santiago, Chile

P. Pontarotti (ed.), *Evolutionary Biology*, DOI 10.1007/978-3-319-41324-2_19

information must be major building blocks from which cellular genomes evolved. We present evidence supporting all six of these apparently unlikely predictions. Our conclusion is that life is not about the evolution of genes, but the evolution of the kinds of networked interactions through complementarity that characterize ecologies: Genes evolved merely as storage units to "back up" ribosomal functions. This same complementarity-based approach may help to explain why functional traits, rather than genetic populations, appear to network interactions within higher-order systems such as ecosystems and holobionts.

When Darwin published his *Origin of Species* in (1859), one of the arguments against evolution by natural selection that he addressed in his book was the so-called watchmaker argument put forth a few decades earlier by William Paley. In the first three chapters of *Natural Theology* (1824), Paley proposed a thought experiment. Suppose, he wrote that one finds a watch lying in the middle of a field, which could have lain there, for all one knows, forever. One would have no difficulty arguing from the close relationship between the structure and the function of the watch that it was the work of a clever designer. Why, Paley, asked, can we not make the same argument from the evidence of living things found in the paleontological record, and even more eloquently, presented in our own living form, that all life also had a Designer? Darwin, of course, found this argument to be full of logical and evidential holes (e.g., every instance of poor or failed design was evidence of an incompetent designer), but the watchmaker analogy had, by then, taken on a life of its own. Today, evolutionary biologists sometimes refer to nature as a "blind watchmaker," working randomly to produce flawed, imperfect, and often untenable "watches" that all too often go extinct.

Thus, today, in contrast to Paley, we who study the origins of life may ask ourselves: How did a blind watchmaker create a selfish gene that yielded life? Richard Dawkins has proposed some of the iconic thoughts on this matter (Dawkins 1976). Dawkins's ideas, and those of most other evolutionary biologists, center on what Watson and Crick called "the secret of life": the gene. The gene, goes the modern dogma, is the master molecule containing all the necessary information needed to evolve life, natural selection acting on random mutations becoming an inanimate, unseeing replacement for an all-seeing designer. Current research on the RNA-world scenario falls in line with this master molecule approach to evolving life (reviewed in Strobel 2001; Neveu et al. 2013).

There are difficulties with master molecule theories, however (Bowman et al. 2015; Caetano-Anolles and Seufferheld 2013; Galadino et al. 2012; Norris et al. 2012). One is that RNA and DNA cannot do anything by themselves. Only in the most contrived and controlled conditions do these molecules do anything other than just sit. They were selected to store information, and to move it about, not to evolve it. Genes require an integrated metabolic system to have a function. So one thing that master molecule theories fail to do is to explain how the metabolic system that makes possible the functions of DNA and RNA evolved. Beyond that, living systems are characterized by their very high degree of interactivity. A cell is highly

networked in the many ways that we now characterize as interactomes. Every component of a cell interacts more or less specifically with several, and often many, other components. No component functions in isolation. Again, master molecule theories cannot explain how interactomes evolved (Braakman and Smtih 2013; Gross et al. 2014; Mann 2012).

But even beyond the problems of whether master molecules could explain life, there is the experimental problem that polynucleic acids are very difficult to synthesize and certainly never, under any reasonable conditions, evolved in the absence of every other type of molecule associated with life. When Miller (1953) performed his iconic experiment in 1953 adding energy to ammonia, methane, hydrogen, and water, he made amino acids, not nucleic acids. One reason is that he had no source of phosphates. We have recently recreated Miller's experiment but with new conditions.

Where Miller used distilled, deionized water in his apparatus, we substituted what is effectively sea or hot spring water that contained sodium chloride, potassium salts, calcium and magnesium phosphates, and sulfates. We also modified the apparatus so that we could regularly regas it so that we were not limited by the initial concentrations of components. Like Miller, we, too, have produced significant amounts of amino acids. Unlike Miller, we have also produced sugars, nucleic acid bases, nucleic acids, fatty acids, and other metabolites associated with living systems (Root-Bernstein, unpublished data). All this in a "one pot" synthesis over a period of six weeks! The point is that no system that mimics a real-world prebiotic scenario produces just one class of chemical (Johnson et al. 2008; Bada 2013). All such experiments produce many molecules of many classes, and this is almost certainly the ecological context within which prebiotic evolution had to occur.

Recognizing the existence of this diverse chemical ecology forces us to think about prebiotic evolution in new and different ways. In the real world, every chemical reaction that can occur will and does occur. So we need to rethink the way in which evolution proceeded from the first chemical reactions to the first cell. There was never one type of master molecule that evolved first or all alone and could therefore dominate chemical evolution. What evolved was the chemical ecosystem itself as it explored the possibilities and limits of all available prebiotic chemistries (Root-Bernstein and Dillon 1997; Hunding et al. 2006; Root-Bernstein 2012). Some of these molecules were quickly degraded; others that were synthesized interacted stereospecifically with each other to protect themselves against degradation and to promote catalytic reactions from which even more complex molecules with new properties could emerge.

Such stereospecific, reversible (i.e., non-covalent) types of interaction are subsumed under by the term "molecular complementarity." Molecular complementarity usually involves hydrogen bonds, ionic bonds, pi–pi overlap bonds, charge-transfer complexes, or van der Waals attractions. Molecular complementarity mediates most of the important interactions within and among cellular components and is the basis of all interactome studies: polynucleotide specificity; enzyme specificity; receptor specificity; antibody–antigen specificity; and porphyrin- and heme-ion specificity. In fact, molecular complementarity is absolutely ubiquitous among the components of living systems.

Why?

Why is molecular complementarity so important to evolution? Why is all life characterized by extremely tightly regulated interactomes?

A partial explanation for these phenomena has been offered in a very different context by economics Nobel laureate Simon (1969). Simon argued that in complex systems, complementary underlies modularity and modularity promotes system efficiency. Like Paley and Dawkins, Simon, too, imagined a watchmaker, or in this case two of them. Imagine, says Simon, that one watchmaker makes his watches one piece at a time in such a way that until the entire watch is completed, it is unstable and will fall apart. Imagine, however, that the other watchmaker makes her watches in small, stable modules and that once each module is completed, she can set it aside and work on the next without losing her work. Who will make more watches?

Simon worked out an equation describing the work of the two watchmakers. The equation shown here is $C = 1/(1 - p) \exp S$ where

C is the probability of completing a watch
p is the probability of being interrupted
S is the "span" of that unit (i.e., the # of components making up a unit).

If each watch is composed of 1000 pieces and it takes one minute, on average, to put a piece in place, then it will take watchmaker 1 17 h to make a watch. During that time, if he is interrupted by a customer, has to go to the bathroom, and takes a break to eat a meal, or anything else, he loses all his work. He will complete very few watches during his life. Watchmaker 2, however, assembles her watches in stable modules composed of 10 parts, so every 10 min she can complete a stable module if she is not interrupted. Assuming several interruptions per day, she will still complete a watch every 2.5 days. That is a lot more watches!

There is a close analogy between Simon's watchmakers and origins of life problems. If we assume that we need a particular enzyme to carry out a necessary function such as RNA ligation or DNA replication in order for genes to function, and if we assume that this enzyme is of a reasonable length for such proteins—say, 400 amino acids long—then we run into a terrible conundrum in trying to account for the origins of life because there are 400^{20} possible permutations of our protein that will not perform the function we need to perform. Even with the entire history of the universe in which to evolve it, this protein is unlikely to evolve when and where it is needed. And we must multiply this problem for every other gene and protein that a living cell might need. Without stable modularity by which to bootstrap macromolecules, evolution, Simon is telling us, will not work. With modularity, however, evolution becomes much more likely.

Unfortunately, two missing elements marred Simon's economic analysis of systems evolution. One is that he had no mechanism for making modular components stable. What plays that role in prebiotic evolution? The second problem is that he fell into what we call the "engineering fallacy," which is that he assumed that the watchmakers know what components to select at each step and how to fit

them together. The watchmakers whom Simon analyzes certainly knew exactly how to put all 1000 pieces of their watches together and in what order to do so. A blind evolutionary watchmaker such as Dawkins imagines would not have that miraculous knowledge!

So we need to rethink Simon's watchmaker analogy to add in the evolutionary complexities he left out. We need to make him actually blind—and not only blind, but insensate so that he (or she!) cannot feel which part is being connected to which other. In other words, we have to make the watchmaking process truly random in the way that so many biologists talk about organismal evolution.

Now, in order to carry out this new analysis and to make the numbers manageable, we have elected to reduce the number of parts per watch from 1000 to 25. If a blind watchmaker without any knowledge of watch assembly were to produce a watch from 25 pieces in one sitting, he would now have to explore 25 factorial, or more than 10^{25} permutations. Talk about an impossibility! Scale this up to 400 pieces or a 1000, and the problem becomes intractable.

We assert that molecular complementarity is the solution to the permutation disaster. Molecular complementarity provides the necessary stabilizing mechanism for the combined components, automatically selecting which components go together and which do not. And molecular complementarity also does the assembling for the watchmaker. Utilizing molecular complementarity to aggregate only self-selected sets of components, then stable modules result naturally.

Now, if the second watchmaker, the female one, makes her watch in stable modules of, say, five parts, then she only needs to explore five factorial, or 120 permutations, at each stage of evolution. From this process, she needs to make only five sets of five modules. And then, she needs to perform a mere five factorial, or 120, additional explorations to discover which way these five sets of modules need to be assembled to form a functioning 25-part watch. That is only 720 total permutations.

So do the math! 10 to the 25th power permutations or 720 permutations? Scale it up, and the savings become even greater (Root-Bernstein and Dillon 1997; Root-Bernstein 2012).

Obviously, the permutation savings will vary to some extent with the number of stable modules that need to be explored at any given level of organization. The calculation using five parts per stable module given above is merely an example. There is nothing magical about five parts per module rather than three or six, nor need every stable module be comprised of the same number of components—Nature will determine that for herself. We have assumed five-part modules simply because that makes the calculations a bit more transparent. Nor does the selection have to be perfect. In real systems, molecular complementarity will produce stabilities ranging from very stable to hardly stable at all, so the results will not be as "clean" as just described and the permutation "savings" less extensive. But the general principle holds: The production of stable modules by molecular complementary selection can drastically cut down the permutations that need to be explored, perhaps not to the extent just calculated, but certainly by many, many orders of magnitude!

Molecular complementarity and modularity are therefore the heroes of evolution!

Now, we assert that Simon's watchmaker problem is very similar to various origins of life problems. Quite a few eminent scientists (including Jacobson 1955; Hoyle 1983; Crick 1981) have fallen into the trap of thinking that evolution must elaborate every possible permutation of an RNA or protein sequence in order to evolve a functional molecule. Simon's analysis indicates that such a purely stochastic approach to evolution is extremely inefficient and unlikely to succeed. But selection for stable, complementary modules makes the problem tractable and the evolution of complex systems highly likely. And we actually know from the studies of both genes and proteins that modularity is as ubiquitous as molecular complementarity in living systems.

Molecular complementarity also solves another one of life's mysteries, which is the ubiquity of interactomes. The necessary result of complementarity-based modularity is distributed networks or systems of interactive molecular complexes. Some modules may utilize the products of other modules; some may combine structurally with others to increase stability or to yield new functions (such as catalytic activity). Molecular complementarity will mediate these functional interactions and provide the selective drive for ever-increasing organization (Root-Bernstein and Dillon 1997; Hunding et al. 2006; Root-Bernstein 2012).

In short, we do not need a selfish gene; complementarity in any of its diverse forms and expressed through any set or sets of molecules—not single master molecule replication—drives evolution. One might argue that after DNA emerged, it took over evolution, assembling a heterogenous mix of interactomes into a system designed to replicate genes. But none of its molecular characteristics really support this view. DNA is not, by itself, active. It requires proteins to drive its unwinding, its separation, and its replication. By itself, DNA just sits. It is a storage unit, not a functional one. Furthermore, the emphasis on DNA replication ignores the key role of transcription, which is arguably even more fundamental (you cannot carry out gene replication if no transcription to create a functioning cell has taken place). So even if we accept genes as some kind of master molecules, how did these master molecules evolve and what really drives DNA functioning?

Meredith Root-Bernstein came up with a surprising suggestion: If anything wants to "be selfish" by promoting functionality, it is the ribosome! Ribosomes mediate all cellular functions by producing the proteins that are necessary to cellular life. In addition, all evolutionary studies of the origins of the ribosome have reached a perplexing conclusion that the origins of the ribosome significantly predate the evolution of cells (Mushegian 2008; Wang et al. 2009; Fox 2010). Thus, genetic storage of information, protein synthesis, and basic energy metabolism all had to exist before cells. How?

Well, if the ribosome is the functionally "selfish" unit that predates and gives rise to life, there is only one possibility. What, Meredith proposed, if protoribosomes could self-replicate? What if ribosomes were effectively Turing machines, which is to say, machines capable of replicating themselves using the structures and information encoded in their own components? What if, therefore, ribosomes were key

missing links between prebiotic compositional chemistries and the first cells (Root-Bernstein and Root-Bernstein 2015)?

To say that Meredith's idea is radical, anti-dogmatic, and even heretical, might be an understatement. Ribosomal RNA is generally agreed to contain no genes, so how could a ribosome produce the proteins to make its own structure? The ribosome would need messenger RNAs to encode those proteins—where do those come from since the rRNA supposedly has not genetic information? The ribosome would need transfer RNAs in order to translate these hypothetical mRNAs into proteins: What possible source could there be for these tRNAs? And on top of it all, the ribosomal RNA would need to be able to replicate itself!

We decided to assume, as a testable working hypothesis, that all of these things were not only possible, but that ribosomal RNA incorporates all of these functions. This working hypothesis led to the following testable predictions.

First, the ribosomal RNA would have to contain genetic information encoding its own proteins. In other words, it would have to act as messenger RNA as well. This proposition is somewhat heretical in that ribosomal RNA is generally considered to play only one role and that it acts as a scaffold upon which ribosomal proteins are arranged.

Second, these proteins would need to bind to the ribosomal RNA in order to form the functional ribosomal structure, but that also means that since the rRNA is also functioning as mRNA, the ribosomal proteins should bind to their own mRNA as well. In other words, the mRNAs encoding ribosomal proteins would need to be homologous to rRNA.

Third, the ribosomal RNA would need to encode all of the transfer RNAs required for the translation of its genetic information, another heretical notion.

Fourth, if rRNA encodes all tRNAs, then tRNAs may be the functional modules from which rRNA itself evolved.

Fifth, one should therefore find that the ribosomal RNA is pleiofunctional, integrating genetic, protein, translational, and structural information often in the same or overlapping sequences.

And finally, if the ribosome preceded cellular life and represents the first self-replicating, evolvable entity, then one should find that rRNA-like genetic information is a major basis for cellular genomes. This possibility has never been explored in relation to the origins of cellular life.

We now present evidence supporting all six of these predictions, much of which has actually been accumulating for decades without the meaning having become apparent.

If the ribosome is a missing link between a prebiotic chemical ecology and the first cells, then our first prediction is that ribosomal RNA should contain genes encoding proteins related to ribosomal functions and ones related to RNA transcription and replication. This is the case for *E. coli* K12, the bacterial species upon which we have carried out most of our tests. We will simply note that the same kind of data, often in an even more robust form, characterizes other bacterial and archaeal species we have examined. The 5S, 16S, and 23S ribosomal RNAs encode about 3–5 times as many ribosome-related protein sequences than do random

mRNA sequences from the *E. coli* genome, which is a very highly statistically significant difference. These proteins encoded in rRNA also contain significantly more known active sites than the protein fragments encoded by control mRNA sequences (55 % probability vs. 12–22 %) (Root-Bernstein and Root-Bernstein 2015). Even more surprisingly, it has been known for some years that some rRNA- and rDNA-encoded sequences are actually transcribed and translated as functional proteins. Tenson and Mankin (1995, 1996) and Dam et al. (1996) have demonstrated in various bacteria that there is a "short open reading frame in the 23 S rRNA that encodes a pentapeptide (E-peptide) whose expression in vivo renders cells resistant to erythromycin" and which binds directly to the 23S rRNA. Chevalier and Stoddard (2011) have reviewed studies demonstrating that over 250 homing endonucleases having transposase activity are encoded in rDNA and bind back upon the rRNA–mRNA sequences encoding them. Two additional proteins, ribin (Kermekchiev and Ivanova 2001; Barthélémy et al. 2010) and Tar1p (transcript antisense to ribosomal RNA: Coelho et al. 2002; Bonawitz et al. 2008; Galopier and Hermann-Le Denmat 2011), are also known to be encoded by the rDNA sequence encoding rRNA (reviewed in Root-Bernstein and Root-Bernstein 2016).

Our second prediction is that ribosomal RNA-encoded proteins would, in many cases, need to be able to bind to their rRNA in order to self-assemble a ribosome, but since the rRNA is functioning simultaneously as mRNA, these proteins would also autoregulate their own synthesis by binding to their encoding mRNA. This prediction also turns out to be correct. One or more groups have demonstrated that each of the small subunit ribosomal proteins saves for S14, S16, and S17, and most of the large subunit ribosomal proteins as well directly bind not only to the ribosomal RNA but also to their own mRNA resulting in autogenous translation regulation (e.g., Nomura et al. 1980; Olins and Nomura 1981; Gimautdinova et al. 1981; Gregory et al. 1988; Robert and Brakier-Gingras 2001; reviewed in Root-Bernstein and Root-Bernstein 2016). In most cases, it has further been demonstrated that the rRNA and mRNA to which the ribosomal proteins bind are homologous. In addition, many tRNA synthetases not only bind to their tRNA to add the appropriate amino acid, but also bind to their own mRNA to produce autogenous regulation of their activity (e.g., Fukuda et al. 1978; Dykxhoorn et al. 1996; Passador and Linn 1992; Steward and Linn 1992; Van Gemen et al. 1989; Steinmetz et al. 2001; Roth et al. 2005). Surprisingly, none of these data have ever been collated before nor applied to understand the origins of the ribosome itself. The only exception to this oversight seems to be the work of Brandman et al. (2012) on rRNA–ribosomal protein codon usage, which further supports our hypothesis that ribosomal proteins coevolved with rRNA.

Our third prediction is that rRNA contains genes that encode transfer RNAs. Surprisingly, all twenty types of tRNA are encoded in the 16S rRNA at least once by direct cleavage of the rRNA and again in the complementary strand. The tRNAs are encoded yet again, at least once, in the 23S rRNA sequence (Root-Bernstein and Root-Bernstein 2015). In every case, the tRNA showed at least 65 % identity with modern tRNA, and statistical studies demonstrated that these homologies were

unlikely to be a result of chance. Notably, many of the tRNA homologue sequences that we identified in the rRNA can be shown to fold (theoretically) into the cloverleaf forms typical of modern tRNA (Root-Bernstein and Root-Bernstein 2015).

Given the multiple copies of tRNAs found in rRNAs, it is not surprising to find that other investigators have also noticed this fact. Thus, several groups have previously proposed that rRNA evolved by cobbling together tRNA-like modules, thereby supporting our fourth prediction. Our fourth prediction is that since rRNA incorporates all of the tRNAs necessary to function in protein translation, tRNA actually may be the structural and functional modules from which rRNA itself evolved. Bloch et al. (1983, 1984, 1989), De Farias (2013), De Farias et al. (2014), Caetano-Anollés (2002), and Caetano-Anollés and Caetano-Anollés (2015) have each put forward evidence supporting such a proposition. We will return to this point below, because the tRNA-like basis of rRNA means that both tRNAs and rRNAs can act as "genetic" modules from which other genes could have evolved. tRNAs, in short, may have evolved as virus-like particles that worked in tandem to produce functionally active peptides and proteins that were the precursors of ribosomal proteins.

Our fifth prediction, which is that rRNA is pleiofunctional, is also supported by our research. We have demonstrated that rRNA encodes not only tRNA in more than one reading frame, but also, in all six possible reading frames, protein modules that represent active sites of modern proteins. These encodings are present at statistically significantly higher rates than are found in random sets of mRNAs drawn from the *E. coli* K12 genome, suggesting that the very high information density present in the rRNAs is not by chance (Root-Bernstein and Root-Bernstein 2015). In fact, every portion of the 5S, 16S, and 23S rRNAs contains some combination of peptides and proteins and/or tRNAs. These data are consistent with the modern findings of rRNA- and rDNA (i.e., complementary rRNA)-encoded functional proteins summarized above.

Our sixth and final testable prediction regarding self-replicating ribosomes is that such entities would have become the basis for the first cellular genomes. It has already been established above that ribosomes evolved prior to cellular life (Mushegian 2005; Wang et al. 2009; Fox 2010), making it plausible to now postulate that cells and their genomes evolved to protect, stabilize, and optimize ribosome function. Thus, modern genomes should be composed, to an unexpected degree, of rRNA-like sequences. This prediction also appears to be correct, and the evidence for it has once again been available for some years without its fundamental meaning being understood. Mauro and Edelman (1997) discovered that more than 20 % of eukaryotic genomes, when read in all six possible reading frames, can be related to a high degree of homology to rRNA-like sequences. Use of RNA hybridization techniques demonstrated that these homologies were also highly correlated with the production of functional proteins (Mauro and Edelman 1997). Subsequent studies verified these results in a wide range of organisms (Mignone and Pisole 2002; Kong et al. 2008). In addition, it has now been well established that Alu transposable elements (named after the *Arthrobacter luteus*

restriction endonuclease), and other short interspersed nucleotide elements (SINEs) are derived from tRNA. Alu and SINES each make up about 11–15 % of mammalian genomes (Smit and Riggs 1995; Schmitz 2012; Treangen and Salzberg 2012). In short, there is absolutely no doubt that rRNA-like and tRNA-like sequences make up a very large and quite unexpected proportion of modern genomes, but no one seems to have previously asked "Why?"

So to summarize to this point, we would like to suggest that our modular, networked, ecological approach to thinking about the origins of life clearly leads to verifiable predictions about how a self-replicating ribosome-like entity may have emerged as the basis from which cellular life with its associated DNA-encoded genomes then evolved. We further suggest that this unusual approach may have much broader implications for understanding evolutionary processes. Far from randomly exploring every possibility, evolution builds from what has already been selected, mixing and matching only those surviving modules. Because those modules have been selected for interactivity through their complementarity (which also explains their specific functions relative to each other), subsequent uses of these modules ensure some degree of future interactivity as well (Root-Bernstein and Dillon 1997; Hunding et al. 2006; Root-Bernstein 2012). Thus, molecular complementarity has its equivalents at every level of evolutionary organization, as Peter Csermely has made abundantly obvious in his book *Weak Links* (Csermely 2009), an examination of how reversible interactions characterize all living systems. We would therefore like to expand on the implications of utilizing the concept of complementary weak linkers to understand not just the emergence of the first self-replicating entities from a chemical ecosystem, but also the subsequent evolution and networking of living systems after the first cells emerged into functional ecosystems and multispecies symbiotic units such as holobionts (Rosenburg and Zilber-Rosenburg 2008; Gilbert et al. 2010; Chiu and Gilbert 2015).

In general, ecological approaches to complementarity are quite disparate and only provide vague guidance for analogies to how a ribosome might have evolved within a molecular ecology and contributed to the evolution of highly organized cellular life. This partly reflects a lack of engagement with evolutionary questions by the majority of ecologists.

From the origins of ecology as a discipline, researchers have attempted to find associations in the assortment of species and their patterns of distribution at small and large scales (Real and Brown 1991; Lomolino et al. 2004; Kissling and Schleuning 2015; Vilhena and Antonelli 2015). The idea that the structure–function link may underlie ecological units, as well as evolutionary units, has largely been ignored. Concepts such as habitat, biome, ecosystem, biogeographic region, or food web, are largely based on the ideas of pattern and scale. However, advances in functional ecology allow us to make predictions about the form and dynamics of ecological units based on structure–function interactions.

The best-known strong ecological interactions with a structure/function relationship include species–species interactions along the mutualism/commensalism/parasitism continuum such as pollination mutualisms and ant enslavement (Thrall et al. 2007). Along the facilitation/competition gradient, one also finds weaker but

similar examples such as directed seed dispersal (Wenny and Levey 1998). In these cases, the interaction of the two species, or two sets of species, shows a clear functional complementarity or coupling based on their biological structures or features. We also know that all ecosystems have a trophic structure with, at a minimum, primary producers and decomposers, and potentially up to six or seven levels of herbivores, omnivores, and carnivores. Between them, these phenomena are robust but do not explain ecological patterning.

Modularity, compartmentalization, and nestedness have emerged as key concepts within network analysis of pollination and seed dispersal mutualisms as well as predator–prey and parasite–host systems, and may also be a feature of food webs (Bascompte et al. 2003; Guimarães et al. 2007; Graham et al. 2009; Stouffer and Bascompte 2011). Nestedness here refers to the finding that specialist species' networks are subsets of those of generalist species. Modularity refers to the distribution of interactions into groups with few links between them, while compartmentalism relatedly refers to the existence of sets of species with the same interaction subnetwork. These studies reveal interesting observations about trophic and non-trophic interactions, but have not been applied to explaining ecological patterning into, e.g., habitats or ecoregions.

Functional ecology examines the distribution of functional traits within and across ecosystems to understand how ecosystem processes and ecosystem services are provided (Díaz et al. 2004; McGill et al. 2006; Violle et al. 2007). The innovation of this approach is to analyze functional traits separately from the species that have those traits. The radical implications of this theoretical dissociation of the traits from the genotypes/phenotypes generating/possessing them have hardly been discussed. Until now, ecologists have assumed, along with evolutionary theory, that the species and the individual are the primary units underlying higher orders of organization. The functional trait approach allows us to define traits that have clear functions (ecological interactions underlying ecosystem processes) and to look for patterns, regardless of their assortment into species. The common finding that both phylogenetic closeness and convergent evolution underlie many observed modules is consistent with the interpretation that functional traits underlie interaction networks, since both related and especially convergent species are likely to share many functional traits (Krasnov et al. 2012; Schleuning et al. 2014). Similarly, species attributes are correlated with modules in a pollination network (González et al. 2012).

The vast majority of functional ecology research focuses on plants in grasslands (e.g., Díaz et al. 2004; De Bello et al. 2010). Because of these research gaps, from a purely functional ecology point of view, it is difficult to predict how functional traits in general assort into networks (but see Werner and Peacor 2003 for "trait-mediated indirect interactions"). The theory of niche construction may provide a guide for what to expect. Niche construction proposes that ecological processes of, e.g., disturbance and ecosystem engineering can influence selection, thus having an effect on evolution while also transforming the environment at evolutionary timescales (Odling-Smee et al. 2003). One way in which niche construction differs from a classical account of selection is in essentially describing all selection as

co-evolution *sensu lato*. This has caused some theoretical problems, not well resolved, because it is not clear from a classical position how coevolution can occur if only some of the interactants have genes. Thus, niche construction has largely relied on the transfer of information as the way in which we can trace, e.g., the effect of a worm with a given genotype on the soil and the effect the soil has on the genotype of bacteria (Odling-Smee et al. 2013). This is largely unsatisfying for a number of reasons including the non-measurability of information in almost any relevant context. Functional ecology offers a simpler solution, which is to assume that nothing (beyond obvious cases of biomass displacement) is "passed between" the worm and the soil, but that they have a structural interaction that changes the state of one or both, a change that contributes to an ecological process. Functional ecology in turn has been resistant to include the "traits" or properties of abiotic ecosystem components in analyses, but many researchers have included these community ecology variables in any case due to their explanatory power (e.g., Maestre et al. 2012).

Niche construction networks have the following characteristics: (1) similar sets of interactions and coevolved traits/properties across similar habitats or ecoregions. Corenblit et al. (2009) argue that almost all riparian systems are characterized by a set of species traits and hydrogeomorphic properties that have coevolved through niche construction. Moreover, they relate riparian niche construction directly to the successional ecosystem process cycle. (2) Second is frequent interspersion of abiotic properties within coevolutionary interactions. Although this does not appear to have been quantified and could represent a bias of research interest in niche construction, many examples of niche construction involve a relationship between a biotic trait, an abiotic property, and another biotic trait (Odling-Smee et al. 2003). Thus, the tight evolutionary coupling of mutualisms might be unusual precisely because they are not attenuated by an abiotic intermediary.

An unresolved mystery of the functional trait approach, assuming it is as robust as it currently appears, is why functional traits, rather than species, should explain ecological functioning. This implies that beyond the evolution of the body or organism, at the ecological level, evolutionary processes disaggregate traits and assort them independently of the organismal integration of traits encoded in the DNA. Briefly, one possibility is that a hierarchical phase shift occurs in which further integration beyond the level of the body is unstable. Without trying to go into why such a phase shift should occur then and not at smaller or larger scales, it is clear that among multicellular organisms, only eusocial species achieve high levels of integration at the group level; group selection is contentious; and although multiple individuals' genotypes and phenotypes can contribute to cultural phenomena, other forms of niche construction, and epigenetic phenomena, this is far from forming a superorganism (although structure is present, e.g., Allen and Holling 2008).

This overview of ecological patterning, though not forming a unified body of theory, leads to some tentative implications for the evolution of the ribosome from a molecular ecology to highly structured and stabilized cellular life forms. First, the evolution of cellular life can be thought of in terms of niche construction. Following

the observations above of the two common features of niche-constructed networks, this leads to the following implications: Certain chemical cycles (analogous to successional and other ecological processes) could drive similar communities of prebiotic chemical ecologies toward similar niche construction outcomes, and those outcomes will be highly integrated with inorganic or non-prebiotic chemical interactions, e.g., substrates. Presumably, abiotic materials have highly different characteristic rates of chemical cycles and transformations, which may stabilize or buffer the biotic or prebiotic ecologies interacting with them. Secondly, if we take a clue from eusocial species, the only species that come close to forming a tight integration between bodies as a kind of superorganism, we note that they do this by enforcing a tight division of labor among individuals, with a single reproducing queen who creates multiple versions of herself (not actually clones of herself, but nearly clones of each other) to carry out specific tasks within a colony. Can this provide an analogy to the evolution of RNA into rRNA, mRNA, tRNA, and DNA? This again raises a question: If the weak-link-attenuated networks of niche construction are typical of "ecologies" whether chemical or organismal, but the strong structure–function integration of gene-encoded cells, tissues, organs, and bodies is typical of life, what is the difference between the two? Might organismal ecologies be in a phase of evolution equivalent to the highly distributed, compositionally linked prebiotic chemical ecology that gave rise to the ribosome, therefore foreshadowing some future integration of ecological function that will more strongly link functional trait networks (Csermely 2009)? Is some of this integration already apparent in holobionts (Rosenberg and Zilber-Rosenberg 2008; Gilbert et al. 2010; Chiu and Gilbert 2015) that integrate microbiota as obligatory symbionts with organisms such as corals, termites, ruminants, and even people?

Acknowledgements We would like to thank the US National Science Foundation for granting support and Professors Patrick F. Dillon and Adam W. Brown of Michigan State University and Professors Vic Norris and Corinne Loutellier-Bourhis of the University of Rouen, as well as two of our students, Tyler Rhinesmith and Andrew Baker, for their invaluable assistance and their feedback during the development of this research.

References

Allen CR, Holling CS (eds) (2008) Discontinuities in ecosystems and other complex systems. Columbia University Press, New York

Bada JL (2013) New insights into prebiotic chemistry from Stanley Miller's spark discharge experiments. Chem Soc Rev 42(5):2186–2196. doi:10.1039/c3cs35433d

Barthélémy RM, Grino M, Casanova JP, Faure E (2010) Ribin-like protein expression in the chaetognath Spadella cephaloptera. Int J Genet Mol Biol 2(2):20–29

Bascompte J, Jordano P, Melián C, Olesen J (2003) The nested assembly of plant–animal mutualistic networks. Proc Natl Acad Sci 100:9383–9387

Bloch DP, McArthur B, Widdowson R, Spector D, Guimaraes RC, Smith J (1983) tRNA-rRNA sequence homologies: evidence for a common evolutionary origin? J Mol Evol 19(6):420–428

Bloch D, McArthur B, Mirrop S (1984) tRNA-rRNA sequence homologies: evidence for an ancient modular format shared by tRNAs and rRNAs. BioSystems 17:209–225. doi:10.1016/0303-2647(85)90075-9

Bloch DP, McArthur B, Guimarães RC, Smith J, Staves MP (1989) tRNA-rRNA sequence matches from inter- and intraspecies comparisons suggest common origins for the two RNAs. Braz J Med Biol Res 22(8):931–944

Bonawitz ND, Chatenay-Lapointe M, Wearn CM, Shadel GS (2008) Expression of the rDNA-encoded mitochondrial protein Tar1p is stringently controlled and responds differentially to mitochondrial respiratory demand and dysfunction. Curr Genet 54(2):83–94. doi:10.1007/s00294-008-0203-0

Bowman JC, Hud NV, Williams LD (2015) The ribosome challenge to the RNA world. J Mol Evol 80(3–4):143–161. doi:10.1007/s00239-015-9669-9

Braakman R, Smith E (2013) The compositional and evolutionary logic of metabolism. Phys Biol 10:011001. doi:10.1088/1478-3975/10/1/011001

Brandman R, Brandman Y, Pande VS (2012) Sequence coevolution between RNA and protein characterized by mutual information between residue triplets. PLoS ONE 7(1):e30022. doi:10.1371/journal.pone.0030022

Caetano-Anollés G (2002) Tracing the evolution of RNA structure in ribosomes. Nucleic Acids Res 30(11):2575–2587

Caetano-Anollés D, Caetano-Anollés G (2015) Ribosomal accretion, apriorism and the phylogenetic method: a response to Petrov and Williams. Front Genet 6:194. doi:10.3389/fgene.2015.00194

Caetano-Anolles G, Seufferheld MJ (2013) The coevolutionary roots of biochemistry and cellular organization challenge the RNA world paradigm. J Mol Microbiol Biotechnol 23(1–2):152–177

Chevalier BS, Stoddard BL (2011) Homing endonucleases: structural and functional insight into the catalysts of intron/intein mobility. Nucleic Acids Res 29(18):3757–3774

Chiu L, Gilbert S (2015) The birth of the holobiont. Multi-species birthing through mutual scaffolding and niche construction. Biosemiotics 8:191–210. doi:10.1007/s12304-015-9232-5

Coelho PS, Bryan AC, Kumar A, Shadel GS, Snyder M (2002) A novel mitochondrial protein, Tar1p, is encoded on the antisense strand of the nuclear 25S rDNA. Genes Dev 16(21):2755–2760

Corenblit D, Steiger J, Gurnell AM, Naiman RJ (2009) Plants intertwine fluvial landform dynamics with ecological succession and natural selection: a niche construction perspective for riparian systems. Global Ecol Biogeogr 18(4):507–520

Crick FHC (1981) Life itself. Simon and Schuster, New York

Csermely, P (2009) Weak links. The universal key to the stability of networks and complex systems. Berlin, Springer

Dam M, Douthwaite S, Tenson T, Mankin AS (1996) Mutations in domain II of 23 S rRNA facilitate translation of a 23 S rRNA-encoded pentapeptide conferring erythromycin resistance. J Mol Biol 259(1):1–6

Darwin Charles (1859) On the origin of species. Murray, London

Dawkins Richard (1976) The selfish gene. Oxford University Press, Oxford

De Bello F, Lavorel S, Díaz S, Harrington R, Cornelissen JH, Bardgett RD, Harrison PA (2010) Towards an assessment of multiple ecosystem processes and services via functional traits. Biodivers Conserv 19(10):2873–2893

De Farias ST (2013) Suggested phylogeny of tRNAs based on the construction of ancestral sequences. J Theor Biol 335(245–248):2013. doi:10.1016/j.jtbi.06.033

De Farias ST, do Rêgo TG, José MV (2014) Evolution of transfer RNA and the origin of the translation system. Front Genet 28(5):303–313. http://dx.doi.org/10.3389/fgene.2014.00303

Díaz S, Hodson JG, Thompson K, Cabido M, Cornelissen JHC, Jalili A et al (2004) The plant traits that drive ecosystems: evidence from three continents. J Vegetation Sci 15:295–304

Dykxhoorn DM, St. Pierre R, Linn T (1996) Synthesis of the beta and beta' subunits of *Escherichia coli* RNA polymerase is autogenously regulated in vivo by both transcriptional and translational mechanisms. Mol Microbiol 19(3):483–493

Fox GE (2010) Origin and evolution of the ribosome. Persp Biol, Cold Spring Harb. doi:10.1101/cshperspect.a003483

Fukuda R, Taketo M, Ishihama A (1978) Autogenous regulation of RNA polymerase beta subunit synthesis in vitro. J Biol Chem 253(13):4501–4504

Galadino R, Botta G, Pino S, Costanzo G, DiMauro E (2012) Genetics first or metabolism first? The formamide clue. Chem Soc Rev 41(16):5526–5565

Galopier A, Hermann-Le Denmat S (2011) Mitochondria of the yeasts Saccharomyces cerevisiae and Kluyveromyces lactis contain nuclear rDNA-encoded proteins. PLoS ONE 6(1):e16325. doi:10.1371/journal.pone.0016325

Gilbert SF, McDonald E, Boyle N, Buttino N, Gyi L, Mai M, Prakash N, Robinson J (2010) Symbiosis as a source of selectable epigenetic variation: taking the heat for the big guy. Phil Trans Roy Soc 365:671–678

Gimautdinova OI, Karpova GG, Knorre DG, Kobetz ND (1981) The proteins of the messenger RNA binding site of *Escherichia coli* ribosomes. Nucl Acids Res 9(14):3465–3481

González A, Allesina S, Rodrigo A, Bosch J (2012) Drivers of compartmentalization in a Mediterranean pollination network. Oikos 121:2001–2013

Graham S, Hassan H, Burkett-Cadena N, Guyer C, Unnasch T (2009) Nestedness of ectoparasite-vertebrate host networks. PLoS ONE 4

Gregory RJ, Cahill PB, Thurlow DL, Zimmermann RA (1988) Interaction of *Escherichia coli* ribosomal protein S8 with its binding sites in ribosomal RNA and messenger RNA. J Mol Biol 204(2):295–307

Gross R, Fouxon I, Lancet D, Markovitch O (2014) Quasispecies in population of compositional assemblies. BMC Evol Biol 14:265

Guimarães Jr P, Rico-Gray V, Oliveira P, Izzo T, Reis S, Thompson J (2007) Interaction intimacy affects structure and coevolutionary dynamics in mutualistic networks. Curr Biol 17

Hoyle F (1983) Intelligent universe. Holt, Rinehart and Winston, New York

Hunding A, Kepes F, Lancet D, Minsky A, Norris V, Raine D, Sriram K, Root-Bernstein R (2006) Compositional complementarity and prebiotic ecology in the origin of life. BioEssays 28 (4):399–412

Jacobson H (1955) Information, reproduction and the origin of life. Am Scientist 43:119–127

Johnson AP, Cleaves HJ, Dworkin JP, Glavin DP, Lazcano A, Bada JL (2008) The Miller volcanic spark discharge experiment. Science 322(5900):404. doi:10.1126/science.1161527

Kermekchiev M, Ivanova L (2001) Ribin, a protein encoded by a message complementary to rRNA, modulates ribosomal transcription and cell proliferation. Mol Cell Biol 21(24):8255–8263

Kissling WD, Schleuning M (2015) Multispecies interactions across trophic levels at macroscales: retrospective and future directions. Ecography 38(4): 346–357

Kong Q, Stockinger MP, Chang Y, Tashiro H, Lin CL (2008) The presence of rRNA sequences in polyadenylated RNA and its potential functions. Biotechnol J 3(8):1041–1046. doi:10.1002/biot.200800122

Krasnov B, Fortuna M, Mouillot D, Khokhlova I, Shenbrot G, Poulin R (2012) Phylogenetic signal in module composition and species connectivity in compartmentalized host-parasite networks. Am Naturalist 179:501–511

Lomolino MV, Sax DF, Brown JH (2004) Foundations of biogeography: classic papers with commentaries. University of Chicago Press, Chicago

Maestre FT, Quero JL, Gotelli NJ, Escudero A, Ochoa V, Delgado-Baquerizo M, Val J (2012) Plant species richness and ecosystem multifunctionality in global drylands. Science 335 (6065):214–218

Mann S (2012) Systems of creation: the emergence of life from nonliving matter. Acc Chem Res 45:2131–2141

Mauro VP, Edelman GM (1997) rRNA-like sequences occur in diverse primary transcripts: implications for the control of gene expression. Proc Natl Acad Sci U.S.A. 94(2):422–427
McGill BJ, Enquist BJ, Weiher E, Westoby M (2006) Rebuilding community ecology from functional traits. Trends Ecol Evol 21(4):178–185
Mignone F, Pesole G (2002) rRNA-like sequences in human mRNAs. Appl Bioinform 1(3): 145–154
Miller SL (1953) Production of amino acids under possible primitive earth conditions. Science 117(3046):528–529. doi:10.1126/science.117.3046.528
Mushegian A (2005) Protein content of minimal and ancestral ribosome. RNA 11:1400–1406
Mushegian A (2008) Gene content of LUCA, the last universal common ancestor. Front Biosci 13:4657–4666. http://www.ncbi.nlm.nih.gov/pubmed/18508537
Neveu M, Kim HJ, Benner SA (2013) The "strong" RNA world hypothesis: fifty years old. Astrobiology 13(4):391–403
Nomura M, Yates JL, Dean D, Post LE (1980) Feedback regulation of ribosomal protein gene expression in *Escherichia coli*: structural homology of ribosomal RNA and ribosomal protein mRNA. Proc Natl Acad Sci U.S.A. 77:7084–7088
Norris V, Loutelier-Bourhis C, Thierry A (2012) How did metabolism and genetic replication get married? Orig Life Evol Biosph 42(5):487–495
Odling-Smee FJ, Laland KN, Feldman MW (2003) Niche construction: the neglected process in evolution. Princeton University Press, Princeton, NJ
Odling-Smee FJ, Erwin DH, Palkovacs EP, Feldman MW, Laland K (2013) Niche construction theory: a practical guide for ecologists. Q Rev Biol 88(1):3–28
Olins PO, Nomura M (1981) Translational regulation by ribosomal protein S8 in *Escherichia coli*: structural homology between rRNA binding site and feedback target on mRNA. Nucleic Acids Res 9(7):1757–1764
Paley W (1824) Natural theology. S. King, New York
Passador L, Linn T (1992) An internal region of rpoB is required for autogenous translational regulation of the beta subunit of *Escherichia coli* RNA polymerase. J Bacteriol 174(22): 7174–7179
Real LA, Brown JH (eds) (1991) Foundations of ecology: classic papers with commentaries. University of Chicago Press, Chicago
Robert F, Brakier-Gingras L (2001) Ribosomal protein S7 from *Escherichia coli* uses the same determinants to bind 16S ribosomal RNA and its messenger RNA. Nucleic Acids Res 29 (3):677–682
Root-Bernstein R (2012) A modular hierarchy-based theory of the chemical origins of life based on molecular complementarity. Acc Chem Res 45(12):2169–2177. doi:10.1021/ar200209k; http://pubs.acs.org/toc/achre4/45/12
Root-Bernstein RS, Dillon PF (1997) Molecular complementarity 1: The molecular complementarity theory of the origin and evolution of life. J Theor Biol 188:447–479
Root-Bernstein M, Root-Bernstein R (2015) The ribosome as a missing link in the evolution of life. J Theor Biol 367:130–158. doi:10.1016/j.jtbi.2014.11.025
Root-Bernstein RS, Root-Bernstein MM (2016) The ribosome as a missing link in prebiotic evolution II: ribosomes encode ribosomal proteins that bind to common regions of their own mRNAs and rRNAs. J Theor Biol 397:115–127
Rosenburg E, Zilber-Rosenburg I (2008) From bacterial bleaching to the hologenome theory of evolution. Proceedings of 11th annual coral reef symposium session, vol 9. pp 269–273
Roth KM, Wolf MK, Rossi M, Butler JS (2005) The nuclear exosome contributes to autogenous control of NAB2 mRNA levels. Mol Cell Biol 25:1577–1585
Schleuning M, Ingmann L, Strauß R, Fritz S, Dalsgaard B, Dehling M, Plein M, Saavedra F, Sandel B, Svenning J, Böhning-Gaese K, Dormann C (2014) Ecological, historical and evolutionary determinants of modularity in weighted seed-dispersal networks. Ecol Lett 17:454–463
Schmitz J (2012) SINEs as driving forces in genome evolution. Genome Dyn. 7:92–107. doi: 10.1159/000337117

Simon H (1969) The sciences of the artificial. M.I.T. Press, Cambridge, M. A
Smit AF, Riggs AD (1995) MIRs are classic, tRNA-derived SINEs that amplified before the mammalian radiation. Nucleic Acids Res. 23(1):98–102
Steinmetz EJ, Conrad NK, Brow DA, Corden JL (2001) RNA-binding protein Nrd1 directs poly (A)-independent 3'-end formation of RNA polymerase II transcripts. Nature 413:327–331
Steward KL, Linn T (1992) Transcription frequency modulates the efficiency of an attenuator preceding the rpoBC RNA polymerase genes of *Escherichia coli*: possible autogenous control. Nucleic Acids Res 20(18):4773–4779
Stouffer D, Bascompte J (2011) Compartmentalization increases food-web persistence. Proc Natl Acad Sci U.S.A. 108:3648–3652
Stouffer DB, Sales-Pardo M, Leicht EA, Newman M (2010) Origin of compartmentalization in food webs. Ecology 91(10):2941–2951
Strobel SA (2001) Repopulating the RNA world. Nature 411:1003–1006
Tenson T, Mankin A (1995) Comparison of functional peptide encoded in the *Escherichia coli* 23S rRNA with other peptides involved in cis-regulation of translation. Biochem Cell Biol 73(11–12):1061–1070
Tenson T, DeBlasio A, Mankin A (1996) A functional peptide encoded in the Escherichia coli 23S rRNA Proc. Natl Acad Sci USA 93:5641–5646
Thrall P, Hochberg M, Burdon J, Bever J (2007) Coevolution of symbiotic mutualists and parasites in a community context. Trends Ecol Evol 22:120–126
Treangen TJ, Salzberg SL (2012) Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nat Rev Genet 13:36–46. doi:10.1038/nrg3117
Van Gemen B, Twisk J, van Knippenberg PH (1989) Autogenous regulation of the *Escherichia coli* ksgA gene at the level of translation. J Bacteriol 171(7):4002–4008
Vilhena D, Antonelli A (2015) A network approach for identifying and delimiting biogeographical regions. Nature Commun 6:6848
Violle C, Navas ML, Vile D, Kazakou E, Fortunel C, Hummel I, Garnier E (2007) Let the concept of trait be functional! Oikos 116(5):882–892
Wang J, Dasgupta I, Fox GE (2009) Many nonuniversal archaeal ribosomal proteins are found in conserved gene clusters. Archaea 2(4):241–251
Wenny DG, Levey DJ (1998) Seed dispersal by bellbirds in a tropical cloud forest. Proc Nat Acad Sci U.S.A. 95(11):6204–6207
Werner E, Peacor S (2003) A review of trait-mediated indirect interactions in ecological communities. Ecology 84:1083–1100

Part IV
Methods

Chapter 20
Inference Methods for Multiple Merger Coalescents

Bjarki Eldon

Abstract Some populations, including a diverse group of marine populations such as Pacific oysters and Atlantic cod, are highly fecund. Models of high fecundity—coupled with a skewed offspring distribution—have coalescent processes, which admit (simultaneous) multiple mergers of ancestral lineages associated with them. In contrast, the celebrated and extensively employed Kingman's coalescent only admits pairwise mergers of ancestral lineages. We review multiple merger coalescent models derived from population models, which admit high fecundity and skewed offspring distribution. Inference methods that have been developed based on these multiple merger coalescent models will also be reviewed. In fact, multiple merger coalescent models are able to predict the excess singletons (relative to Kingman's coalescent predictions) observed in the commercially important Atlantic cod. These models may be applicable to a wide range of natural populations—including a diverse group of marine organisms, viruses, and plants which distribute seeds—with significant implications.

20.1 Introduction

The systematic study of genetic variation among individuals based on molecular data began with Harris (1966), Hubby and Lewontin (1966), Lewontin and Hubby (1966), who studied variation of allozymes in humans (Harris 1966), and in *Drosophila pseudoobscura* (Hubby and Lewontin 1966; Lewontin and Hubby 1966). However, as pointed out by Kreitman (1983), variation in allozymes represent phenotypes, not genotypes, and a deeper study of genetic variation in natural populations naturally requires genetic data. Once DNA sequencing technology had become available, Kreitman (1983) sequenced the *Adh* gene (alcohol dehydrogenase) in eleven *D. melanogaster* lines and found considerable amount of silent

B. Eldon (✉)
Museum Fuer Naturkunde, Leibniz Institut Fuer Evolutions- Und Biodiversitaetsforschung, Berlin, Germany
e-mail: bjarki.eldon@mfn-berlin.de

P. Pontarotti (ed.), *Evolutionary Biology*, DOI 10.1007/978-3-319-41324-2_20

polymorphism. The works of Harris (1966), Hubby and Lewontin (1966), Lewontin and Hubby (1966), Kreitman (1983) on the molecular side, and those of Ewens (1972), Karlin and McGregor (1972), Watterson (1975), Kingman (1982a, b, c), Hudson (1983), Tajima (1983) on the theoretical side, initiated an important shift in focus in population genetics, namely a shift to ancestral—or coalescent—processes, properties of gene genealogies and inference based on a sample. In fact, Kreitman (1983), Kingman (1982a, b, c), Hudson (1983) and Tajima (1983) were all published around the same time (1982–83). Kingman (2000) reviews the ideas that eventually led to the Kingman coalescent.

The coalescent approach is concerned with describing the random ancestral relations of gene copies sampled from a natural population. In biological applications, a coalescent process is often derived from a mathematically tractable population model, which describes how genetic information is passed on between successive generations.

Many natural populations may be characterised as low-fecundity populations, in which each individual contributes only a very small number of offspring, relative to the total population size. Indeed, bacteria are a prime example, in which each individual cell contributes only two offspring in a single reproduction event. The classical Wright–Fisher and Moran population models of how genetic information is passed on from parents to offspring seem well suited to model low-fecundity populations. However, as reviewed by Hedgecock and Pudovkin (2011) in relation to marine populations, some are highly fecund. In fact, a single Atlantic cod female may lay millions of eggs during spawning (May 1967; Oosthuizen and Daan 1974). The Pacific oyster *Crassostrea gigas* may also lay millions of eggs during reproduction (Li and Hedgecock 1998). This high fecundity counteracts the hostile environment in which larvae must survive, and in which most likely perish (a Type III survivorship curve). Occasionally, though, a huge number of larvae derived from few parents may survive. This (potential) skewness, coupled with the high fecundity, is not captured by the classical population models. Population models of high fecundity and skewed offspring distributions, and associated (Fleming-Viot measure-valued) diffusions and multiple merger coalescent processes, have received considerable attention by mathematicians [cf. e.g. Berestycki (2009), Birkner and Blath (2009), Etheridge (2011)]; but relatively less attention by biologists (however, see Hedgecock and Pudovkin (2011), Tellier and Lemaire (2014)).

We briefly review multiple merger coalescent processes, some of the population models, which have these coalescent processes as their ancestral limit processes, and inference methods derived from stated coalescent processes. The mathematical theory and inference methods of multiple merger coalescent processes are in active development. Hence, we do not aim at a comprehensive review, but rather at arguing that multiple merger coalescent processes are highly relevant for the study of genetic diversity.

20.2 The Classical Kingman's Coalescent

In order to learn about the evolutionary history of natural populations, and to find out which genes may be under some kind of natural selection, we sample individuals at random from the population under investigation, sequence part(s) of their genome and compare the genetic variation we observe in the sample to predictions of our model. The celebrated Kingman's coalescent (Kingman 1982a, b, c) has been a cornerstone of inference in population genetics since its introduction. A coalescent process describes the random ancestral relations of a sample of gene copies (or DNA sequences) taken from a natural population. Inference methods can then be developed based on a given coalescent model. A key to the success of the coalescent approach is that it is sample based. Thus, one can compare genetic variation

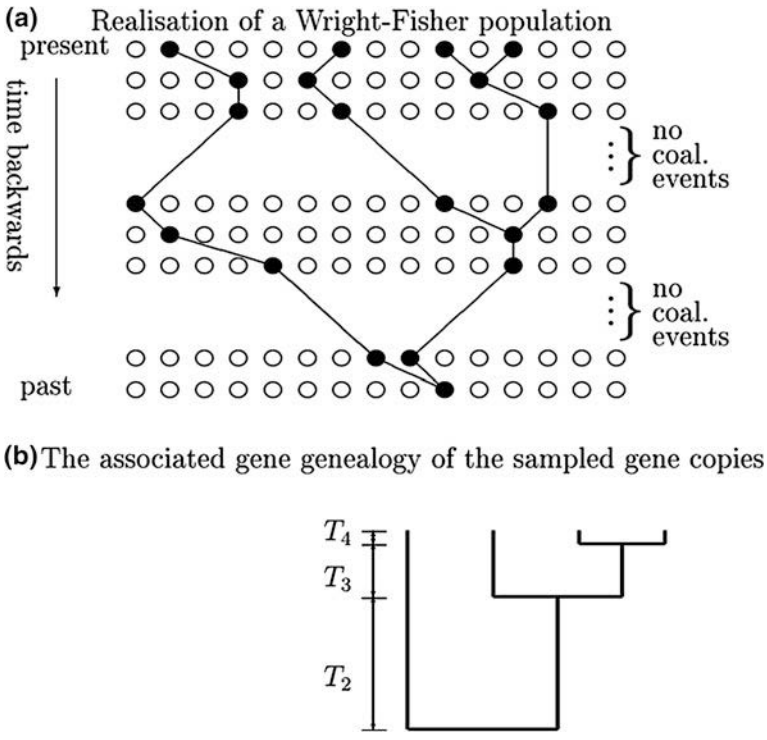


Fig. 20.1 Example of the ancestral relation of 4 sampled gene copies. In **a**, the ancestral lines of the 4 sampled genes are traced back in time through each generation of a haploid Wright–Fisher population. The *closed circles* represent the sampled gene copies and the ancestral ones. The *vertical dotted lines* represent generations in which coalescence events between the ancestral gene copies *do not* occur. In **b**, the corresponding gene genealogy of the four sampled gene copies is shown, with times between coalescence events denoted by T_4, T_3 and T_2; the T_j are independent exponentials with rate $\binom{j}{2}$

observed in a sample to predictions based on a given coalescent model. A very readable basic description of the theory behind the Kingman coalescent can be found in Wakeley (2007). Figure 20.1 illustrates how the Kingman coalescent arises from a low-fecundity population model—for example, the haploid Wright–Fisher model—of size N. One can think of Wright–Fisher reproduction as one in which each individual formes a huge number of viable offspring. The next generation of individuals is then formed by multinomial sampling from the pool of viable offspring. This is equivalent to forming the relations between parents and offspring by allowing the offspring to choose their parents independently and uniformly at random (for clarity, relations between parents and offspring are only shown for the sampled gene copies in Figure 20.1a). The probability that a given group of k offspring have the same parent is $1/N^{k-1}$. Thus, in a large population, large offspring numbers are quite unlikely. The gene genealogy of the sampled gene copies is shown in Figure 20.1b serves to illustrate how much the coalescent approach simplifies computations. Instead of first simulating the ancestral relations of a large population forwards in time, and then extracting the gene genealogy of the sampled gene copies, one can equivalently, and much more efficiently, simulate only the gene genealogy of the sampled gene copies. Indeed, Kelleher et al. (2015) have recently made significant improvements in simulating gene genealogies of whole chromosomes with recombination, even for large sample sizes.

20.3 Multiple Merger Coalescent Models

Coalescent processes which admit (simultaneous) multiple mergers of ancestral lineages can be derived from population models of *h*igh *f*ecundity and *s*kewed *o*ffspring *d*istribution HFSOD (Sagitov 1999, 2003; Möhle and Sagitov 2001, 2003; Schweinsberg 2003; Sargsyan and Wakeley 2008; Eldon and Wakeley 2006; Birkner et al. 2013). The classical Wright–Fisher and Moran models do not allow for HFSOD. Before we review population models, which allow for HFSOD, we introduce multiple merger coalescent processes.

20.3.1 Λ-Coalescents

Pitman (1999), Sagitov (1999), Donnelly and Kurtz (1999) independently introduce Λ-coalescents. Coalescent processes, which allow any number of ancestral lineages to merge to a common ancestor each time, but only one such merger occurs each time, are referred to as Λ-coalescents. Sagitov (1999) derives a Λ-coalescent in general form from a population model, which admits HFSOD. Pitman (1999) derives Λ-coalescents from a Poisson process construction, and Donnelly and Kurtz (1999) show that Λ-coalescents can be derived from their particle representation

(look-down construction) of an infinite population. In the most general form, the rate at which a given group of k lineages merges is given by Pitman (1999), Sagitov (1999), Donnelly and Kurtz (1999)

$$\lambda_{b,k} = \int_0^1 x^k (1-x)^{b-k} x^{-2} \Lambda(\mathrm{d}x) \tag{20.1}$$

where $\Lambda(\mathrm{d}x)$ denotes a finite measure on (Borel subsets of) the unit interval ($[0, 1]$). One can think of the rate (20.1) as x representing the random fraction of the population replaced by the offspring of a single individual, and $\Lambda(\mathrm{d}x)$ as the probability distribution associated with the stated fraction. In applying (20.1) to actual populations, one would need specific forms of $\Lambda(\mathrm{d}x)$ derived from population models, which admit HFSOD.

Figure 20.2 shows an example of the gene genealogy of four sampled gene copies in a HFSOD population. In Figure 20.2, time is counted backwards, and we assume a large population size. In generation 1, one individual contributes huge number of offspring to generation 0, to which three of the sampled lineages belong.

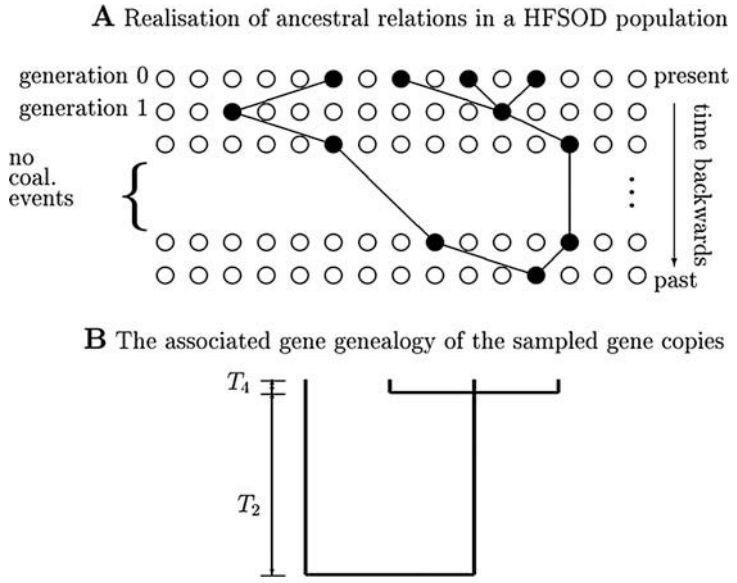


Fig. 20.2 Example of the ancestral relations of 4 sampled gene copies in a haploid HFSOD population. In **a**, the ancestral lines of the 4 sampled genes are traced back in time through each generation of the population. The *closed circles* represent the sampled gene copies and the ancestral ones. The *vertical dotted lines* represent generations in which coalescence events between the ancestral gene copies *do not* occur. In **b**, the corresponding gene genealogy of the four sampled gene copies is shown, with times between coalescence events denoted by T_4 and T_2. The times T_j are independent exponentials with rate determined by (20.1)

Thus, we see a 3-merger in generation 1. The associated gene genealogy (Fig. 20.2b) never displays 3 ancestral lineages due to the multiple merger; i.e. the time $T_3 = 0$ (almost surely). As in Figure 20.1a, we do not, for clarity, show all the relations between parents and offspring. Gene genealogies associated with Λ-coalescents can be efficiently simulated (cf. e.g. Birkner and Blath 2008). Indeed, one could also (quite likely) adapt the algorithm of Kelleher et al. (2015) to multiple merger coalescents.

20.3.2 Ξ-Coalescents

Schweinsberg (2000) and Mohle and Sagitov (2001) introduced *simultaneous multiple merger coalescents* (Ξ-coalescents), in which many distinct multiple mergers may occur at the same time (Fig. 20.3). In Figure 20.3a, a large offspring number event occurs in generation 1, in which two ancestral gene copies are parents to *at least* the ancestral lineages as shown. The two 'lucky' parent gene copies (chromosomes) could belong to the same diploid parent pair (see Fig. 20.4). Ξ-coalescents are obtained from models of repeated severe bottlenecks (Birkner et al. 2009), selection (Durrett and Schweinsberg 2004, 2005) and polyploidy in which selfing is excluded (Birkner et al. 2013a; Blath et al. 2016). We refrain from stating the general coalescence rate of a Ξ-coalescent, as it needs a bit of notation, and refer to Schweinsberg (2000) for details (see also Blath et al. (2016) for a precise description of Kingman-, Λ- and Ξ-coalescents). An example of the coalescent rate when associated with polyploidy (Blath et al. 2016) is given below. Ξ-coalescents are the most general form of time-homogeneous coalescent processes; they include both the Λ-coalescent and the Kingman's coalescent as special cases. In Figure 20.3, which gives an example of a gene genealogy of six sampled gene copies, the first merger is a simultaneous merger of two groups of three ancestral lineages each. Even though all the lineages are involved in the first merger event, we still have two ancestral lineages left.

20.3.3 Multiple Merger Coalescent Processes and Selection

The success of the coalescent approach as a modelling and inference approach is that one can 'separate' the (selectively neutral) mutation process from the coalescent process, which generates the gene genealogy; i.e. the topology of the gene genealogy is statistically independent from the mutation process (usually modelled as a Poisson process on a given gene genealogy). Thus, in simulations, one usually constructs a gene genealogy (as in Figs. 20.1b, 20.2b and 20.3) and then generates simulated genetic variation by applying a mutation process to the given gene genealogy. When modelling selection, things are not quite as simple (cf. e.g. Krone and Neuhauser 1997; Neuhauser and Krone 1997; Etheridge and Griffiths 2009;

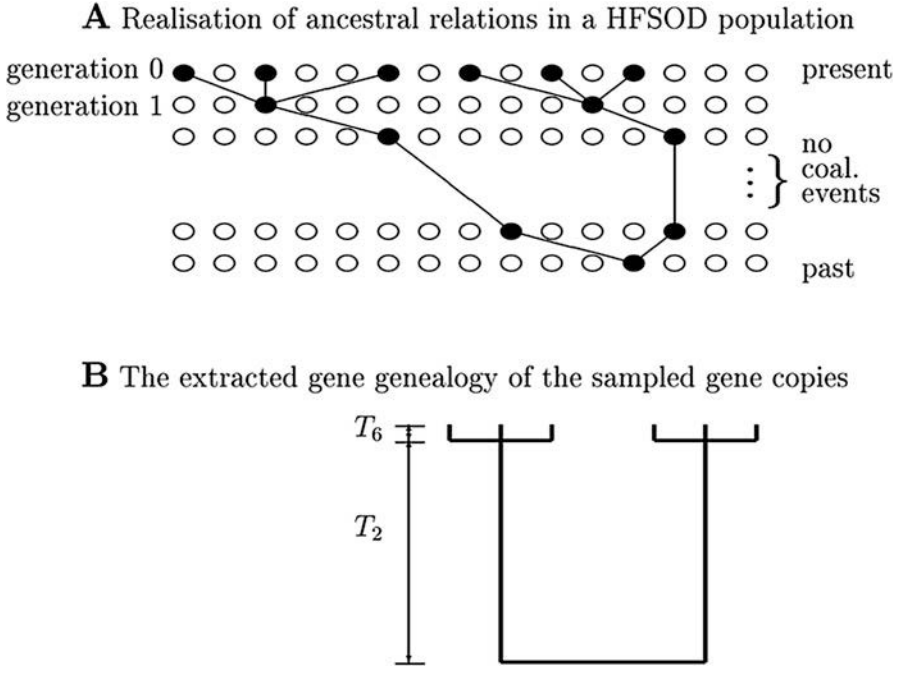


Fig. 20.3 Example of a gene genealogy associated with a Ξ-coalescent (sample size $n = 6$). Times between coalescence events are denoted by T_6 and T_2; they are independent exponentials with rate determined by the associated Ξ-coalescent. In **a**, a large offspring number event occurs in generation 1, involving all the ancestral lineages as shown. Only the parent–offspring relations associated with the sampled gene copies are shown. The associated gene genealogy is shown in **b**

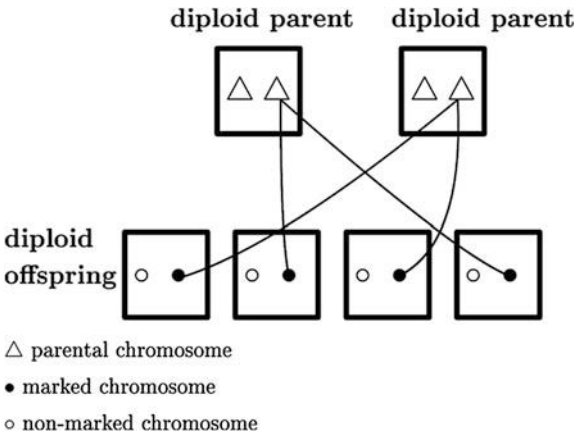


Fig. 20.4 An illustration of how a simultaneous merger occurs in a diploid population. The parents have at least 4 diploid offspring, and each offspring carries a 'marked' chromosome (*closed circles*), or a chromosome ancestral to (some of) the sampled chromosomes. Since the marked chromosomes share parental chromosomes (Δ) as shown, a simultaneous merger—two simultaneous pairwise mergers—occurs. The origin of the non-marked chromosomes (open circles) is not shown for clarity

Etheridge et al. 2010), since then we need to model genetic types which have different propensity to survive and/or reproduce. Etheridge et al. (2010), in particular, derive a coalescent process for a locus under viability selection, modelled on continuous-time Moran model of high fecundity. Individuals contribute viable offspring irrespective of their genetic type. However, the offspring, which inherit the type of their parent, have different survival probabilities, which depend on their type. Ancestral selection graphs (ASG) are different from graphs associated with selectively neutral loci since the propensity to survive and/or reproduce depends on the genetic type of each individual. This type dependence introduces branching events in ancestral selection graphs. They are therefore more difficult to (efficiently) simulate than selectively neutral graphs.

Even though the interaction of selection and skewed offspring distribution is of great interest, the study of this important topic is still in its infancy. Der et al. (2011) consider 'generalised' Wright–Fisher (GWF) models with types, in which the first two moments of the type frequency matches that of the classical Wright–Fisher (up to a constant for the variance), but the higher moments may differ from the Wright–Fisher model. This corresponds to shifting the transition probability distributions of the type frequencies. Der et al. (2011) study examples of GWF models, which correspond to skewed offspring distribution, the interaction of selection and skewed offspring distribution (or the associated drift) and show that these particular forms of drift may greatly amplify the effect of selection. Foucart (2013) considers the evolution of the frequency of a deleterious allele in a population with skewed offspring distribution and shows that under certain conditions, the deleterious allele will vanish from the population (with probability 1). In contrast, in a classical

Wright–Fisher population, a (weakly) deleterious allele always has a chance of surviving. One could also ask if skewed offspring distribution should be modelled as a neutral process, or as selection, possibly some kind of ephemeral selection.

Multiple mergers of ancestral lineages also occur when sample size is large relative to the effective population size. To give an idea of how this would work in a haploid Wright–Fisher population, assume we have some fixed number n of balls, and we throw these balls, one by one, uniformly at random into N boxes. The balls are offspring gene copies, and they are choosing their parents independently and uniformly at random according to Wright–Fisher sampling. If $n \ll N$, we will occasionally see a box with two balls (coalescence of two gene copies), but almost never a box with three or more balls (multiple merger). Now assume $n = O(N)$ (for example, $n = N/4$). Now we are very likely to see a number of boxes with at least two balls (simultaneous mergers), since we violate the assumption $n \ll N$.

20.3.4 Large Sample Size

In deriving a coalescent process from a population model, one usually assumes that sample size is small relative to the population size (Sagitov 1999; Mohle and Sagitov 2001, 2003). Another key mathematical assumption is that population size can be arbitrarily large. However, with recent advances in DNA sequencing technology, it is becoming increasingly feasible to sequence thousands of individuals (e.g. Nelson et al. 2012; Flannick et al. 2014). Such large sample sizes may be on the order of the effective population size, thus violating the assumption of small sample size relative to effective population size. An effective population size is the population size of a hypothetical (usually Wright–Fisher) population, which carries the same amount of genetic variation as observed in a given sample.

Wakeley and Takahashi (2003) consider the case when sample size is on the order, or even larger, than the effective size, and give evidence that the expected amount of singletons (mutations observed in one copy in a sample) increases with relative sample size (relative to effective size). Singletons necessarily occur in ancestral lineages before they coalesce, and hence, (most of) these mutations would be very recent. This effect of large sample size would probably be enhanced if the sample was drawn from a HFSOD population, i.e. a population with huge census size but relatively small effective size. By way of an extreme example, if one individual contributed all the offspring, then they would all carry singleton mutations (if we assume the infinitely many sites mutation model), and so our sample (really no matter how large in this case) would consist only of singletons.

Fu (2006) derives a continuous-time 'exact' coalescent for the haploid Wright–Fisher model (but still requires the condition $N(N-1)\ldots(N-n+1)N^{-n} \gg 0$ to hold, where n is sample size and N is population size) and compares the exact coalescent with the classical Kingman coalescent. The Kingman coalescent, when based on the Wright–Fisher model, is a good approximation when sample size

$n < \sqrt{2N}$, where N denotes the size of the Wright–Fisher population. When sample size becomes large enough, the exact coalescent admits multiple mergers (Fu 2006). Bhaskar et al. (2014) consider the impact of large sample size in a growing population, and also observe that the expected number of singletons increases with sample size. Wakeley and Takahashi (2003), Fu (2006) and Bhaskar et al. (2014) all remark that multiple mergers will be observed more frequently in the gene genealogy as sample size increases.

20.3.5 Single-locus HFSOD Models

Now we consider population models, which admit HFSOD, and which have multiple merger coalescents as their limiting (as population size tends to infinity) ancestral processes. Mohle and Sagitov (2001, 2003) give general conditions for the offspring distribution under which multiple merger coalescent processes are obtained from haploid (Mohle and Sagitov 2001) or diploid (Möhle and Sagitov 2003) population models.

20.3.5.1 The Model by Schweinsberg (2003)

Most HFSOD models are single-locus models. Schweinsberg (2003) considers the following haploid discrete-generations model of fixed size N. In each generation, each individual independently contributes a random number X of viable offspring to a common pool of offspring. From this pool, the subsequent generation of N individuals is formed by sampling without replacement. The skewness is modelled in the following way (Schweinsberg 2003):

$$\mathbb{P}(X \geq k) \sim Ck^{-\alpha}; \quad C, \alpha > 0. \tag{20.2}$$

Equation (20.2) specifies how the probability of having *at least* k offspring decays as k increases. The symbol '$\sim$' means that $\mathbb{P}(X \geq k)$ is approximated by $Ck^{-\alpha}$ only for very large k (in fact, as $k \to \infty$). To ensure that the pool of viable offspring is at least of size N, one requires that each individual contributes, on average, more than one viable offspring (i.e. $\mathbb{E}[X] > 1$). The constant C is simply a normalising constant; the key parameter in model (20.2) is α, as it dictates the magnitude of the skewness.

If $\alpha \geq 2$, the associated coalescent process is the classical Kingman coalescent. In the case $\alpha \geq 2$, the probability of having very many offspring decays sufficiently rapidly that in the limit of a large population size ($N \to \infty$), such events become negligible. If $1 \leq \alpha < 2$, a multiple merger coalescent [specific version of a Λ-coalescent (20.1)] is obtained. Indeed, if $1 \leq \alpha < 2$, when we have $b \geq 2$ ancestral

lines in total, the rate at which a given group of k ancestral lines coalesce is given by Schweinsberg (2003)

$$\lambda_{b,k} = \frac{B(k-\alpha, b-k+\alpha)}{B(2-\alpha,\alpha)}, \quad 2 \le k \le b; \tag{20.3}$$

where $B(\cdot,\cdot)$ is the beta function. The coalescent with rate (20.3) is a specific example of a Λ-coalescent, and we will refer to it as the *beta-coalescent*. It has received some attention among mathematicians (Birkner et al. 2005; Berestycki et al. 2007, 2008; Kersting 2012; Dahmer et al. 2014; and Etheridge 2011) and has been applied to population genetic data on Atlantic cod (Birkner and Blath 2008; Birkner et al. 2013b, c; and Árnason and Halldórsdóttir 2015).

Equation (20.3) also holds when $\alpha = 1$, in which case, the coalescent process is referred to as the *Bolthausen–Sznitman coalescent*, which was first studied by Bolthausen and Sznitman (1998). Finally, when $0 < \alpha < 1$, a discrete-time Ξ-coalescent is obtained (Schweinsberg 2003). However, discrete-time coalescent processes may not be biologically realistic, since coalescence events likely occur too fast for mutations to generate genetic variation. We will return to this point in the section on **timescales**.

20.3.5.2 The Model by Eldon and Wakeley (2006)

In the classical Moran model, a single individual (the parent) contributes exactly one new offspring in each reproduction event. Eldon and Wakeley (2006) consider a discrete-time modified Moran model as follows. In each reproduction event, the parent contributes 1 offspring with probability $1 - \varepsilon_N$. With probability ε_N, the number of offspring is a fixed fraction ψ $(0 < \psi < 1)$ of the total population. In other words—with probability ε_N, the number of offspring is on the order of the population size. In order to keep the population size constant, the same number of individuals (excluding the offspring) must perish. Eldon and Wakeley (2006) let $\varepsilon_N = N^{-\gamma}$ for some parameter $\gamma > 0$. If $\gamma > 2$, the classical Kingman coalescent results. If $\gamma \le 2$, the coalescent process admits multiple mergers. Indeed, if $1 < \gamma < 2$, a given group of k ancestral lines, when we have b active ancestral lines, merge with rate (Eldon and Wakeley 2006)

$$\lambda_{b,k} = \psi^k (1-\psi)^{b-k}, \quad 2 \le k \le b. \tag{20.4}$$

The model by Eldon and Wakeley (2006) is certainly a simplification, since it assumes that exactly the same fraction of the population is replaced in each reproduction event. It is possible to allow ψ to be random, but then one is left with the problem of finding the probability distribution for ψ appropriate for the population under consideration.

20.3.5.3 Other HFSOD Models

Sargsyan and Wakeley (2008) consider a haploid model of fixed size N in which, with probability $1 - \varepsilon_N$, a single individual is replaced by one offspring. With probability ε_N, a random number X_N of adults is replaced by the same number of offspring of a random number Y_N of parents. Under different general conditions on the offspring distribution, Kingman's coalescent, Λ-coalescents or Ξ-coalescents are obtained (see Sargsyan and Wakeley 2008 for details). Sargsyan and Wakeley (2008) do not consider specific forms for the offspring distribution, but give simulation algorithms to sample gene genealogies from the different processes.

Huillet (2014) considers models related to the model of Schweinsberg (2003) in which the offspring distribution follows Pareto distribution, i.e. where $\mathbb{P}(X_i > x) = x^{-\alpha}$, $\alpha > 0$, and $x \geq 1$. If a certain size biasing of the offspring numbers is involved, which depends on an additional parameter β, multiple merger (Λ)-coalescent processes are obtained (Huillet 2014) with rates, which depend on both α and β. Huillet and Mohle (2013) consider an extended Moran model (only one parent in each reproduction event), give conditions on the offspring distribution under which continuous-time Λ-coalescents are obtained and give examples. Mohle (2011), Huillet and Mohle (2011) also consider certain classes of HFSOD models, which can yield multiple merger coalescent processes. However, the biological interpretation of the models considered by Mohle (2011) and Huillet and Mohle (2011) seems, at first sight, less clear.

20.3.6 Timescales

When a coalescent process is derived from a given discrete-time population model, the time is rescaled (or 'speeded up') in order to obtain convergence to a continuous-time process. We can see why this makes sense from Figure 20.1. As the ancestral lines (the black circles) become fewer, the (expected) time between coalescence events increases. It may take many generations (in a large population) until the last two lines coalesce. It is therefore much more convenient to work with rescaled time. It is natural (Sagitov 1999; Mohle and Sagitov 2001) to rescale time with the probability that two distinct individuals, sampled at the same time, derive from the same parent present in the previous generation. This probability is usually denoted by c_N. The probability c_N can be very different for different population models, even though the models lead to the same coalescent process.

By way of example, the haploid discrete-time Wright–Fisher and Moran population models both have the Kingman coalescent as their ancestral process. However, $c_N = 1/N$ when associated with the haploid Wright–Fisher model, while $c_N = 1/\binom{N}{2}$ when associated with the discrete-time Moran model. Indeed, the Wright–Fisher and Moran models are only two examples of a large class of

Cannings (1974) exchangeable population models, all of which have the Kingman coalescent as their ancestral process (Mohle and Sagitov 2001). The same is likely true for population models, which admit HFSOD and lead to multiple merger coalescent processes, even though specific examples are not yet available.

The issue of timescale is important for inference. Eldon and Wakeley (2006) show that for some population models, the probability c_N can be too high for mutations to accumulate. In other words, coalescence events happen so quickly that mutation has no chance of generating variation between individuals. Eldon et al. (2015) remind us that θ, the rescaled mutation rate, is not always proportional to $N\mu$, but is always necessarily proportional to μ/c_N, where μ is the per-locus per-generation mutation rate. And c_N can be a non-linear function of the population size (Eldon and Wakeley 2006). In HFSOD models, c_N can also be a function of the associated coalescent parameters. Indeed, in the model introduced by Schweinsberg (2003), c_N is a function of the parameter α. Most population parameters, such as the migration rate (Eldon and Wakeley 2009) and recombination rate (Eldon and Wakeley 2008; Birkner et al. 2013), must also be scaled with c_N. Thus, while the rate of migration, say, could be high when measured over N generations, it could be low when measured over the (possibly) much shorter time $1/c_N$. Skewed offspring distribution, coupled with collective dispersal in some marine populations, could therefore generate 'chaotic genetic patchiness' (Johnson and Black 1982, 1984; Broquet et al. 2013), which is temporally unstable genetic heterogeneity observed over short geographic distances.

20.3.7 Multi-loci HFSOD Models

20.3.7.1 Diploidy and Polyploidy

All the models hitherto discussed are single-locus models. With recent advances in DNA sequencing technology, it is now feasible to sequence whole genomes (Li and Durbin 2011; Gronau et al. 2011; Zhao et al. 2012; Hearn et al. 2013; McManus et al. 2014; Halldórsdóttir and Árnason 2015). To analyse genomic data with a coalescent approach, multi-loci coalescent processes, which admit recombination between or within loci, are needed. In fact, one of the first applications of the coalescent approach did involve recombination (Hudson 1983). To model multi-loci dynamics properly, one must take into account diploidy (or polyploidy). Indeed, Mohle and Sagitov (2003) consider a single-locus diploid population model, which admits skewed offspring distribution. If the skewness is 'strong enough', a Ξ-coalescent process is obtained, in which multiple mergers may occur in up to four groups. In the model of Mohle and Sagitov (2003), the diploid parents form pairs, and the pairs produce diploid offspring. Each diploid parent pairs with exactly one other diploid parent. Each diploid offspring associated with a given parent pair is then formed (independently of all other offspring) by sampling two

chromosomes (uniformly), one from each parent. A fourfold Ξ-coalescent necessarily follows a large reproduction event, since 4 parent chromosomes participate in generating the offspring of a given parent pair (see Figure 20.4). Eventually, the two parent chromosomes ancestral to the four marked chromosomes in Figure 20.4 will coalesce to the same ancestral chromosome. However, for the derivation of a coalescent process from a (diploid) population model, the ancestral relations over only one generation are usually considered. Blath et al. (2016) also give an example of how a Ξ-coalescent, under arbitrary level of ploidy, can be obtained. Indeed, let $\underline{k} = (k_1, \ldots, k_r)$, with $k_1 \geq \ldots \geq k_r \geq 2$ and $|\underline{k}| := k_1 + \cdots + k_r$, denote the sizes of r simultaneous mergers, i.e. k_j ancestral lineages merge in group j for $1 \leq j \leq r$. The number M denotes the maximum number of such groups. Then, when we have m active ancestral lineages, a $\underline{k}$ simultaneous merger occurs with rate (Blath et al. 2016)

$$\lambda_{m,\underline{k}} = \sum_{\ell=0}^{(m-|\underline{k}|)\wedge(M-r)} \binom{m-|\underline{k}|}{\ell} (M)_{r+\ell} M^{-(|\underline{k}|+\ell)} \int_{[0,1]} x^{|\underline{k}|+\ell-2} (1-x)^{m-(|\underline{k}|+\ell)} \Lambda(\mathrm{d}x), \tag{20.5}$$

where Λ denotes a probability measure on the unit interval, and $(a)_n := a(a-1)\ldots(a-n+1)$ denotes the falling factorial. A related model is the collection of multi-loci ancestral recombination graphs (ARGs) derived by Birkner et al. (2013). The coalescent with rate (20.5) can be recovered from the ARG of Birkner et al. (2013) if one restricts to one locus and excludes recombination.

20.3.7.2 Recombination

Birkner et al. (2013) consider a multi-loci discrete-time modified Moran population model of a diploid population in which selfing is excluded. Recombination only occurs between loci, and only one crossover may occur each time. The main result of Birkner et al. (2013) is a multi-loci ancestral recombination graph, which admits simultaneous multiple mergers (in up to four groups). The ancestral recombination graph is then used to study correlation in coalescence times between two loci. Birkner et al. (2013) show that if the ancestral recombination graph admits multiple mergers, even unlinked loci remain dependent. A similar result was also obtained by Eldon and Wakeley (2008), who were mainly concerned with linkage disequilibrium in HFSOD populations. The dependence between unlinked loci in HFSOD populations may possibly be used in inference, and in study design, i.e. determining how many loci, and how many sequences per locus one should obtain genetic variation for in order to be able to distinguish between models with 'adequate' statistical power (cf. e.g. Felsenstein 2006).

20.3.8 Population Structure

Very few HFSOD models take population substructure into account, and this subfield remains to be explored much more. Limic and Sturm (2006) show that a spatial Λ-coalescent exists as a mathematical object, where space is structured in a classical island-model fashion, and the migration rate is fixed at 1. Eldon (2009) derives a structured coalescent process for the classical island model with conservative migration mechanism, in which in each subpopulation reproduction follows the model of Eldon and Wakeley (2006). The resulting structured coalescent process is therefore a special case of the general result of Limic and Sturm (2006), where the coalescence rate (in each subpopulation) takes a specific form of a Λ-coalescent. Taylor and Véber (2009) consider an island model composed of infinitely many subpopulations, which are subject to recurrent extinction followed by recolonisation, and obtain a Ξ-coalescent process under certain conditions.

More recent and promising work (Barton et al. 2010, 2013) involves populations evolving in continuous space, in which reproduction events involve only individuals confined at the time within a certain area. In particular, one might use models of populations evolving in continuous space as approximations to models involving infinitely many demes. However, one might then need to develop new F_{ST}-like statistics, as the classical ones are derived from models on discrete space.

Eldon and Wakeley (2009) derive expressions for F_{ST} in a simple island model of a population which admits HFSOD. Depending on the model, and how time is rescaled, F_{ST} depends explicitly (or implicitly) on the associated coalescent parameters [such as α in the beta-coalescent (20.3)]. Since F_{ST} is based on pairwise comparisons between sequences, one may need to derive new statistics which compare a larger group of sequences (three or four) to fully capture and adequately account for the effect of skewed offspring distribution on substructure statistics.

20.4 Inference Methods Based on Coalescent Models

The coalescent approach is sample based, i.e. it is concerned with the ancestral relations of a sample of gene copies, or DNA sequences, from a natural population. Hence, it is very well suited for inference. A significant volume of work is concerned with developing inference methods based on the Kingman coalescent; see e.g. Wakeley (2007) for an excellent review. Inference methods for multiple merger coalescents can broadly be divided into likelihoods based on the full DNA sequence data and approximate likelihoods based on the site-frequency spectrum.

20.4.1 Exact Inference Methods

Griffiths and Tavarà (1994a, b, c) introduce importance sampling into population genetics, by deriving importance sampling schemes for the Kingman coalescent. Birkner and Blath (2008) extend the works of Ethier and Griffiths (1987), Griffiths and Tavaré (1994a, b, c, 1995) to general Λ-coalescents, i.e. Birkner and Blath derive a recursion for the likelihood for observed DNA sequence data under the infinitely many sites mutation model (Kimura 1969; Watterson 1975). The importance sampling scheme of Birkner et al. (2011) extends the works of Griffiths and Tavaré (1994), Stephens and Donnelly (2000), Hobolth et al. (2008) to Λ-coalescents. Koskela et al. (2015) develop importance sampling algorithms for both Λ- and Ξ-coalescents. However, the computational burden of importance sampling algorithms can be substantial for large sample sizes and high number of mutations. Methods introduced by Li and Stephens (2003) might also be applied to Λ- and Ξ-coalescents to improve the efficiency Koskela et al. (2015). To tackle data sets consisting of large number of sequences (≥ 1000), which share a high number of polymorphic sites, approximate inference methods may be more feasible.

20.4.2 Approximate Inference Methods

Approximate inference methods rely on summary statistics of the full DNA sequence data. A simple summary statistic of DNA sequence data which gives valuable information about variation among individuals is the site-frequency spectrum.

20.4.2.1 The Site-Frequency Spectrum

Let $\xi_i^{(n)}$ denote the number of polymorphic sites at which one variant is observed in i copies, with $1 \leq i < n$, where n denotes sample size (20.6). The (unfolded) site-frequency spectrum (SFS) is then the random vector $\underline{\xi}^{(n)} = \left(\xi_1^{(n)}, \ldots, \xi_{n-1}^{(n)}\right)$. By way of example, (20.6) shows four polymorphic (segregating) sites for 3 DNA sequences.

$$\begin{array}{l} 1 : 00\mathbf{2}0000000 \\ 2 : 00\mathbf{2}0000000 \\ 3 : 0000\mathbf{1}00\mathbf{1}0\mathbf{1} \end{array} \tag{20.6}$$

If we assume that the ancestral state ['0' in (20.6)] is known, the unfolded site-frequency spectrum for the three sequences in (20.6) is $\underline{\xi}^{(3)} = (3, 1)$.

Fu (1995) derived expressions for the expected value and (co)variances of the site-frequency spectrum when associated with the Kingman coalescent. Indeed, if θ denotes the scaled mutation rate, the expected value is given by Fu (1995)

$$\mathbb{E}^{(K)}\left[\xi_i^{(n)}\right] = \frac{\theta}{i}, \quad 1 \leq i < n. \tag{20.7}$$

Closed-form expressions for the expected site-frequency spectrum are quite hard to obtain when associated with Λ- or Ξ-coalescents due to the multiple merger property. Birkner et al. (2013) obtain recursions for the expected value and (co)-variances of the site-frequency spectrum when associated with Λ-coalescents. Blath et al. (2016) use similar arguments to derive recursions for the expected values when associated with Ξ-coalescents. However, the recursions of Blath et al. (2016) do not allow for efficient computation of the expected values when associated with Ξ-coalescents. Using a different approach, Spence et al. (2016) derive an efficient method to compute the expected site-frequency spectrum when associated with Ξ-coalescents (and therefore also Λ-coalescents). In this context, it is worth to mention the work of Polanski and Kimmel (2003), who obtain efficient and numerically stable recursions to compute the expected site-frequency spectrum when associated with population growth (i.e. continuous increase in population size). Together, the methods of Spence et al. (2016) and Polanski and Kimmel (2003) allow for efficient inference, and estimation of statistical power (given sample size) to distinguish between different multiple merger and population growth models.

20.4.2.2 Approximate Likelihood Approach

The individual components $\xi_i^{(n)}$ of the site-frequency spectrum (SFS) are dependent. However, Kersting and Stanciu (2015) show that, in the limit of large sample size, the joint distribution of the SFS converges to that of independent Poissons with mean $2\theta/i$, when associated with the Kingman coalescent. Hence, one could consider the likelihood function

$$L(\theta, \underline{y}^{(n)}) = \prod_{i=1}^{n-1} \frac{(2\theta/i)^{y_i}}{y_i!} \mathrm{e}^{-2\theta/i} \tag{20.8}$$

where $\underline{y}^{(n)} = (y_1^{(n)}, \ldots, y_{n-1}^{(n)})$ denotes the observed SFS, for a large enough sample size.

A similar convergence result, as Kersting and Stanciu (2015) obtained in association with the Kingman coalescent, is not available for multiple merger coalescents. Hence, one needs to make simplifying assumptions. Let $B_i^{(n)}$ denote the random length of branches, which subtend i leaves ($1 \leq i < n$). The random total

length of the genealogy is then $B^{(n)} = B_1^{(n)} + \cdots + B_{n-1}^{(n)}$. Define the relative length $R_i^{(n)} = B_i^{(n)}/B^{(n)}$. Eldon et al. (2015) consider the likelihood function, based on the site-frequency spectrum,

$$L(\Pi, \underline{y}^{(n)}, s) = \mathbb{E}^{(\Pi)}\left[\binom{s}{y_1^{(n)} \ldots y_{n-1}^{(n)}} \prod_{i=1}^{n-1} \left(R_i^{(n)}\right)^{y_i}\right] \tag{20.9}$$

where Π denotes a coalescent process, $\underline{y}^{(n)}$ the observed SFS, $s = y_1^{(n)} + \cdots + y_{n-1}^{(n)}$ is the observed number of segregating sites, and $\mathbb{E}^{(\Pi)}$ denotes expectation with respect to Π. We do not have a way to write the likelihood function (20.8), or even the individual moments $\mathbb{E}^{(\Pi)}\left[\left(R_i^{(n)}\right)^{y_i}\right]$, as a simple function of the relevant parameters. The approximation that Eldon et al. (2015) apply is to replace the random ratio $R_i^{(n)}$ in (20.8) with the quantity

$$\varphi_i^{(n,\Pi)} = \frac{\mathbb{E}^{(\Pi)}\left[B_i^{(n)}\right]}{\mathbb{E}^{(\Pi)}\left[B^{(n)}\right]}, \quad 1 \le i < n; \tag{20.10}$$

the normalised expected site-frequency spectrum. Obviously, $\varphi_i^{(n,\Pi)}$ is not the same as $\mathbb{E}^{(\Pi)}\left[R_i^{(n)}\right]$, but is a decent approximation for some coalescents when sample size is not too small (Eldon et al. 2015). Since $\mathbb{E}^{(\Pi)}\left[R_i^{(n)}\right]$ is a good approximation of $\mathbb{E}^{(\Pi)}\left[\zeta_i^{(n)}\right]$, where $\zeta_i^{(n)} := \xi_i^{(n)}/\left(\xi_1^{(n)} + \cdots + \xi_{n-1}^{(n)}\right)$, the normalised SFS, if the scaled mutation rate θ is not too small (Eldon et al. 2015), it follows that $\varphi_i^{(n,\Pi)}$ is an acceptable approximation of $\mathbb{E}^{(\Pi)}\left[\zeta_i^{(n)}\right]$. The approximate likelihood function is then

$$\tilde{L}(\Pi, \underline{y}^{(n)}, s) = \binom{s}{y_1 \ldots y_{n-1}} \prod_{i=1}^{n-1} \left(\varphi_i^{(n,\Pi)}\right)^{y_i}. \tag{20.11}$$

Since $\varphi_i^{(n,\Pi)}$ can be computed efficiently for any time-homogeneous Ξ-coalescents (Spence et al. 2016), and many population growth models (Polanski and Kimmel 2003), we can now make efficient inference even for large sample size. By way of example, Eldon et al. (2015) show that the beta-coalescent (20.3) can be distinguished from exponential population growth with 'acceptable' power, if sample size and mutation rate are large enough. The hypotheses tested by Eldon et al. (2015) were interval hypotheses, which included the Kingman coalescent. By implication, the Kingman coalescent can be distinguished from the beta-coalescent (regarded as the alternative hypothesis) when the coalescent parameter (α) is not too close to 2.

20.4.2.3 Statistics Based on the SFS

Since $\mathbb{E}^{(\Pi)}\left[R_i^{(n)}\right]$ is a good approximation of $\mathbb{E}^{(\Pi)}\left[\zeta_i^{(n)}\right]$, and (20.9) is not a function of the mutation rate θ, a natural distance statistic, based on the normalised SFS $\underline{\zeta}^{(n)}$, would be

$$X^{(n,\Pi)} = \sqrt{\sum_{i=1}^{n-1} \frac{\left(\zeta_i^{(n)} - \varphi_i^{(n,\Pi)}\right)^2}{\mathbb{V}^{(\Pi)}\left[\zeta_i^{(n)}\right]}} \tag{20.12}$$

in which $\mathbb{V}^{(\Pi)}\left[\zeta_i^{(n)}\right]$ denotes the variance of $\zeta_i^{(n)}$ computed with respect to Π. However, we do not have a way to write down the variance of $\zeta_i^{(n)}$ as a simple function of the associated coalescent parameters or sample size. The statistic $X^{(n,\Pi)}$ is similar to the G_ξ statistic of Fu (1996), except G_ξ is based on the un-normalised SFS $\underline{\xi}^{(n)}$.

20.5 Comparison with Real Data

Multiple merger coalescents generally predict a higher number of singletons than the Kingman coalescent (Birkner et al. 2013b, Blath et al. 2016). Excess number of singletons (relative to the Kingman coalescent) is, in particular, observed in Atlantic cod (Árnason 2004, Árnason and Halldórsdóttir 2015). In fact, Árnason and Halldórsdóttir (2015) observe relative amount of singletons to all polymorphic sites, to be about 0.7 and 0.8 for the *Myg* and *HbA2* polymorphisms, respectively. All the singletons observed after sequencing for the *Myg* and *HbA2* genes were confirmed by additional molecular work (Árnason and Halldórsdóttir 2015). Thus, the excess number of singletons is real, at least in Atlantic cod. This is important, since next-generation sequencing methods introduce sequencing errors (e.g. DePristo et al. (2011)). Methods to distinguish between true variants and artefacts in next-generation sequencing data should therefore receive added significance. Even more so for the reason that an excess of singletons is the most prominent signal of the SFS that distinguishes multiple merger coalescents from the Kingman coalescent.

Blath et al. (2016) obtain different parameter estimates of corresponding Λ- and Ξ-coalescents when applied to the Atlantic cod data of Árnason and Halldórsdóttir (2015). As previously discussed (see section **diploidy and polyploidy**), Ξ-coalescents are obtained when one models diploidy in HFSOD populations. Blath et al. (2016) conclude that one should apply Λ-coalescents to data inherited in haploid fashion, such as mtDNA, and Ξ-coalescent models to autosomal data obtained from polyploid populations.

20.6 Conclusions

Even though some coalescent models and inference methods exist with which to identify skewed reproductive success, much work still remains. By way of example, the interaction of skewed reproductive success with other demographic forces, and/or with various forms of natural selection, remains largely unexplored. By way of example, we do not have multiple merger coalescent models (MMCM), which take into account continuous population growth. We also do not have many MMCM, which also admit selection. Whole-genome data are currently being generated; to analyse these whole-genome data, multi-loci Ξ-coalescent models with selection could be helpful. Then, there are the various forms of population structure to take into account.

Heuer and Sturm (2013) consider a population evolving in two dimensions (a two-dimensional torus) and show that the genealogy of individuals sampled far enough apart converges to that of the non-spatial Kingman's coalescent, even though the variance in offspring distribution is large. The study by Heuer and Sturm (2013) highlights the limits of our models; different demographic factors may cancel each other out. Another limitation worth mentioning is implicit in the way coalescent processes are usually derived from population models: as scaling limits of probabilities of offspring numbers over one generation. This approach ignores possible family and pedigree structure, as has been highlighted by Wakeley et al. (2012).

Hedgecock and Pudovkin (2011) review the empirical evidence for sweepstakes reproductive success (SRS) in marine populations and conclude that SRS has significant impact on marine biodiversity. Skewed reproduction could be much more common than just among some marine populations and might also be detected among certain terrestrial [e.g. forest trees (Ingvarson 2010)] populations. We have efficient (approximate) statistical methods (cf. e.g. Blath et al. 2016; Eldon et al. 2015; Spence et al. 2016) derived from multiple merger models with which to analyse population genetic data; hence, we propose that multiple merger coalescents should be considered for any population, which displays potential for skewed reproductive success; skewed reproductive success could have a significant rôle in shaping biodiversity.

Acknowledgements I thank Einar Árnason for helpful comments. The financial support of the DFG Priority Programme SPP1590 'Probabilistic Structures in Evolution' through grant BL 1105/3-1 to Jochen Blath at TU Berlin, and Matthias Birkner at JGU Mainz, is acknowledged. As is the support of the DFG Priority Programme SPP 1819 'Rapid Evolutionary Adaptation' through DFG grant STE 325/17-1 to Wolfgang Stephan. The generous hospitality of TU Berlin is warmly acknowledged.

References

Árnason E (2004) Mitochondrial cytochrome *b* variation in the high-fecundity Atlantic cod: trans-Atlantic clines and shallow gene genealogy. Genetics 166:1871–1885

Árnason E, Halldórsdóttir K (2015) Nucleotide variation and balancing selection at the *Ckma* gene in Atlantic cod: analysis with multiple merger coalescent models. PeerJ 3:e786. doi: 10.7717/peerj.786, URL http://dx.doi.org/10.7717/peerj.786

Barton NH, Etheridge AM, Véber A (2010) A new model for evolution in a spatial continuum. Electron J Probab 7:162–216

Barton NH, Etheridge AM, Véber A (2013) Modelling evolution in a spatial continuum. J Stat Mech 2013:1002

Berestycki N (2009) Recent progress in coalescent theory. Ensaios Mathématicos 16:1–193

Berestycki J, Berestycki N, Schweinsberg J (2007) Beta-coalescents and continuous stable random trees. Ann Probab 35:1835–1887

Berestycki J, Berestycki N, Schweinsberg J (2008) Small-time behavior of beta coalescents. Ann Inst H Poincaré Probab Statist 44:214–238

Bhaskar A, Clark A, Song Y (2014) Distortion of genealogical properties when the sample size is very large. PNAS 111:2385–2390

Birkner M, Blath J (2008) Computing likelihoods for coalescents with multiple collisions in the infinitely many sites model. J Math Biol 57:435–465

Birkner M, Blath J (2009) Measure-valued diffusions, general coalescents and population genetic inference. In: Blath J, Mörters P, Scheutzow M (eds) Trends in stochastic analysis. Cambridge University Press, Cambridge, pp 329–363

Birkner M, Blath J, Capaldo M, Etheridge AM, Möhle M, Schweinsberg J, Wakolbinger A (2005) Alpha-stable branching and beta-coalescents. Electron J Probab 10:303–325

Birkner M, Blath J, Möhle M, Steinrücken M, Tams J (2009) A modified lookdown construction for the Xi-Fleming-Viot process with mutation and populations with recurrent bottlenecks. ALEA Lat Am J Probab Math Stat 6:25–61

Birkner M, Blath J, Steinrücken M (2011) Importance sampling for Lambda-coalescents in the infinitely many sites model. Theor Popul Biol 79:155–173

Birkner M, Blath J, Eldon B (2013a) An ancestral recombination graph for diploid populations with skewed offspring distribution. Genetics 193:255–290

Birkner M, Blath J, Eldon B (2013b) Statistical properties of the site-frequency spectrum associated with Λ-coalescents. Genetics 195:1037–1053

Birkner M, Blath J, Steinrücken M (2013c) Analysis of DNA sequence variation within marine species using Beta-coalescents. Theor Popul Biol 87:15–24

Blath J, Cronjäger M, Eldon B, Hammer M (2016) The site-frequency spectrum associated with Ξ-coalescents. http://biorxiv.org/content/early/2015/08/28/025684

Bolthausen E, Sznitman A (1998) On Ruelle's probability cascades and an abstract cavity method. Comm Math Phys 197:247–276

Broquet T, Viard F, Yearsley JM (2013) Genetic drift and collective dispersal can result in chaotic genetic patchiness. Evolution 67(6):1660–1675. doi:10.1111/j.1558-5646.2012.01826.x, url <Go to ISI>://WOS:000319874800012

Cannings C (1974) The latent roots of certain Markov chains arising in genetics: a new approach, I. Haploid models. Adv Appl Probab 6:260–290

Dahmer I, Kersting G, Wakolbinger A (2014) The total external length of Beta-coalescents. Comb Prob Comp 23:1010–1027

DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, McKenna A, Fennell TJ, Kernytsky AM, Sivachenko AY, Cibulskis K, Gabriel SB, Altshuler D, Daly MJ (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet 43(5):491–498, doi:10.1038/ng.806, URL http://dx.doi.org/10.1038/ng.806

Der R, Epstein CL, Plotkin JB (2011) Generalized population models and the nature of genetic drift. Theoret Popul Biol 80(2):80–99. doi:10.1016/j.tpb.2011.06.004, URL http://dx.doi.org/10.1016/j.tpb.2011.06.004

Donnelly P, Kurtz TG (1999) Particle representations for measure-valued population models. Ann Probab 27:166–205

Durrett R, Schweinsberg J (2004) Approximating selective sweeps. Theor Popul Biol 66:129–138

Durrett R, Schweinsberg J (2005) A coalescent model for the effect of advantageous mutations on the genealogy of a population. Stoch Proc Appl 115:1628–1657

Eldon B (2009) Structured coalescent processes from a modified Moran model with large offspring numbers. Theor Popul Biol 76:92–104

Eldon B, Wakeley J (2006) Coalescent processes when the distribution of offspring number among individuals is highly skewed. Genetics 172:2621–2633

Eldon B, Wakeley J (2008) Linkage disequilibrium under skewed offspring distribution among individuals in a population. Genetics 178:1517–1532

Eldon B, Wakeley J (2009) Coalescence times and F_{st} under a skewed offspring distribution among individuals in a population. Genetics 181:615–629

Eldon B, Birkner M, Blath J, Freund F (2015) Can the site-frequency spectrum distinguish exponential population growth from multiple-merger coalescents. Genetics 199:841–856

Etheridge A (2011) Some mathematical models from population genetics. Springer, Berlin. doi:10.1007/978-3-642-16632-7, URL http://dx.doi.org/10.1007/978-3-642-16632-7

Etheridge A, Griffiths R (2009) A coalescent dual process in a Moran model with genic selection. Theor Popul Biol 75:320–330

Etheridge AM, Griffiths RC, Taylor JE (2010) A coalescent dual process in a Moran model with genic selection, and the Lambda coalescent limit. Theor Popul Biol 78:77–92

Ethier S, Griffiths R (1987) The infinitely-many sites model as a measure-valued diffusion. Ann Probab 15:515–545

Ewens WJ (1972) The sampling theory of selectively neutral alleles. Theor Popul Biol 3(1):87–112

Felsenstein J (2006) Accuracy of coalescent likelihood estimates: do we need more sites, more sequences, or more loci. Mol Biol Evol 23:691–700

Flannick J, Thorleifsson G, Beer NL, Jacobs SBR, Grarup N, Burtt NP, Mahajan A, Fuchsberger C, Atzmon G, Benediktsson R, Blangero J, Bowden DW, Brandslund I, Brosnan J, Burslem F, Chambers J, Cho YS, Christensen C, Douglas DA, Duggirala R, Dymek Z, Farjoun Y, Fennell T, Fontanillas P, Forsén T, Gabriel S, Glaser B, Gudbjartsson DF, Hanis C, Hansen T, Hreidarsson AB, Hveem K, Ingelsson E, Isomaa B, Johansson S, Jørgensen T, Jørgensen ME, Kathiresan S, Kong A, Kooner J, Kravic J, Laakso M, Lee JY, Lind L, Lindgren CM, Linneberg A, Masson G, Meitinger T, Mohlke KL, Molven A, Morris AP, Potluri S, Rauramaa R, Ribel-Madsen R, Richard AM, Rolph T, Salomaa V, Segrè AV, Skärstrand H, Steinthorsdottir V, Stringham HM, Sulem P, Tai ES, Teo YY, Teslovich T, Thorsteinsdottir U, Trimmer JK, Tuomi T, Tuomilehto J, Vaziri-Sani F, Voight BF, Wilson JG, Boehnke M, McCarthy MI, Njølstad PR, Pedersen O, Groop L, Cox DR, Stefansson K, Altshuler D (2014) Loss-of-function mutations in SLC30a8 protect against type 2 diabetes. Nat Genet 46(4):357–363. doi:10.1038/ng.2915, URL http://dx.doi.org/10.1038/ng.2915

Foucart C (2013) The impact of selection in the λ-wright-fisher model. Electron Commun Probab 18:1–10

Fu Y (1995) Statistical properties of segregating sites. Theor Popul Biol 48:172–197

Fu Y (1996) New statistical tests of neutrality for DNA samples from a population. Genetics 143:557–570

Fu Y (2006) Exact coalescent for the Wright-Fisher model. Theor Popul Biol 69:385–394

Griffiths R, Tavaré S (1994a) Ancestral inference in population genetics. Stat Sci 9:307–319

Griffiths R, Tavaré S (1994b) Sampling theory for neutral alleles in a varying environment. Phil Trans R Soc London B 344:403–410

Griffiths R, Tavaré S (1994c) Simulating probability distributions in the coalescent. Theor Popul Biol 46:131–159
Griffiths R, Tavaré S (1995) Unrooted genealogical tree probabilities in the infinitely-many-sites model. Math Biosci 127:77–98
Gronau I, Hubisz MJ, Gulko B, Danko CG, Siepel A (2011) Bayesian inference of ancient human demography from individual genome sequences. Nat Genet 43(10):1031–1034. doi:10.1038/ng.937, URL http://dx.doi.org/10.1038/ng.937
Halldórsdóttir K, Árnason E (2015) Whole-genome sequencing uncovers cryptic and hybrid species among Atlantic and Pacific cod-fish. doi:10.1101/034926, http://dx.doi.org/10.1101/034926
Harris H (1966) Enzyme polymorphisms in man. Proc R Soc Lond B Biol Sci 164(995):298–310
Hearn J, Stone GN, Bunnefeld L, Nicholls JA, Barton NH, Lohse K (2013) Likelihood-based inference of population history from low-coverage de novo genome assemblies. Mol Ecol 23(1):198–211. doi:10.1111/mec.12578, URL http://dx.doi.org/10.1111/mec.12578
Hedgecock D, Pudovkin AI (2011) Sweepstakes reproductive success in highly fecund marine fish and shellfish: a review and commentary. Bull Marine Science 87:971–1002
Heuer B, Sturm A (2013) On spatial coalescents with multiple mergers in two dimensions. Theor Population Biol 87:90–104. doi:10.1016/j.tpb.2012.11.006, URL http://dx.doi.org/10.1016/j.tpb.2012.11.006
Hobolth A, Uyenoyama M, Wiuf C (2008) Importance sampling for the infinite sites model. Stat Appl Genet Mol Biol 7, article 32
Hubby J, Lewontin R (1966) A molecular approach to the study of genic heterozygosity in natural populations. I. The number of alleles at different loci in *Drosophila pseudoobscura*. Genetics 54:577–594
Hudson RR (1983) Properties of a neutral allele model with intragenic recombination. Theor Popul Biol 23:183–201
Huillet TE (2014) Pareto genealogies arising from a Poisson branching evolution model with selection. J Math Biol 68(3):727–761
Huillet T, Möhle M (2011) Population genetics models with skewed fertilities: forward and backward analysis. Stoch Models 27:521–554
Huillet T, Möhle M (2013) On the extended Moran model and its relation to coalescents with multiple collisions. Theor Popul Biol 87:5–14
Ingvarson PK (2010) Nucleotide polymorphism, linkage disequilibrium and complex trait dissection in *Populus*. In: Genetics and genomics of *Populus*. Springer, Berlin, pp 91–111
Johnson M, Black R (1982) Chaotic genetic patchiness in an intertidal limpet, *Siphonaria* sp. Mar Biol 70:157–164
Johnson M, Black R (1984) Pattern beneath the chaos: the effect of recruitment on genetic patchiness in an intertidal limpet. Evolution 38:1371–1383
Karlin S, McGregor J (1972) Addendum to paper of W. Ewens. Theor Popul Biol 3:113–116
Kelleher J, Etheridge AM, McVean G (2015) Efficient coalescent simulation and genealogical analysis for large sample sizes. Technical report, University of Oxford. doi:10.1101/033118, URL http://dx.doi.org/10.1101/033118
Kersting G (2012) The asymptotic distribution of the length of Beta-coalescent trees. Ann Appl Probab 22:2086–2107
Kersting G, Stanciu I (2015) The internal branch lengths of the Kingman coalescent. Ann Appl Probab 25:1325–1348
Kimura M (1969) The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. Genetics 61:893–903
Kingman JFC (1982a) The coalescent. Stoch Proc Appl 13:235–248
Kingman JFC (1982b) Exchangeability and the evolution of large populations. In: Koch G, Spizzichino F (eds) Exchangeability in probability and statistics. North-Holland, Amsterdam, pp 97–112
Kingman JFC (1982c) On the genealogy of large populations. J App Probab 19A:27–43
Kingman J (2000) Origins of the coalescent: 1974–1982. Genetics 156:1461–1463

Koskela J, Jenkins P, Spanò D (2015) Computational inference beyond Kingman's coalescent. J Appl Probab 52:519–537

Kreitman M (1983) Nucleotide polymorphism at the alcohol dehydrogenase locus of *Drosophila melanogaster*. Nature 304:412–417

Krone SM, Neuhauser C (1997) Ancestral processes with selection. Theor Popul Biol 51:210–237

Lewontin R, Hubby J (1966) A molecular approach to the study of genic heterozygosity in natural populations. II. Amount of variation and degree heterozygosity in natural populations of *Drosophila pseudoobscura*. Genetics 54:595–609

Li H, Durbin R (2011) Inference of human population history from individual whole-genome sequences. Nature 475(7357):493–496. doi:10.1038/nature10231, URL http://dx.doi.org/10.1038/nature10231

Li G, Hedgecock D (1998) Genetic heterogeneity, detected by PCR-SSCP, among samples of larval Pacific oysters (*Crassostrea gigas*) supports the hypothesis of large variance in reproductive success. Can J Fish Aquat Sci 55(4):1025–1033. doi:10.1139/f97-312, URL http://dx.doi.org/10.1139/f97-312

Li N, Stephens M (2003) Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. Genetics 165:2213–2233

Limic V, Sturm A (2006) The spatial Λ-coalescent. Electron J Probab 11:363–393

May AW (1967) Fecundity of Atlantic cod. J Fish Res Brd Can 24:1531–1551

McManus KF, Kelley JL, Song S, Veeramah KR, Woerner AE, Stevison LS, Ryder OA, Project GAG, Kidd JM, Wall JD, Bustamante CD, Hammer MF (2014) Inference of gorilla demographic and selective history from whole-genome sequence data. Mol Biol Evol 32(3):600–612. doi:10.1093/molbev/msu394, URL http://dx.doi.org/10.1093/molbev/msu394

Möhle M (2011) Coalescent processes derived from some compound Poisson population models. Elect Comm Probab 16:567–582

Möhle M, Sagitov S (2001) A classification of coalescent processes for haploid exchangeable population models. Ann Probab 29:1547–1562

Möhle M, Sagitov S (2003) Coalescent patterns in diploid exchangeable population models. J Math Biol 47:337–352

Nelson MR, Wegmann D, Ehm MG, Kessner D, Jean PS, Verzilli C, Shen J, Tang Z, Bacanu SA, Fraser D, Warren L, Aponte J, Zawistowski M, Liu X, Zhang H, Zhang Y, Li J, Li Y, Li L, Woollard P, Topp S, Hall MD, Nangle K, Wang J, Abecasis G, Cardon LR, Zollner S, Whittaker JC, Chissoe SL, Novembre J, Mooser V (2012) An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. Science 337(6090):100–104. doi:10.1126/science.1217876, URL http://dx.doi.org/10.1126/science.1217876

Neuhauser C, Krone SM (1997) The genealogy of samples in models with selection. Genetics 145:519–534

Oosthuizen E, Daan N (1974) Egg fecundity and maturity of North Sea cod, *Gadus morhua*. Neth J Sea Res 8(4):378–397

Pitman J (1999) Coalescents with multiple collisions. Ann Probab 27:1870–1902

Polanski A, Kimmel M (2003) New explicit expressions for relative frequencies of single-nucleotide polymorphisms with application to statistical inference on population growth. Genetics 165:427–436

Sagitov S (1999) The general coalescent with asynchronous mergers of ancestral lines. J Appl Probab 36:1116–1125

Sagitov S (2003) Convergence to the coalescent with simultaneous mergers. J Appl Probab 40:839–854

Sargsyan O, Wakeley J (2008) A coalescent process with simultaneous multiple mergers for approximating the gene genealogies of many marine organisms. Theor Pop Biol 74:104–114

Schweinsberg J (2000) Coalescents with simultaneous multiple collisions. Electron J Probab 5:1–50

Schweinsberg J (2003) Coalescent processes obtained from supercritical Galton-Watson processes. Stoch Proc Appl 106:107–139

Spence JP, Kamm JA, Song YS (2016) The site frequency spectrum for general coalescents. Genetics 202(4):1549–1561. doi:10.1534/genetics.115.184101, URL http://www.genetics.org/content/202/4/1549, http://www.genetics.org/content/202/4/1549.full.pdf

Stephens M, Donnelly P (2000) Inference in molecular population genetics. J R Stat Soc Ser B Stat Methodol 62:605–655

Tajima F (1983) Evolutionary relationships of DNA sequences in finite populations. Genetics 105:437–460

Taylor J, Véber A (2009) Coalescent processes in subdivided populations subject to recurrent mass extinctions. Electron J Probab 14:242–288

Tellier A, Lemaire C (2014) Coalescence 2.0: a multiple branching of recent theoretical developments and their applications. Mol Ecol 23:2637–2652

Wakeley J (2007) Coalescent theory. Roberts & Co

Wakeley J, Takahashi T (2003) Gene genealogies when the sample size exceeds the effective size of the population. Mol Biol Evol 20:208–2013

Wakeley J, King L, Low BS, Ramachandran S (2012) Gene genealogies within a fixed pedigree, and the robustness of Kingman's coalescent. Genetics 190(4):1433–1445

Watterson GA (1975) On the number of segregating sites in genetical models without recombination. Theor Pop Biol 7:256–276

Zhao S, Zheng P, Dong S, Zhan X, Wu Q, Guo X, Hu Y, He W, Zhang S, Fan W, Zhu L, Li D, Zhang X, Chen Q, Zhang H, Zhang Z, Jin X, Zhang J, Yang H, Wang J, Wang J, Wei F (2012) Whole-genome sequencing of giant pandas provides insights into demographic history and local adaptation. Nat Genet 45(1):67–71. doi:10.1038/ng.2494, URL http://dx.doi.org/10.1038/ng.2494

Chapter 21
From Sequence Data Including Orthologs, Paralogs, and Xenologs to Gene and Species Trees

Marc Hellmuth and Nicolas Wieseke

Abstract Phylogenetic reconstruction aims at finding plausible hypotheses of the evolutionary history of genes or species based on genomic sequence information. The distinction of orthologous genes (genes having a common ancestry and diverged after a speciation) is crucial and lies at the heart of many genomic studies. However, existing methods that rely only on 1:1 orthologs to infer species trees are strongly restricted to a small set of allowed genes that provide information about the species tree. The use of larger gene sets that additionally consist of non-orthologous genes (e.g., so-called paralogous or xenologous genes) considerably increases the information about the evolutionary history of the respective species. In this work, we introduce a novel method to compute species phylogenies based on sequence data including orthologs, paralogs, or even xenologs.

21.1 Introduction

Sequence-based phylogenetic approaches heavily rely on initial data sets to be composed of 1:1 orthologous sequences only. To this end, alignments of protein or DNA sequences are employed whose evolutionary history is believed to be congruent to that of the respective species, a property that can be ensured most easily in the absence of gene duplications or horizontal gene transfer. Phylogenetic studies

M. Hellmuth (✉)
Department of Mathematics and Computer Science, University of Greifswald, Walther- Rathenau-Strasse 47, 17487 Greifswald, Germany
e-mail: mhellmuth@mailbox.org

M. Hellmuth
Center for Bioinformatics, Saarland University, Building E 2.1, 151150, 66041 Saarbrücken, Germany

N. Wieseke
Department of Computer Science, Parallel Computing and Complex Systems Group, Leipzig University, Augustusplatz 10, 04109, Leipzig, Germany
e-mail: wieseke@informatik.uni-leipzig.de

P. Pontarotti (ed.), *Evolutionary Biology*, DOI 10.1007/978-3-319-41324-2_21

thus judiciously select families of genes that rarely exhibit duplications (such as rRNAs, most ribosomal proteins, and many of the housekeeping enzymes).

In the presence of gene duplications, however, it becomes necessary to distinguish between the evolutionary history of genes (gene trees) and the evolutionary history of the species (species trees) in which these genes reside.

One way to address this problem is based on the concept of gene tree/species tree reconciliation (Goodman et al. 1979). Here, a gene tree is embedded into a species tree while predicting gene duplications and gene losses as evolutionary events that take place during the divergence of ancestral genes. Subsequent approaches extended this concept by considering additional evolutionary events such as horizontal gene transfer, incomplete lineage sorting, or deep coalescence. See (Doyon et al. 2011; Eulenstein et al. 2010; Szöllösi et al. 2014) for an overview. Instead of constructing a species tree, these methods are used to predict the evolutionary event history of gene families, with respect to a given species tree. This reveals a circular problem: The reconstruction of the species tree requires identifying events of gene family evolution, such as duplications or horizontal gene transfers, and the reconstruction of event-labeled gene trees requires a known species tree (Boussau et al. 2013; Ullah et al. 2015). In recent years, several approaches were introduced using tree reconciliation, to overcome this problem (Bansal and Eulenstein 2013; Boussau et al. 2013; Chang et al. 2013; Chaudhary et al. 2013; Szöllösi et al. 2013; Ullah et al. 2015). Instead, for a given set of gene trees, these methods search for a "most suitable" species tree w.r.t. certain optimization criteria, in which the gene trees can be embedded. Those methods might be probabilistic or heuristic-guided approaches, to find such a species tree in a given search space covering all putative species trees. In comparison with the state-of-the-art sequence-based phylogenetic reconstruction approaches, this allows to predict a species tree, the event history, and the reconciliation between the gene trees and the species trees. In all latter-mentioned methods, the accuracy strongly depends on the predicted gene trees and the methods (together with an underlying evolutionary model) to reconcile the gene trees with the respective species tree.

In this contribution, the opposite way is introduced. Instead of using gene trees as source of information, solely the knowledge about the evolutionary events is utilized in order to infer event-labeled gene trees and the respective species trees. Recent advances in mathematical phylogenetics, based on the theory of symbolic ultrametrics (Böcker and Dress 1998), have indicated that meaningful phylogenetic information and in particular species trees can be reconstructed from the evolutionary event types between pairs of genes only, provided orthologs and paralogs can be distinguished with a degree of certainty (Hellmuth et al. 2013, 2015; Hernandez-Rosales et al. 2012).

We examine a novel approach and explain the conceptional steps for the inference of species trees and event-labeled gene trees, based on the knowledge of orthologs, paralogs, or even xenologs. Whenever the evolutionary events between any two genes (e.g., speciation, duplication, horizontal gene transfer) are known with certainty, this method provides a polynomial-time approach to compute the accurate "least resolved" gene trees, i.e., the topology of these trees does not

pretend a higher resolution than actually supported by the data. Furthermore, in the presence of orthologs and paralogs only, a correct species tree, with which all the computed gene trees can be reconciled with, can also be computed in polynomial time. In particular, this method does not rely on a model of event costs or event probabilities, and moreover, it is independent from a maximum parsimony or maximum-likelihood assumption. Thus, the only possible hindering factor is the correct estimate of the evolutionary events. Since the evolutionary events can only be estimated from sequence data and events as horizontal gene transfer usually complicates the finding of a correct species tree, there are in practice three NP-hard optimization problems to be solved. To solve these problems, an exact implementation as integer linear program (ILP) is used to demonstrate the potential of the approach without confounding it with computational approximations. This method is provided in the software tool `ParaPhylo` (Hellmuth et al. 2015). Although the resolution is very poor for individual gene families, we show that genomewide data sets are sufficient to generate fully resolved phylogenetic trees, even in the presence of horizontal gene transfer.

21.2 Preliminaries

We give here a brief summary of the main definitions and concepts that are needed.

Graphs, Gene Trees, and Species Trees
An *(undirected) graph* G is a pair (V, E) with non-empty vertex set V and edge set E containing two-element subsets of V. A class of graphs that will play an important role in this contribution are cographs. A graph $G = (V, E)$ is a *cograph* iff G does not contain an induced path on four vertices (see (Corneil et al. 1981, 1985) for more details).

A *tree* $T = (V, E)$ is a connected, cycle-free graph. We distinguish two types of vertices in a tree: the *leaves* which are contained in only one edge and the *inner* vertices which are contained in at least two edges. In order to avoid uninteresting trivial cases, we will usually assume that T has at least three leaves.

A *rooted* tree is a tree in which one special (inner) vertex is selected to be the root. The *last common ancestor* $\mathrm{lca}_T(x, y)$ of two vertices x and y in a rooted tree T is the first (unique) vertex that lies on the path from x to the root and y to the root. We say that a tree T contains the triple $\mathsf{xy}|\mathsf{z}$ if x, y, and z are leaves of T and the path from x to y does not intersect the path from z to the root of T. A set of triples $\mathscr{R}$ is consistent if there is a rooted tree that contains all triples in $\mathscr{R}$.

An *event-labeled* tree, usually denoted by the pair (T, t), is a rooted tree T together with a map $t : V \to M$ that assigns to each inner vertex an event $m \in M$. For two distinct leaves x and y of an event-labeled tree (T, t), its last common ancestor $\mathrm{lca}_T(x, y)$ is therefore marked with an event $t(\mathrm{lca}_T(x, y)) = m$, which we denote for simplicity by $\mathrm{lca}_T(x, y) \overset{\wedge}{=}_t m$.

In what follows, the set $\mathbb{S}$ will always denote a set of species and the set $\mathbb{G}$ a set of genes. We write $x \in X$ if a gene $x \in \mathbb{G}$ resides in the species $X \in \mathbb{S}$.

A *species tree* (for $\mathbb{S}$) is a rooted tree T with leaf set $\mathbb{S}$. A *gene tree* (for $\mathbb{G}$) is an event-labeled tree (T, t) that has as leaf set $\mathbb{G}$.

We refer the reader to (Semple and Steel 2003) for an overview and important results on phylogenetics.

Binary Relations and its Graph and Tree Representations

A *(binary) relation* R over (an underlying set) $\mathbb{G}$ is a subset of $\mathbb{G} \times \mathbb{G}$. We will write $\lfloor \mathbb{G} \times \mathbb{G} \rfloor_{\text{irr}} := (\mathbb{G} \times \mathbb{G}) \backslash \{(x, x) | x \in \mathbb{G}\}$ to denote the irreflexive part of $\mathbb{G} \times \mathbb{G}$.

Each relation R has a natural representation as a graph $G_R = (\mathbb{G}, E_R)$ with vertex set $\mathbb{G}$ and edges connecting two vertices whenever they are in relation R. In what follows, we will always deal with *irreflexive symmetric* relations, which we call for simplicity just relations. Therefore, the corresponding graphs G_R can be considered as undirected graphs without loops, that is, $\{x\} \notin E_R$, and additionally, $\{x, y\} \in E_R$ iff $(x, y) \in R$ (and thus, $(y, x) \in R$).

While graph representations G_R of R are straightforward and defined for all binary relations, tree representations of R are a bit more difficile to derive and, even more annoying, not every binary relation does have a tree representation. For each tree representing a relation R over $\mathbb{G}$, the leaf set $L(T)$ is $\mathbb{G}$ and a specific event label is chosen so that the last common ancestor of two distinct elements $x, y \in \mathbb{G}$ is labeled in a way that uniquely determines whether $(x, y) \in R$ or not. That is, an event-labeled tree (T, t) with events "0" and "1" on its inner vertices represents a (symmetric irreflexive) binary relation R if for all $(x, y) \in \lfloor \mathbb{G} \times \mathbb{G} \rfloor_{\text{irr}}$ it holds that $\text{lca}_T(x, y) \stackrel{\triangle}{=}_t 1$ if and only if $(x, y) \in R$.

The latter definitions can easily be extended to arbitrary disjoint (irreflexive symmetric) relations $R_1, \ldots, R_k$ over $\mathbb{G}$: An edge-colored graph $G_{R_1,\ldots,R_k} = (\mathbb{G}, E := \cup_{i=1}^{k} E_{R_i})$ represents the relations $R_1, \ldots, R_k$ if it holds that $(x, y) \in R_i$ if and only if $\{x, y\} \in E$ and the edge $\{x, y\}$ is colored with "i". Analogously, an event-labeled tree (T, t) with events "0" and "$1, \ldots, k$" on its inner vertices represents the relations $R_1, \ldots, R_k$ if for all $(x, y) \in \lfloor \mathbb{G} \times \mathbb{G} \rfloor_{\text{irr}}$ it holds that $\text{lca}_T(x, y) \stackrel{\triangle}{=}_t i$ if and only if $(x, y) \in R_i$, $1 \leq i \leq k$. The latter implies that for all pairs (x, y) that are in none of the relations R_i, we have $\text{lca}_T(x, y) \stackrel{\triangle}{=}_t 0$. In what follows, we will assume that $U_i R_i = \lfloor \mathbb{G} \times \mathbb{G} \rfloor_{\text{irr}}$.

In practice, the disjoint relations correspond to the evolutionary relationship between genes contained in $\mathbb{G}$, as the disjoint relations R_o and R_p that comprise the pairs of orthologous and paralogous genes, respectively.

Paralogy, Orthology, and Xenology

The current flood of genome sequencing data poses new challenges for comparative genomics and phylogenetics. An important topic in this context is the reconstruction of large families of homologous proteins, RNAs, and other genetic elements. The distinction between orthologs, paralogs, and xenologs is a key step in any

research program of this type. The distinction between orthologous and paralogous gene pairs dates back to the 1970s: Two genes whose last common ancestor in the gene tree corresponds to duplication are paralogs; if the last common ancestor was a speciation event, they are orthologs (Fitch 1970). The importance of this distinction is twofold: On the one hand, it is informative in genome annotation, and on the other hand, the orthology (or paralogy) relation conveys information about the events corresponding to internal nodes of the gene tree (Hellmuth et al. 2013) and about the underlying species tree (Hellmuth et al. 2015; Hernandez-Rosales et al. 2012). We are aware of the controversy about the distinction between orthologous and paralogous genes and their consequence in the context of gene function; however, we adopt here the point of view that homology, and therefore, also orthology and paralogy refer only to the evolutionary history of a gene family and not to its function (Gabaldón and Koonin 2013; Gerlt and Babbitt 2000).

In contrast to orthology and paralogy, the definition of xenology is less well established and by no means consistent in the biologic literature. Xenology is defined in terms of *horizontal gene transfer* (*HGT*) that refers to the transfer of genes between organisms in a manner other than traditional reproduction and across species. The most commonly used definition stipulates that two genes are *xenologs* if their history since their common ancestor involves horizontal gene transfer of at least one of them (Fitch 2000; Roy 2001). In this setting, both orthologs and paralogs may at the same time be xenologs (Roy 2001). Importantly, the mathematical framework established for evolutionary "event" relations, as the orthology relation (Böcker and Dress 1998; Hellmuth et al. 2013), naturally accommodates more than two types of events associated with the internal nodes of the gene tree. It is appealing, therefore, to think of a HGT event as different from both speciation and duplication, in line with (Gray and Fitch 1983) where the term "xenologous" was originally introduced.

In this contribution, we therefore will consider the terms orthologs, paralogs and xenologs solely by means of the events on last common ancestors. To this end, note that for a set of genes $\mathbb{G}$, the evolutionary relationship between two homologous genes contained in $\mathbb{G}$ is entirely explained by the *true* evolutionary gene history of these genes. More precisely, if T is a (known) tree reflecting the true gene history together with the events that happened, that is, the labeling t that tags the inner vertices of T as a speciation, duplication, or HGT event, respectively, then we can determine the three disjoint relations R_o, R_p and R_x comprising the pairs of the so-called lca-orthologous, lca-paralogous, and lca-xenologous genes, respectively, as follows: Two genes $x, y \in \mathbb{G}$ are

- lca-*orthologous*, if $\mathrm{lca}_T(x, y) \overset{\wedge}{=}_t$ speciation;
- lca-*paralogous*, if $\mathrm{lca}_T(x, y) \overset{\wedge}{=}_t$ duplication; and
- lca-*xenologous*, if $\mathrm{lca}_T(x, y) \overset{\wedge}{=}_t$ HGT.

The latter also implies the edge-colored graph representation G_{R_o, R_p, R_x} (see Fig. 21.1 for an illustrative example).

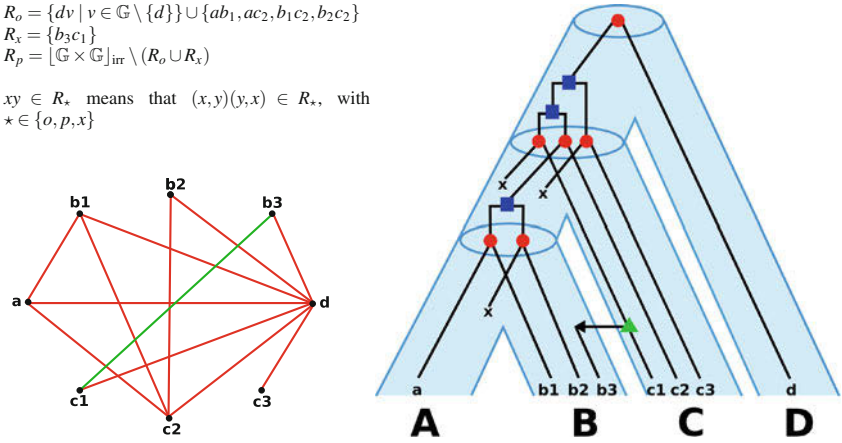


Fig. 21.1 Example of an evolutionary scenario showing the "true" evolution of a gene family evolving along the species tree (shown as *blue* tubelike tree). The corresponding true gene tree T appears embedded in the species tree S. The speciation vertices in the gene tree (*red circuits*) appear on the vertices of the species tree (*blue ovals*), while the duplication vertices (*blue squares*) and the HGT vertices (*green triangles*) are located on the edges of the species tree. Gene losses are represented with "x". The true gene tree T uniquely determines the relationships between the genes by means of the event at $\text{lca}_T(x, y)$ of distinct genes $x, y \in \mathbb{G}$. The pairs of lca-orthologous, -paralogous, and -xenologous genes are comprised in the relations R_o, R_p and R_x, respectively. The graph representation G_{R_o,R_p,R_x} is shown in the *lower left* part. Non-drawn edges indicate the paralogous genes. This graph clearly suggests that the orthology relation R_o is not a complete subgraph and thus *does not* cluster or partition the input gene set $\mathbb{G}$. However, in all cases, the subgraphs G_{R_o}, G_{R_p} and G_{R_x} are the so-called cographs, cf. Theorem 21.1

In the absence of horizontal gene transfer, the relations lca-orthologs and lca-paralogs are equivalent to orthologs and paralogs as defined by Fitch (2000).

We are aware of the fact that this definition of lca-"events" leads to a loss of information of the direction of the HGT event, i.e., the information of donor and acceptor. However, for the proposed method and to understand the idea of representing estimates of evolutionary relationships in an event-labeled tree, this information is not necessarily needed. Nevertheless, generalizations to tree representations of non-symmetric relations or a mathematical framework for xenologs w.r.t. the notion of Fitch might improve the proposed methods.

Remark 1 If there is no risk of confusion and if not stated differently, we call lca-orthologs, lca-paralogs, and lca-xenologs simply orthologs, paralogs, and xenologs, respectively.

Clearly, evolutionary history and the events of the past cannot be observed directly and hence must be inferred, using algorithmic and statistical methods, from the genomic data available today. Therefore, we can only deal with the estimates of the relations R_o, R_p, and R_x. In this contribution, we use those estimates to reconstruct (a hypothesis of) the evolutionary history of the genes and, eventually, the history of the species the genes reside in.

We wish to emphasize that the three relations R_o, R_p, and R_x (will) serve as illustrative examples and the cases $R_p = \emptyset$ or $R_x = \emptyset$ are allowed. In practice, it is possible to have more than these three relations. By way of example, the relation containing the pairs of paralogous genes might be more refined, since gene duplications have several different mechanistic causes that are also empirically distinguishable in real data sets. Thus, instead of having a single relation R_p that comprises all paralogs, we could have different types of paralogy relations that distinguish between events such as local segmental duplications, duplications by retrotransposition, or whole-genome duplications (Zhang 2003).

21.3 From Sequence Data to Species Trees

In this section, we provide the main steps in order to infer event-labeled gene trees and species trees from respective estimated event relations. An implementation of these steps by means of integer linear programming is provided in the software tool `ParaPhylo` (Hellmuth et al. 2015).

The starting point of this method is an estimate of the (true) orthology relation R_o. From this estimate, the necessary information of the event-labeled gene trees and the respective species trees will be derived.

21.3.1 Orthology Detection

The inference of the orthology relation R_o lies at the heart of many reconstruction methods. Orthology inference methods can be classified based on the methodology they use to infer orthology into *tree-based* and *graph-based* methods (for an overview, see, e.g., Altenhoff and Dessimoz 2009; Dalquen et al. 2013; Gabaldón 2008; Kristensen et al. 2011; Trachana et al. 2011).

Tree-based orthology inference methods rely on the reconciliation of a constructed gene tree (without event labeling) from an alignment of homologous sequences and a given species tree (see, e.g., Arvestad et al. 2003; Shi et al. 2011; Hubbard et al. 2007; Van der Heijden et al. 2007; Wapinski et al. 2007).

Although tree-based approaches are often considered as very accurate given a species tree, they suffer from high computational costs and are hence limited in practice to a moderate number of species and genes. A further limitation of those tree reconciliation methods is that for many scenarios, the species tree is not known with confidence, and in addition, all practical issues that complicate phylogenetic inference (e.g., variability of duplication rates, mistaken homology, or HGT) limit the accuracy of both the gene and the species trees.

With ***graph-based orthology inference methods***, it is possible to detect pairs of orthologous genes *without* constructing either gene or species trees. With the recent

advances in graph-based methods, the accuracy of inferred orthology relationships became comparable to that of tree-based methods (Altenhoff and Dessimoz 2012). Many tools of this type have become available over the last decade. To name only a few, COG (Tatusov et al. 2000), OMA (Altenhoff et al. 2011; Schneider et al. 2007), eggNOG (Jensen 2008), OrthoMCL (Chen et al. 2006; Li 2003), InParanoid (Östlund et al. 2010), Roundup 2.0 (Todd 2012), EGM2 (Mahmood et al. 2012) or ProteinOrtho (Lechner et al. 2011), and its extension PoFF (Lechner et al. 2014). Graph-based methods detect orthologous genes for two (pairwise) or more (multiple) species. These methods consist of a *graph construction phase* and, in some cases, a *clustering phase* (Trachana et al. 2011). In the graph construction phase, a graph is inferred where vertices represent genes and (weighted) edges the (confidence of) orthology relationships. The latter rely on pairwise sequence similarities (e.g., Basic Local Alignment Search Tool (BLAST) or Smith-Waterman) calculated between all sequences involved and an operational definition of orthology, for example, reciprocal best hit (RBH), bidirectional best hit (BBH), symmetrical best hit (SymBeT), or reciprocal smallest distance (RSD). In the clustering phase, clusters or groups of orthologs are constructed, using, e.g., single-linkage, complete-linkage, spectral clustering, or Markov cluster algorithm. However, orthology is a symmetric, but not a transitive relation, i.e., it does in general not represent a partition of the set of genes $\mathbb{G}$. In particular, a set $\mathbb{G}'$ of genes can be orthologous to another gene $g \in \mathbb{G} \backslash \mathbb{G}'$, but the genes within $\mathbb{G}'$ are not necessarily orthologous to each other. In this case, the genes in $\mathbb{G}'$ are called co-orthologs to gene g (Koonin 2005). It is important to mention that therefore, the problem of orthology detection is fundamentally different from clustering or partitioning of the input gene set.

In addition to OMA and ProteinOrtho, only Synergy, EGM2, and InParanoid attempt to resolve the orthology relation at the level of gene pairs. The latter two tools can only be used for the analysis of two species at a time, while Synergy is not available as stand-alone tool and therefore cannot be applied to arbitrary user-defined data sets. In particular, the use of orthology inference tools is often limited to the species offered through the databases published by their authors. An exception is provided by ProteinOrtho (Lechner et al. 2011) and its extension PoFF (Lechner et al. 2014), methods that we will use in our approach. These stand-alone tools are specifically designed to handle large-scale user-defined data and can be applied to hundreds of species containing millions of proteins at once. In particular, such computations can be performed on off-the-shelf hardware (Lechner et al. 2011). ProteinOrtho and PoFF compare similarities of given gene sequences (the bit score of the BLAST alignment) that together with an E-value cutoff yield an edge-weighted directed graph. Based on reciprocal best hits, an undirected subgraph is extracted (graph construction phase) on which spectral clustering methods are applied (clustering phase), to determine significant groups of orthologous genes. To enhance the prediction accuracy, the relative order of genes (synteny) can be used as additional feature for the discrimination between orthologs and paralogs.

To summarize, graph-based methods have in common that the output is a set of (pairs of) putative orthologous genes. In addition, orthology detection tools often report some weight or confidence value $w(x, y)$ for x and y to be orthologs or not. This gives rise to a symmetric, irreflexive binary relation

$$\widehat{R}_o = \{(x, y) | x, y \in \mathbb{G} \text{ are estimated orthologs}\} \tag{21.1}$$

$$= \{(x, y) | \mathrm{lca}_T(x, y) \stackrel{\wedge}{=}_t \text{speciation (in the estimated gene tree } T)\}. \tag{21.2}$$

21.3.2 Construction of Gene Trees

Characterization of Evolutionary Event Relations
Assume we have given a "true" orthology relation R_o over $\mathbb{G}$, i.e., R_o comprises all pairs of "true" orthologs, that is, if the true evolutionary history (T, t) of the genes would be known, then $(x, y) \in R_o$ if and only if $\mathrm{lca}_T(x, y) \stackrel{\wedge}{=}_t$ speciation. As we will show, given such a true relation without the knowledge of the gene tree (T, t), it is possible to reconstruct the "observable discriminating part" of (T, t) using the information contained in R_o, resp., R_p only, at least in the absence of xenologous genes (Hellmuth et al. 2013; Hellmuth and Wieseke 2015). In the presence of HGT events, but given the "true" relations R_o and R_p, it is even possible to reconstruct (T, t) using the information contained in R_o and R_p only (Hellmuth et al. 2013; Hellmuth and Wieseke 2015). Note, for the set of pairs of (lca-)xenologs R_x, we have

$$R_x = \lfloor \mathbb{G} \times \mathbb{G} \rfloor_{\mathrm{irr}} \backslash (R_o \cup R_p).$$

Clearly, since we do not know the true evolutionary history with confidence, we always deal with estimates $\widehat{R}_o$, $\widehat{R}_p$, $\widehat{R}_x$ of these true relations R_o, R_p, R_x. In order to understand under which conditions it is possible to infer a gene tree (T, t) that represents the disjoint estimates $\widehat{R}_o$, $\widehat{R}_p$, $\widehat{R}_x$, we characterize in the following the structure of their graph representation $G_{\widehat{R}_o, \widehat{R}_p, \widehat{R}_x}$. Since we assume that $\widehat{R}_o \cup \widehat{R}_p \cup \widehat{R}_x = \lfloor \mathbb{G} \times \mathbb{G} \rfloor_{\mathrm{irr}}$, the graph $G_{\widehat{R}_o, \widehat{R}_p, \widehat{R}_x}$ is a complete edge-colored graph, i.e., for all distinct $x, y \in \mathbb{G}$ there is an edge $\{x, y\} \in E$ s.t. $\{x, y\}$ is colored with "$\star$" if and only if $(x, y) \in R_\star$, $\star \in \{o, p, x\}$.

The following theorem is based on results established by Böcker and Dress (1998) and Hellmuth et al. (2013).

Theorem 21.1 (Böcker and Dress 1998; Hellmuth et al. 2013) *Let $G_{R_1,\ldots,R_k}$ be the graph representation of the relations $R_1, \ldots, R_k$ over some set $\mathbb{G}$ where $\cup_i R_i = \lfloor \mathbb{G} \times \mathbb{G} \rfloor_{\mathrm{irr}}$. There is an event-labeled gene tree representing $R_1, \ldots, R_k$ if and only if*

(i) *The graph* $G_{R_i} = (\mathbb{G}, E_{R_i})$ *is a cograph for all* $i \in \{1, \ldots, k\}$ *and*
(ii) *For all three distinct genes* $x, y, z \in \mathbb{G}$, *the three edges* $\{x, y\}, \{x, z\}$ *and* $\{y, z\}$ *in* $G_{R_1,\ldots,R_k}$ *have at most two distinct colors.*

Clearly, in the absence of xenologs and thus if $\widehat{R}_p = \lfloor \mathbb{G} \times \mathbb{G} \rfloor_{\text{irr}} \backslash \widehat{R}_o$, we can ignore Condition (ii), since at most two colors occur in $G_{\widehat{R}_o, \widehat{R}_p}$. In the latter case, $G_{\widehat{R}_o}$, resp., $G_{\widehat{R}_p}$ alone provide all information of the underlying gene tree.

Theorem 21.1 implies that whenever we have estimates $\widehat{R}_o$, $\widehat{R}_p$, or $\widehat{R}_x$ and we want to find a tree (T, t) that represents these relations, we must ensure that neither $G_{\widehat{R}_o}$, $G_{\widehat{R}_p}$ nor $G_{\widehat{R}_x}$ contains an induced path on four vertices and that there is no triangle (a cycle on three vertices) in $G_{\widehat{R}_o, \widehat{R}_p, \widehat{R}_x}$ where each edge is colored differently. However, due to noise in the data or mispredicted events of pairs of genes, the graph $G_{\widehat{R}_o, \widehat{R}_p, \widehat{R}_x}$ will usually violate Condition (i) or (ii). A particular difficulty arises from the fact that we usually deal with the estimate $\widehat{R}_o$ only and do not know how to distinguish between the paralogs and xenologs.

One possibility to correct the initial estimates $\widehat{R}_o$, $\widehat{R}_p$, $\widehat{R}_x$ to the "closest" relations R_o^*, R_p^*, R_x^* so that there is a tree representation of R_o^*, R_p^*, R_x^*, therefore, could be the change of a minimum number of edge colors in $G_{\widehat{R}_o, \widehat{R}_p, \widehat{R}_x}$ so that $G_{R_o^*, R_p^*, R_x^*}$ fulfills Conditions (i) and (ii). This problem was recently shown to be NP-complete (Hellmuth and Wieseke 2015a, b; Liu et al. 2012).

Inference of Local Substructures of the Gene Tree

Assume we have given (estimated or true) relations R_o, R_p, R_x so that the graph representation G_{R_o, R_p, R_x} fulfills Conditions (i) and (ii) of Theorem 21.1. We show now briefly how to construct the tree (T, t) that represents R_o, R_p, R_x.

Here, we utilize the information of triples that are extracted from the graph G_{R_o, R_p, R_x} and that must be contained in any gene tree (T, t) representing R_o, R_p, R_x. More precisely, given the relations $R_1, \ldots, R_k$, we define the set of triples $\mathscr{T}_{R_1,\ldots,R_k}$ as follows: For all three distinct genes $x, y, z \in \mathbb{G}$, we add the triple $\mathsf{xy|z}$ to $\mathscr{T}_{R_1,\ldots,R_k}$ if and only if the colors of the edge $\{y, z\}$ and $\{x, z\}$ are identical but distinct from the color of the edge $\{x, y\}$ in $G_{R_1,\ldots,R_k}$. In other words, for the given evolutionary relations R_o, R_p, R_x, the triple $\mathsf{xy|z}$ is added to $\mathscr{T}_{R_o,R_p,R_x}$ iff the two genes x and z, as well as y and z, are in the same evolutionary relationship, but different from the evolutionary relation between x and y.

Theorem 21.2 (Böcker and Dress 1998; Hellmuth et al. 2013) *Let* $G_{R_1,\ldots,R_k}$ *be the graph representation of the relations* $R_1, \ldots, R_k$. *The graph* $G_{R_1,\ldots,R_k}$ *fulfills Conditions* (i) *and* (ii) *of Theorem 21.1 (and thus, there is a tree representation of* $R_1, \ldots, R_k$*) if and only if there is a tree* T *that contains all the triples in* $\mathscr{T}_{R_1,\ldots,R_k}$.

The importance of the latter theorem lies in the fact that the well-known algorithm `BUILD` (Aho et al. 1981; Semple and Steel 2003) can be applied to $\mathscr{T}_{R_1,\ldots,R_k}$ to determine whether the set of triples $\mathscr{T}_{R_1,\ldots,R_k}$ is consistent and, if so, constructs a

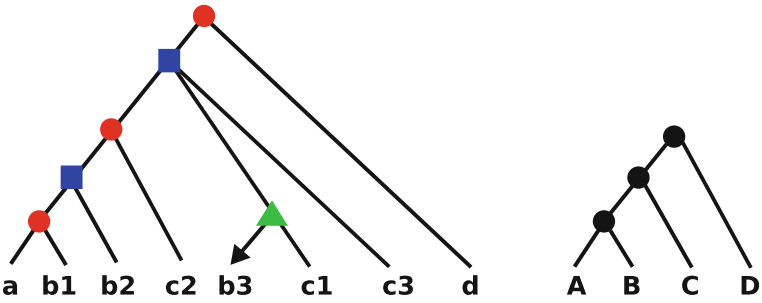


Fig. 21.2 *Left*, the homeomorphic image (T, t) of the observable gene tree (T', t') in Fig. 21.1 is shown. The true gene tree in Fig. 21.1 represents all extant as well as extinct genes, all duplication, HGT, and speciation events. Not all of these events are observable from extant genes' data, however. In particular, extinct genes cannot be observed. Thus, the observable gene tree (T', t') is obtained from the original gene tree in Fig. 21.1 by removing all vertices marked with "x" together with their incident edges and, thereafter, removing all inner vertices that are contained in only two edges. The homeomorphic image (T, t) is obtained from (T', t') by contraction of the edge that connects the two consecutive duplication events. The species triple set $\mathscr{S}$ is $\{\mathsf{AB}|\mathsf{C}_2, \mathsf{AB}|\mathsf{D}_3, \mathsf{AC}|\mathsf{D}_3, \mathsf{BC}|\mathsf{D}_9\}$, where indices indicate the number of gene triples in $\mathscr{R}_o(T)$ that support the respective species triple. In this example, the (unique and thus minimally resolved) species tree S that contains all triples in $\mathscr{S}$ is shown in the right part. The species tree S is identical to the true species tree shown in Fig. 21.1

tree representation in polynomial time. To obtain a valid event label for such a tree T, we can simply set $t(\mathrm{lca}_T(x, y)) = \bigstar$ if the color of the edge $\{x, y\}$ in G_{R_o, R_p, R_x} is "$\bigstar$," $\bigstar \in \{o, p, x\}$ (Hellmuth et al. 2013).

It should be stressed that the evolutionary relations do not contain the full information on the event-labeled gene tree (see Fig. 21.2). Instead, the constructed gene trees (T, t) are homeomorphic images of the (possibly true) observable gene tree (T', t') by collapsing adjacent events of the same type (Hellmuth et al. 2013). That is, in the constructed tree (T, t), all inner vertices that are connected by an edge will have different event labels (see Fig. 21.2). Those trees are also known as discriminating representation, cf. (Böcker and Dress 1998). However, these discriminating representations contain and provide the necessary information to recover the input relations, are unique (up to isomorphism), and do not pretend a higher resolution than actually supported by the data.

21.3.3 Construction of Species Trees

While the latter results have been established for lca-orthologs, lca-paralogs, and lca-xenologs, we restrict our attention in this subsection to orthologous and paralogous genes only and assume that there are no HGT events in the gene trees. We shall see later that in practical computation, the existence of xenologous genes does

not have a large impact on the reconstructed species history, although the theoretical results are established for gene histories without xenologous genes.

In order to derive a species tree S that can be reconciled with or simply spoken "embedded" into a given gene-tree (T, t), (where (T, t) contains only speciation and duplication events), we need to answer the question under which conditions there exists such a species tree for a given gene tree.

A tree $S = (W, F)$ with leaf set $\mathbb{S}$ is a species tree for a gene tree $T = (V, E)$ with leaf set $\mathbb{G}$ if there is a reconciliation map $\mu : V \to W \cup F$ that maps the vertices in V to vertices or edges in $W \cup F$. A reconciliation map μ maps the genes $x \in \mathbb{G}$ in T to the respective species $X \in \mathbb{S}$ in S the gene x resides in so that specific constraints are fulfilled. In particular, the inner vertices of T with label "speciation" are mapped to the inner vertices of S, while the duplication vertices of T are mapped to the edges in W so that the relative "evolutionary order" of the vertices in T is preserved in S. We refer to (Hernandez-Rosales et al. 2012) for the full definition of reconciliation maps. In Fig. 21.1, the reconciliation map μ is implicitly given by drawing the species tree superimposed on the gene tree.

Hence, for a given gene tree (T, t), we wish to efficiently decide whether there is a species tree in which (T, t) can be embedded into and, if so, construct such a species tree together with the respective reconciliation map. We will approach the problem of deriving a species tree from an event-labeled gene tree by reducing the reconciliation map from gene tree to species tree to rooted triples of genes residing in three distinct species. To this end, we define a species triple set $\mathscr{S}$ derived from (T, t) that provides all information needed to efficiently decide whether there is a species tree S for (T, t) or not.

Let $\mathscr{R}_o(T)$ be the set of all triples $\mathsf{ab|c}$ that are contained in T s.t. $a, b, c \in \mathbb{G}$ reside in pairwise different species and $\mathrm{lca}_T(a, b, c) \stackrel{\triangle}{=}_t$ speciation, then set

$$\mathscr{S} := \{\mathsf{AB|C} : \exists \mathsf{ab|c} \in \mathscr{R}_o(T) \text{ with } a \in A, b \in B, c \in C\}.$$

It should be noted that by results established in (Böcker and Dress 1998; Hellmuth et al. 2013), it is possible to derive the triple set $\mathscr{S}$ directly from the orthology relation R_o without constructing a gene tree, cf. (Hellmuth et al. 2013): $\mathsf{AB|C} \in \mathscr{S}$ if and only if

(I) A, B, and C are pairwise different species and there are genes $a \in A$, $b \in B$, $c \in C$ so that *either*

(IIa) $(a, c), (b, c) \in R_o$ and $(a, b) \notin R_o$ *or*

(IIb) $(a, c), (b, c), (a, b) \in R_o$ and there is a gene $d \in \mathbb{G}$ with $(c, d) \in R_o$ and $(a, d), (b, d) \notin R_o$.

Thus, in order to infer species triples, a sufficient number of duplication events must have happened. The following important result was given in (Hernandez-Rosales 2012).

Theorem 21.3 *Let (T, t) be a given gene tree that contains only speciation and duplication events. Then, there is a species tree S for (T, t) if and only if there is a tree containing all triples in* $\mathscr{S}$.

In the positive case, the species tree S and the reconciliation between (T, t) and S can be found in polynomial time.

Interestingly, the latter theorem implies that the gene tree (T, t) can be embedded into any tree that contains the triples in $\mathscr{S}$. Hence, one usually wants to find a species tree with the smallest number of inner vertices, as those trees constitute one of the best estimates of the phylogeny without pretending a higher resolution than actually supported by the data. Such trees are also called minimally resolved tree, and computing such trees is an NP-hard problem (Jansson et al. 2012).

Despite the variance reduction due to cograph editing, noise in the data, as well as the occasional introduction of contradictory triples as a consequence of horizontal gene transfer, is unavoidable. The species triple set $\mathscr{S}$ collected from the individual gene families thus will not always be consistent. The problem of determining a maximum consistent subset of an inconsistent set of triples is NP-hard and also APX-hard (see (Byrka et al. 2010a; Van Iersel 2009)). Polynomial-time approximation algorithms for this problem and further theoretical results are reviewed in (Byrka et al. 2010b).

The results in this subsection have been established for the reconciliation between event-labeled gene trees *without* HGT events and inferred species. Although there are reconciliation maps defined for gene trees that contain xenologs and respective species trees (Bansal et al. 2012, 2013), a mathematical characterization of the species triples $\mathscr{S}$ and the existence of species trees for those gene trees, which might help also to understand the transfer events itself, however, is still an open problem.

21.3.4 Summary of the Theory

The latter results show that it is not necessary to restrict the inference of species trees to 1:1 orthologs. Importantly, orthology information alone is sufficient to reconstruct the species tree, provided that (i) the orthology is known without error and unperturbed by horizontal gene transfer and (ii) the input data contain a sufficient number of duplication events. Although species trees can be inferred in polynomial time for noise-free data, in a realistic setting, three NP-hard optimization problems need to be solved.

We summarize the important working steps to infer the respective gene and species trees from genetic material.

(W1) Compute the estimate $\widehat{R}_o$ and set $\widehat{R}_p = \lfloor \mathbb{G} \times \mathbb{G} \rfloor_{\text{irr}} \backslash \widehat{R}_o$.

(W2) Edit the graph $G_{\widehat{R}_o}$ to the closest cograph with a minimum number of edge edits to obtain the graph G_{R_o}. Note $R_p = \lfloor \mathbb{G} \times \mathbb{G} \rfloor_{\text{irr}} \backslash R_o$.

(W3) Compute the tree representation (T, t) w.r.t. R_o, R_p.

(W4) Extract the species triple set $\widehat{\mathscr{S}}$ from $\mathscr{R}_o(T)$.

(W5) Extract a maximal consistent triple set $\mathscr{S}$ from $\widehat{\mathscr{S}}$.

(W6) Compute a minimally resolved species tree S that contains all triples in $\mathscr{S}$ and, if desired, the reconciliation map μ between (T, t) and S (cf. Theorem 21.3).

In the presence of horizontal transfer, in Step (W1), the xenologous genes x, y are predicted as either orthologs or paralogs.

Furthermore, in Step (W2), it suffices to edit the graph $G_{\widehat{R_o}}$ only, since afterward the graph representation G_{R_p} with $R_p = \lfloor \mathbb{G} \times \mathbb{G} \rfloor_{\text{irr}} \backslash R_o$, and thus, G_{R_o,R_p} fulfills the conditions of Theorem 21.1 (Corneil et al. 1981; Hellmuth et al. 2013). In particular, the graphs G_{R_o,R_p} and G_{R_p} have then been obtained from $G_{\widehat{R_o},\widehat{R_p}}$, resp., $G_{\widehat{R_p}}$ with a minimum number of edge edits. The latter is due to the fact that the complement $\overline{G}_{\widehat{R_o}}$ is the graph $G_{\widehat{R_p}}$ (Corneil et al. 1981).

To extract the species triple set $\widehat{\mathscr{S}}$ in Step (W4), it suffices to choose the respective species triples using Conditions (I) and (IIa)/(IIb), without constructing the gene trees, and thus, Step (W3) can be ignored if the gene history is not of further interest.

21.4 Evaluation

In (Hellmuth et al. 2015), it was already shown that for real-life data sets, the paralogy-based method produces phylogenetic trees for moderately sized species sets. The resulting species trees are comparable to those presented in the literature that are constructed by "state-of-the-art" phylogenetic reconstruction approaches as `RAxML` (Stamatakis 2014) or `MrBayes` (Ronquist et al. 2012). To this end, genomic sequences of eleven *Aquificales* and 19 *Enterobacteriales* species were analyzed. Based on the NCBI gene annotations of those species, an orthology prediction was performed using `ProteinOrtho`. From that prediction, phylogenetic trees were constructed using the aforementioned orthology–paralogy-based approach [working steps (W2)–(W6)] implemented as integer linear program in `ParaPhylo` (Hellmuth et al. 2015). The advantage of this approach is the computation of exact solutions; however, the runtime scales exponentially with the number of input genes per gene family and the number of species.

However, as there is no gold standard for phylogenetic tree reconstruction, three simulation studies are carried out to evaluate the robustness of the method. Using the *artificial life framework* (*ALF*) (Dalquen et al. 2012), the evolution of generated gene sequences was simulated along a given branch length-annotated species tree, explicitly taking into account gene duplication, gene loss, and horizontal transfer events. For realistic species trees, the γ-proteobacteria tree from the OMA project (Altenhoff et al. 2011) was randomly pruned to a size of 10 species while conserving the branch lengths. For additional details on the simulation, see (Hellmuth

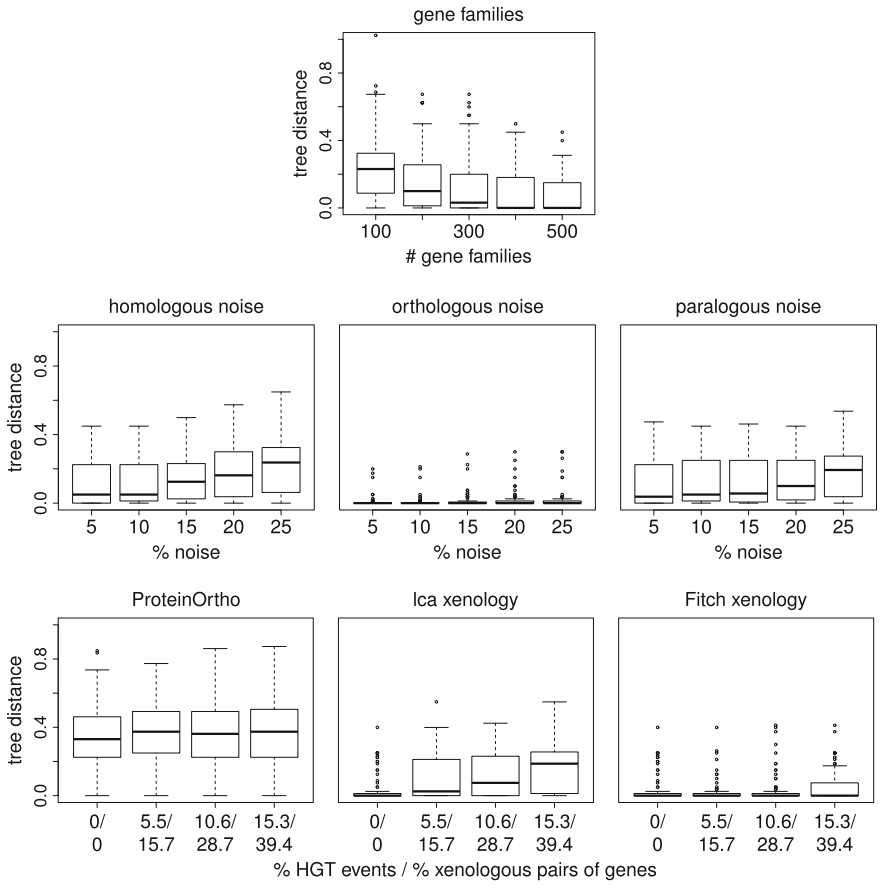


Fig. 21.3 Accuracy of reconstructed species trees (10 species) in simulated data sets; *(Top)* dependence on the number of gene families; *(Middle)* dependence of different noise models; *(Down)* dependence on noise by HGT

et al. 2015). The reconstructed trees are then compared with the initial species trees, using the software `TreeCmp` (Bogdanowicz et al. 2012). In the provided box plots (Fig. 21.3), tree distances are computed according to the triple metric and normalized by the average distance between random Yule trees (see (Hellmuth et al. 2015) for further evaluations).

The three simulation studies are intended to answer three individual questions.

1. How much data are needed to provide enough information to reconstruct accurate species trees? [cf. Fig. 21.3 (top)]
2. How does the method perform with noisy data? [cf. Fig. 21.3 (middle)]
3. What is the impact of horizontal gene transfer on the accuracy of the method? [cf. Fig. 21.3 (down)]

To construct the accurate species trees, the presented method requires a sufficient amount of duplicated genes. Assuming a certain gene duplication rate, the amount of duplicated genes correlates directly with the number of genes per species, respectively, the number of gene families. The first simulation study [Fig. 21.3 (top)] is therefore performed with several numbers of gene families, varying from 100 to 500. The simulation with *ALF* was performed without horizontal gene transfer, and the phylogenetic trees are computed based on the unaltered orthology/paralogy relation obtained from the simulation, that is, the orthologs and paralogs can directly be derived from the simulated gene trees. It turned out that with a duplication rate of 0.005, which corresponds to approximately 8 % of paralogous pairs of genes, 500 gene families are sufficient to produce reliable species trees. With fewer gene families, and hence less duplicated genes, the trees tend to be only poorly resolved.

For the second study, the simulated orthology/paralogy relation of 1000 gene families was perturbed by different types of noise: (i) insertion and deletion of edges in the orthology graph (homologous noise), (ii) insertion of edges (orthologous noise), and (iii) deletion of edges (paralogous noise) (see Fig. 21.3 (middle)). In the three models, an edge is inserted or removed with a probability $p \in \{0.05, 0.1, 0.15, 0.2, 0.25\}$. It can be observed that up to noise of approximately 10 %, the method produces species trees which are almost identical to the initial trees. Particularly, in the case of orthology overprediction (orthologous noise), the method is robust even if 25 % of the input data were disturbed.

Finally, in the third analysis, data sets are simulated with different rates of horizontal gene transfer (see Fig. 21.3 (down)). The number of HGT events in the gene trees is varied up to 15.3 %, which corresponds to 39.4 % of all pairs of genes (x, y) having at least one HGT event on the path from x to y in the generated gene tree, i.e., x and y are xenologous with respect to the definition of Fitch (Fitch 2000). Firstly, the simulated gene sequences are analyzed using `ProteinOrtho`, and the tree reconstruction is then performed based on the resulting orthology/paralogy prediction [Fig. 21.3 (down/left)]. Secondly, we used both definitions of xenology, i.e., lca-xenologous and the notion of Fitch. Note so far the reconstruction of species trees with `ParaPhylo` requires that pairs of genes are either orthologous or paralogous. Hence, we used the information of the lca-orthologs, lca-paralogs, and lca-xenologs derived from the simulated gene trees. Figure 21.3 (down/center) shows the accuracy of reconstructed species trees under the assumption that all lca-xenologs are "mispredicted" as lca-orthologs, in which case all paralogous genes are identified correctly. Figure 21.3 (down/right) shows the accuracy of reconstructed species trees under the assumption that all xenologs w.r.t. the notion of Fitch are interpreted as lca-orthologs. The latter amounts to the "misprediction" of lca-xenologs and lca-paralogs, as lca-orthologs. However, all remaining lca-paralogs are still correctly identified. For the orthology/paralogy prediction based on `ProteinOrtho`, it turned out that the resulting trees have a distance of approximately 0.3 to 0.4 to the initial species tree. Thereby, a distance of 1 refers not to a maximal distance, but to the average distance between random trees. However, the accuracy of the constructed trees appears to be independent from the

amount of horizontal gene transfer. Hence, `ProteinOrtho` is not able to either identify the gene families correctly, or mispredict orthologs and paralogs (due to, e.g., gene loss). In case that all paralogous genes are identified correctly, `ParaPhylo` produces more accurate trees. We obtain even more accurate species trees, when predicting all pairs of Fitch–xenologous genes as lca-orthologs, even with a large amount of HGT events.

21.5 Concluding Remarks

The restriction to 1:1 orthologs for the reconstruction of the evolutionary history of species is not necessary. Even more, it has been shown that the knowledge of only a few correct identified paralogs allows to reconstruct accurate species trees, even in the presence of horizontal gene transfer. Hence, paralogs contain meaningful and valuable information about the gene and the species trees. The information of paralogs is strictly complementary to the sources of information used in phylogenomic studies, which are often based on the alignments of orthologous sequences.

The proposed method computes the event-labeled gene trees in polynomial time, whenever the evolutionary events are known with certainty. In the presence of orthologs and paralogs only, the respective species tree together with the reconciliation map can be found in polynomial time. Therefore, this approach strongly depends on the quality of the estimates of orthologs or paralogs. Future research might therefore focus on the improvements of orthology and paralogy inference tools. In particular, it remains an open question to what extent this "lca-xenology" relation can be inferred directly from sequence similarity data similar to the orthology and paralogy relations. The most commonly used definition of the xenology relation, however, is based on the presence of one or more horizontal transfer events along the unique path in the gene tree that connects two genes. It cannot be expressed in terms of labels at the lowest common ancestor only. This raises the question whether edge-labeled phylogenetic trees give rise to similar systems of relations on the gene set.

If we deal with estimated orthologs or paralogs, or if horizontal gene transfer occurs, then, in practice, three NP-hard problems (cograph or symbolic ultrametric editing, maximum consistent triple set, and minimally resolved tree) need to be solved. Here, the computational tasks are solved exactly for moderate-sized problems by means of an ILP formulation. However, in order to solve the NP-hard problems for large-sized data, efficient and reliable heuristics need to be developed.

For simplifications, horizontal gene transfer is considered as a symmetric event. However, there is a clear distinction between donor and acceptor, and hence, HGT is a directed event. A mathematical framework for event-labeled tree representations of non-symmetric relations is provided in (Hellmuth 2016) and might be used to improve the proposed method.

References

Aho AV, Sagiv Y, Szymanski TG, Ullman JD (1981) Inferring a tree from lowest common ancestors with an application to the optimization of relational expressions. SIAM J Comput 10:405–421

Altenhoff AM, Dessimoz C (2009) Phylogenetic and functional assessment of orthologs inference projects and methods. PLoS Comput Biol 5:e1000262

Altenhoff AM, Dessimoz C (2012) Inferring orthology and paralogy. Evol Genomics Stat Comput Methods 1:259–279

Altenhoff AM, Schneider A, Gonnet GH, Dessimoz C (2011) OMA 2011: orthology inference among 1000 complete genomes. Nucleic Acids Res 39(suppl 1):D289–D294

Arvestad L, Berglund AC, Lagergren J, Sennblad B (2003) Bayesian gene/species tree reconciliation and orthology analysis using mcmc. Bioinformatics 19(suppl 1):i7–i15

Bansal MS, Eulenstein O (2013) Algorithms for genome-scale phylogenetics using gene tree parsimony. Comput Biol Bioinform IEEE/ACM Trans 10(4):939–956

Bansal MS, Alm EJ, Kellis M (2012) Efficient algorithms for the reconciliation problem with gene duplication, horizontal transfer and loss. Bioinformatics 28(12):i283–i291

Bansal MS, Alm EJ, Kellis M (2013) Reconciliation revisited: handling multiple optima when reconciling with duplication, transfer, and loss. J Comput Biol 20(10):738–754

Böcker S, Dress AWM (1998) Recovering symbolically dated, rooted trees from symbolic ultrametrics. Adv Math 138:105–125

Bogdanowicz D, Giaro K, Wróbel B (2012) Treecmp: Comparison of trees in polynomial time. Evol Bioinform Online 8:475

Boussau B, Szöllösi GJ, Duret L, Gouy M, Tannier E, Daubin V (2013) Genome-scale coestimation of species and gene trees. Genome Res 23(2):323–330

Byrka J, Gawrychowski P, Huber KT, Kelk S (2010a) Worst-case optimal approximation algorithms for maximizing triplet consistency within phylogenetic networks. J Discr Alg 8:65–75

Byrka J, Guillemot S, Jansson J (2010b) New results on optimizing rooted triplets consistency. Discr Appl Math 158:1136–1147

Chang WC, Górecki P, Eulenstein O (2013) Exact solutions for species tree inference from discordant gene trees. J Bioinform Comput Biol 11(05):1342005

Chaudhary R, Burleigh JG, Fernandez-Baca D (2013) Inferring species trees from incongruent multi-copy gene trees using the robinson-foulds distance. Algorithms Mol Biol 8:28

Chen F, Mackey AJ, Stoeckert CJ, Roos DS (2006) Orthomcl-db: querying a comprehensive multi-species collection of ortholog groups. Nucleic Acids Res 34(suppl 1):D363–D368

Corneil DG, Lerchs H, Steward Burlingham L (1981) Complement reducible graphs. Discr Appl Math 3:163–174

Corneil DG, Perl Y, Stewart LK (1985) A linear recognition algorithm for cographs. SIAM J Comput 14:926–934

Dalquen DA, Anisimova M, Gonnet GH, Dessimoz C (2012) ALF–a simulation framework for genome evolution. Mol Biol Evol 29(4):1115–1123

Dalquen DA, Altenhoff AM, Gonnet GH, Dessimoz C (2013) The impact of gene duplication, insertion, deletion, lateral gene transfer and sequencing error on orthology inference: a simulation study. PLoS ONE 8(2):e56925

DeLuca TF, Cui J, Jung JY, Gabriel KCS, Wall DP (2012) Roundup 2.0: enabling comparative genomics for over 1800 genomes. Bioinformatics 28(5):715–716

Doyon JP, Ranwez V, Daubin V, Berry V (2011) Models, algorithms and programs for phylogeny reconciliation. Briefings Bioinform 12(5):392–400

Eulenstein O, Huzurbazar S, Liberles DA (2010) Reconciling phylogenetic trees. Evol After Gene Duplication 185–206

Fitch WM (1970) Distinguishing homologous from analogous proteins. Syst Zool 19:99–113

Fitch WM (2000) Homology: a personal view on some of the problems. Trends Genet 16:227–231

Gabaldón T (2008) Large-scale assignment of orthology: back to phylogenetics? Genome Biol 9 (10):235

Gabaldón T, Koonin EV (2013) Functional and evolutionary implications of gene orthology. Nat Rev Genet 14(5):360–366

Gerlt J, Babbitt P (2000) Can sequence determine function? Genome Biol 1(5):reviews0005.1–reviews0005.10

Goodman M, Czelusniak J, William Moore G, Romero-Herrera AE, Matsuda G (1979) Fitting the gene lineage into its species lineage, a parsimony strategy illustrated by cladograms constructed from globin sequences. Syst Biol 28(2):132–163

Gray GS, Fitch WM (1983) Evolution of antibiotic resistance genes: the DNA sequence of a kanamycin resistance gene from *Staphylococcus aureus*. Mol Biol Evol 1:57–66

Hellmuth M, Wieseke N (2015a) On symbolic ultrametrics, cotree representations, and cograph edge decompositions and partitions. In: Xu D, Du D, Du D (eds) Computing and combinatorics. Lecture notes in computer science, vol. 9198. Springer International Publishing, pp 609–623

Hellmuth M, Wieseke N (2015b) On tree representations of relations and graphs: symbolic ultrametrics and cograph edge decompositions. J Comb Opt CoRR abs/1509.05069 (Springer)

Hellmuth M, Hernandez-Rosales M, Huber KT, Moulton V, Stadler PF, Wieseke N (2013) Orthology relations, symbolic ultrametrics, and cographs. J Math Biol 66(1–2):399–420

Hellmuth M, Wieseke N, Lechner M, Lenhof H-P, Middendorf M, Stadler PF (2015) Phylogenomics with paralogs. Proc Natl Acad Sci 112(7):2058–2063

Hellmuth M, Stadler PF, Wieseke N (2016) The mathematics of xenology: Di-cographs, symbolic ultrametrics, 2-structures and tree-representable systems of binary relations. CoRR abs/1603.02467

Hernandez-Rosales M, Hellmuth M, Wieseke N, Huber KT, Moulton V, Stadler PF (2012) From event-labeled gene trees to species trees. BMC Bioinform 13(Suppl 19):S6

Hubbard TJ et al (2007) Ensembl 2007. Nucleic Acids Res 35(suppl 1):D610–D617

Jansson J, Lemence RS, Lingas A (2012) The complexity of inferring a minimally resolved phylogenetic supertree. SIAM J Comput 41:272–291

Jensen RA (2001) Orthologs and paralogs—we need to get it right. Genome Biol 2:8

Jensen LJ, Julien P, Kuhn M, Von Mering C, Muller J, Doerks T, Bork P (2008) eggnog: automated construction and annotation of orthologous groups of genes. Nucleic Acids Res 36 (suppl 1):D250–D254

Koonin EV (2005) Orthologs, paralogs, and evolutionary genomics1. Annu Rev Genet 39(1):309–338

Kristensen DM, Wolf YI, Mushegian AR, Koonin EV (2011) Computational methods for gene orthology inference. Briefings Bioinform 12(5):379–391

Lechner M, Findeiß S, Steiner L, Marz M, Stadler PF, Prohaska SJ (2011) Proteinortho: detection of (co-)orthologs in large-scale analysis. BMC Bioinform 12:124

Lechner M, Hernandez-Rosales M, Doerr D, Wiesecke N, Thevenin A, Stoye J, Hartmann RK, Prohaska SJ, Stadler PF (2014) Orthology detection combining clustering and synteny for very large datasets. PLoS ONE 9(8):e105015

Li L, Stoeckert CJ, Roos DS (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res 13(9):2178–2189

Liu Y, Wang J, Guo J, Chen J (2012) Complexity and parameterized algorithms for cograph editing. Theoret Comput Sci 461:45–54

Mahmood K, Webb GI, Song J, Whisstock JC, Konagurthu AS (2012) Efficient large-scale protein sequence comparison and gene matching to identify orthologs and co-orthologs. Nucleic Acids Res 40(6):e44–e44

Östlund G, Schmitt T, Forslund K, Köstler T, Messina DN, Roopra S, Frings O, Sonnhammer ELL (2010) InParanoid 7: new algorithms and tools for eukaryotic orthology analysis. Nucleic acids Res 38(suppl 1):D196–D203

Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A, Hohna S, Larget B, Liu L, Suchard MA, Huelsenbeck JP (2012) MrBayes 3.2: efficient bayesian phylogenetic inference and model choice across a large model space. Syst Biol 61(3):539–542

Schneider A, Dessimoz C, Gonnet GH (2007) Oma browser—exploring orthologous relations across 352 complete genomes. Bioinformatics 23(16):2180–2182

Semple C, Steel M (2003) Phylogenetics. In: Oxford lecture series in mathematics and its applications, vol. 24. Oxford University Press, Oxford, UK

Shi G, Peng M-C, Jiang T (2011) Multimsoar 2.0: an accurate tool to identify ortholog groups among multiple genomes. PLoS ONE 6(6):e20892

Stamatakis A (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics

Szöllösi GJ, Rosikiewicz W, Boussau B, Tannier E, Daubin V (2013) Efficient exploration of the space of reconciled gene trees. Syst Biol p syt054

Szöllösi GJ, Tannier E, Daubin V, Boussau B (2014) The inference of gene trees with species trees. Syst Biol p syu048

Tatusov RL, Galperin MY, Natale DA, Koonin EV (2000) The cog database: a tool for genome-scale analysis of protein functions and evolution. Nucleic Acids Res 28(1):33–36

Trachana K, Larsson TA, Powell S, Chen WH, Doerks T, Muller J, Bork P (2011) Orthology prediction methods: a quality assessment using curated protein families. BioEssays 33 (10):769–780

Ullah I, Parviainen P, Lagergren J (2015) Species tree inference using a mixture model. Mol Biol Evol 32(9):2469–2482

Van der Heijden R, Snel B, Van Noort V, Huynen M (2007) Orthology prediction at scalable resolution by phylogenetic tree analysis. BMC Bioinform 8(1):83

Van Iersel L, Kelk S, Mnich M (2009) Uniqueness, intractability and exact algorithms: reflections on level-*k* phylogenetic networks. J Bioinf Comp Biol 7:597–623

Wapinski I, Pfeffer A, Friedman N, Regev A (2007) Automatic genome-wide reconstruction of phylogenetic gene trees. Bioinformatics 23(13):i549–i558

Zhang J (2003) Evolution by gene duplication: an update. Trends Ecol Evol 18:292–298

Chapter 22
Oh Brother, Where Art Thou? Finding Orthologs in the Twilight and Midnight Zones of Sequence Similarity

Bianca Hermine Habermann

Abstract Inferring remote orthologs is a persistent challenge in computational biology. The identification of orthologs is necessary for performing evolutionary analyses, comparative genomics, and genome annotation or for functional predictions and sensible planning of experimental studies. If we miss orthologous relationships due to low sequence conservation, we lose a significant amount of information. Given their fast evolutionary rates, remote orthologs can only be identified on protein level. A pair of proteins that has evolved by speciation and has below 30 % sequence identity can be defined as remote orthologs. Their high sequence divergence prevents their unambiguous recognition as orthologous proteins and does not allow a reliable interpretation of their evolutionary relationship. Thus, many remote orthologs remain hidden to date. In this article, I review current methods for remote orthology inference, highlight existing problems in, and discuss potential solutions for discovering remote orthologs.

22.1 Introduction

Deducing orthologous genes is one core task in computational biology. We need knowledge about orthologous relationships for comparative genomics and evolutionary analysis. We further use it to ascertain the conservation of biological processes beyond species barriers and depend on it for genome annotation by transferring molecular functional annotations of orthologs across species.

B.H. Habermann (✉)
Max Planck Institute of Biochemistry, Am Klopferspitz 18,
82152 Martinsried, Germany
e-mail: habermann@biochem.mpg.de

B.H. Habermann
Aix Marseille Université CNRS, IBDM UMR 7288, Marseille 13288, France

P. Pontarotti (ed.), *Evolutionary Biology*, DOI 10.1007/978-3-319-41324-2_22

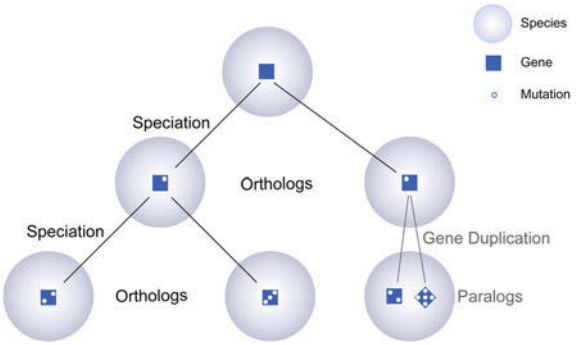


Fig. 22.1 Evolutionary family histories of genes and proteins. Genes/proteins can either evolve by speciation, giving rise to orthologs, or by duplication, which results in paralogs. Xenologs (not shown) are the result of horizontal gene transfer. *Gray shaded* balls show species, *blue rectangles* indicate a gene or protein, and *white balls* designate mutations

Two genes are called homologous, if they have a common ancestor. They are orthologous, if they have evolved by speciation, i.e., the evolution of species. Should they have evolved by gene duplication, they are referred to as paralogous (Fitch 1970) (see Fig. 22.1). Xenologs arise by horizontal gene transfer from another organism (Gray and Fitch 1983). Given these major processes in evolution, family histories of genes can become vastly complicated, which is skillfully reviewed in Kristensen et al. (2011).

There are three main methodologies for orthology inference [reviewed in Kristensen et al. (2011)]: phylogenetic tree analysis, sequence similarity-based best-match approaches, and methods based on synteny. In particular, phylogenetic analysis and sequence-based best-match approaches are powerful and widely used procedures to infer the evolutionary history of a protein family or to decode the relationship between two proteins. They have thus been realized in a number of different pipelines by themselves or in combination (Afrasiabi et al. 2013; Alexeyenko et al. 2006; Ivliev and Sergeeva 2008; Kim et al. 2008; Östlund et al. 2010; Schreiber and Sonnhammer 2013; Shi et al. 2010; Szklarczyk et al. 2012; Wagner et al. 2014).

Kristensen's et al. review (2011) also covers databases of orthologous genes and proteins. There are many resources available with differing content and quality. Virtually all of them use automatic pipelines based on best-match approaches or phylogenetic analysis for assigning orthologous groups. Large genome resource centers including the NCBI or Ensembl have developed and provide their own homology resources, such as Ensembl Compara (Ensembl) (Herrero et al. 2016; Vilella et al. 2009) or HomoloGene (NCBI) (NCBI Resource Coordinators 2016). Altenhoff and Dessimoz (2009) performed a comparative review of methods for ortholog detection and databases of orthologous groups.

Given the amount of available orthology search methods and resources, one could assume that the hunt for orthologs is over. Yet, there are two notable problems: First, complex family histories make the automated assignment of orthologs vulnerable to errors. Gene and genome duplications on all levels of the phylogenetic hierarchy, as well as gene losses, are the major sources of inconsistencies and incorrect orthology assignments in eukaryotes; horizontal gene transfer is an additional complication often encountered in Bacteria and Archaea. Users should therefore carefully evaluate information residing in orthology databases and verify predicted orthologous relationships using an independent method. Second, there are a number of known orthologs, which have undergone significant mutational change over time. Highly diverged, so-called remote orthologs are not easily detected using standard and—especially—automated methods.

Remote orthologs are subject of this review. Methods in remote orthology detection will be discussed, together with their difficulties and limitations. As remote orthologs evolve extremely fast, they only share observable sequence similarity on protein level. Working with remote orthologs is consequently restricted to protein sequences. An 'orthologous pair' will for the remainder of this article refer to a pair of orthologous proteins.

22.2 Defining Remote Orthology: Sequence Similarity Below Statistical Significance

Two proteins can be referred to as remote orthologs, if they have arisen by speciation and have diverged to below 30 % sequence identity at protein level. Below this threshold lies the twilight zone of sequence identity, a term first used by Doolittle in 1986 (Doolittle 1986) to describe a threshold in sequence similarity, at which two proteins are hard to align. For proteins sharing a common fold but hardly detectable sequence similarity, the term midnight zone was introduced by Rost in 1999 (Rost 1999). Protein homologs enter the midnight zone, if they have below 20 % sequence identity with each other (Fig. 22.2).

The fact that two proteins can be homologous even though they show so little sequence conservation became evident as early as 1961 with one of the first structural studies of a polypeptide (Perutz et al. 1960; Watson and Kendrew 1961): sperm whale myoglobin and human hemoglobin have nearly identical folds, even though they share only 26 % sequence identity. Both proteins are members of the globin family, of which vertebrates have at least eight types [reviewed in (Burmester and Hankeln 2014)]. Therefore, they share one common ancestor. This means that these two proteins have arisen by divergent and not convergent evolution. Structural similarity was seen early on as potential evidence that two proteins are related to each other (Fitch 1970). There is meanwhile also compelling

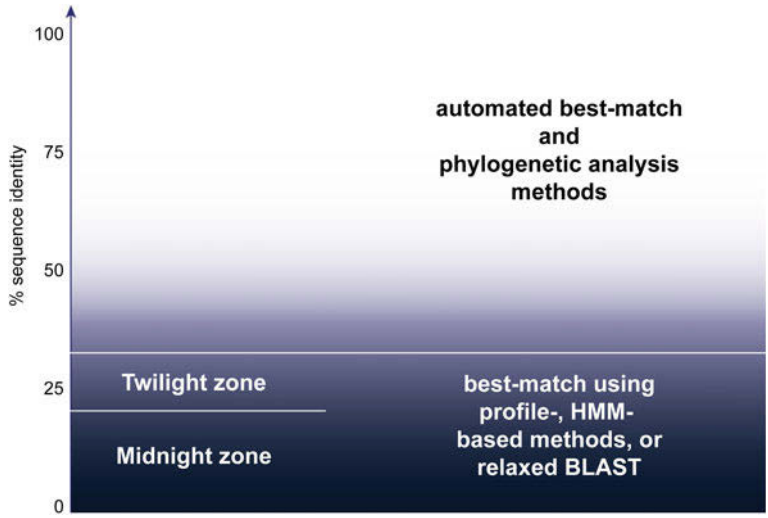


Fig. 22.2 In the twilight and midnight zones of sequence similarity, standard homology search and analysis methods often fail. Given low sequence similarity, profile-based search methods and HMM–HMM comparisons are superior to standard methods

evidence that sequences with similar folds and nearly non-detectable sequence similarity may still have evolved from a common ancestor (Alva et al. 2010).

A much more practical and robust measure of the evolutionary relatedness of two proteins is similarity on the level of their primary structure, hence their DNA or protein sequences (Fitch 1970). In fact, in molecular evolution, we base homology—and consequently orthology and paralogy—on the conservation of sequence, irrespective of potentially divergent functions of the evolutionarily related genes.

The assignment of orthology in the twilight or midnight zones represents an additional layer of complexity. It requires evidence for an orthologous relationship between two proteins from either phylogenetic tree analysis or a sequence-based best-match approach. Myoglobin and hemoglobin are for instance not orthologous to each other: both the genome of the sperm whale, *Physeter catodon,* and the human genome contain myoglobin, as well as hemoglobins. The respective orthologs of myoglobin and hemoglobins are well conserved between the two species and are therefore unproblematic to assign. Yet, there are a number of remote orthologs described in the literature, which have either been found by laborious manual bioinformatics techniques, or experimental studies (see Table 22.1 for some examples of remote orthologs reported in the literature).

Table 22.1 A small[a] collection of remote orthologs reported in literature

Protein	Organism	Comment	Author
Wss1	*S. cerevisiae*	Ortholog of mammalian spartan	Stingele et al. (2015)
Cox20	*S. cerevisiae*	Ortholog of human FAM36A	Szklarczyk et al. (2013)
Collection of mitochondrial sequence organs	*S. cerevisiae, S. pombe*	Collection of human orthologs identified with Ortho-Profile	Szklarczyk et al. (2012)
CDC45 of CMG complex	Eukaryotes	RecJ in Archaea	Makarova et al. (2012)
GAGA factor	*D. melanogaster*	Potential ortholog of mouse Zbtb3	Kumar et al. (2011)
Soronin	*X. laevis*	Ortholog in *D. melanogaster* is Dalmatian	Nishiyama et al. (2010)
AUR1	*S. cerevisiae*	Co-orthologs in *Arabidopsis thaliana* are the IPC synthases IPCS1, IPCS2, and IPCS3	Mina et al. (2010)
Dgt3	*D. melanogaster*	Orthologs of human HAUS3,	Lawo et al. (2009)
Dgt5	*D. melanogaster*	Orthologs of human HAUS5	
Dgt6	*D. melanogaster*	Orthologs of human HAUS6	
Hpc2	*S. cerevisiae*	Ortholog of human UBN1	Banumathy et al. (2009)
Rxt2	*S. cerevisiae*	SPBC428.06c is *S. pombe* ortholog	Shevchenko et al. (2008)
Dep1	*S. cerevisiae*	SPBC21C3.02c is *S. pombe* ortholog	
Swc3	*S. cerevisiae*	SPAC4H3.02c is *S. pombe* ortholog	
Ies4	*S. cerevisiae*	SPAC23G3.04 is *S. pombe* ortholog	
Asa1	*S. cerevisiae*	SPAC1006.02 is *S. pombe* ortholog	
PalC	*A. nidulans*	YGR122w is likely ortholog in *S. cerevisiae*	Galindo et al. (2007)
Cdc26	*S. cerevisiae*	B0511.9 is *Caenorhabditis elegans* ortholog of Cdc26	Dong et al. (2007)
Wapl	*D. melanogaster*	Wapl is human ortholog	Kueng et al. (2006)
Bora	*D. melanogaster*	Bora (LOC79866) is human ortholog	Hutterer et al. (2006)
STA-1	*C. elegans*	STAT is *D. melanogaster* ortholog, Stat5b most likely human ortholog	Wang and Levy (2006)
TPXL-1	*C. elegans*	ortholog of *Xenopus laevis* Tpx2	Özlü et al. (2005)
Spd-2	*C. elegans*	ortholog of human KIAA1569 (Cep192) and CG17286 (spd-2) in *D. melanogaster*	Pelletier et al. (2004)
Swm1/Apc13	*S. cerevisiae*	Ortholog of human Apc13 and *S. pombe* Apc13 (SPBC28E12.01c)	Schwickart et al. (2004)

(continued)

Table 22.1 (continued)

Protein	Organism	Comment	Author
Mei-S332	*D. melanogaster*	Co-orthologs in *S. pombe* are Sgo1 and Sgo2, Sgo1 is human ortholog, and C33H5.15 is *C. elegans* ortholog	Rabitsch et al. (2004)
Sgo1	*S. cerevisiae*	Co-ortholog of *S. pombe* Sgo1 (SPBP35G2.03C) and Sgo2 (SPAC15A10.15)	Kitajima et al. (2004)
Doc1/Apc10	*S. cerevisiae*	Ortholog in human is Doc1	Grossberger et al. (1999)

[a]*Note* I am certain that this table is by far not complete and I am aware of its bias. However, it is difficult to find literature-reported orthologs simply due to the different terms used to describe them, such as 'remote ortholog(ue),' remote homolog(ue), distant relative

22.3 Methods for Detection of Orthologs in the Twilight and Midnight Zones of Sequence Similarity

Assigning orthology in the twilight or midnight zones has the same requirements as inferring orthology for well-conserved orthologs: either a tree-based phylogenetic analysis or a sequence-based best-match approach has to be used to establish an orthologous relationship between two proteins.

Whichever method is chosen, the potential ortholog of a particular species must first be identified as a candidate out of thousands of other proteins from this species —or of millions of proteins in a protein database. This first step usually requires a sequence-based database search, whereby the BLAST family (Altschul et al. 1997) is among the most popular search algorithms readily available for this task.

One severe limitation of using standard BLAST in the twilight and midnight zones is the lack of statistical power to ascertain a true evolutionary relationship between two sequences. Naturally, unusually relaxed BLAST parameters have to be chosen to detect remote orthologs. E-values above 100 are observable for highly diverged orthologs (Wagner et al. 2014). Therefore, BLAST is not necessarily the most powerful method to discover remote orthologs.

To circumvent this limitation of BLAST, more sensitive approaches have been developed. These methods include the meanwhile well-established profile-based database searches (Altschul et al. 1997; Bhadra et al. 2006; Zhang et al. 1998), HMM-based sequence database searches (Eddy 2009; Finn et al. 2015; Kumar and Cowen 2009; Sinha and Lynn 2014), as well as profile–profile or HMM–HMM comparisons (Finn et al. 2015; Ginalski 2003; Johnson et al. 2010; Remmert et al. 2012; Söding et al. 2005, 2006; Yona and Levitt 2002). Other methods make use of structural information (Bedoya and Tischer 2014, 2015; Bernardes et al. 2007; Kim et al. 2009; Kuziemko et al. 2011; Lee et al. 2008; Murzin and Bateman 1997; Wu and Zhang 2008), optimize substitution matrices (Abagyan and Batalov 1997; Blake and Cohen 2001; Vogt et al. 1995; Yamada and Tomii 2014), apply

Table 22.2 Popular and ready-to-use remote homology search tools

Search tool	Web link	Comments
BLAST & PSI-BLAST (Altschul et al. 1997) PHI-BLAST (Zhang et al. 1998)	http://blast.ncbi.nlm.nih.gov/Blast.cgi	NCBI BLAST family; most used; very stable; user-friendly; up-to-date databases
HHblits (Remmert et al. 2012)	http://toolkit.tuebingen.mpg.de/hhblits	User-friendly; very powerful suite for HMM–HMM comparison; several algorithms and databases available; HHpred also useable to search against species-specific proteomes (Biegert et al. 2006)
HHpred (Söding et al. 2005)	http://toolkit.tuebingen.mpg.de/hhpred	
JackHMMER (Johnson et al. 2010)	https://www.ebi.ac.uk/Tools/hmmer/search/jackhmmer	EBI-based, iterative HMM–HMM comparison search engine; several databases and subsets thereof available; organism restriction possible
HMMER web server (Finn et al. 2015)	https://www.ebi.ac.uk/Tools/hmmer/	EBI-based HMM search engines; several databases and subsets thereof; organism restriction possible

machine-learning methods to separate homologs from non-homologs using different sets of features (Bedoya and Tischer 2014, 2015; Bernardes et al. 2011; Comin and Verzotto 2011; Darzentas et al. 2005; Karwath and King 2002; Liu et al. 2014, 2015; Maulik and Sarkar 2013; Muda et al. 2011; Shah et al. 2008; Wieser and Niranjan 2009; Yang et al. 2008), or semantically manipulate the search database (Mudgal et al. 2015; 2014; Sandhya et al. 2012) (see Table 22.2 for popular and ready-to-use web-based search engines for remote homology). For further information, also see the reviews on remote homology detection methods from Dietmann et al. or Fariselli et al. (Dietmann et al. 2002; Fariselli et al. 2007).

After having detected an orthology candidate, one has to choose between a phylogenetic tree analysis and a sequence-based best-match approach to establish the nature of the evolutionary relationship.

For reconstructing a phylogenetic tree, typically a multiple sequence alignment of the remotely conserved orthologs has to be built. Due to low sequence conservation and a resulting poor quality of the multiple sequence alignment, it is difficult to perform a reliable phylogenetic analysis in the twilight and midnight zones. Alignment-free phylogenetic methods can alternatively be chosen, which either infer phylogenetic trees without character-based information from multiple sequence alignments (Bhardwaj et al. 2012; Chang et al. 2008; Gupta et al. 2013; Höhl and Ragan 2007; Höhl et al. 2006; Nelesen et al. 2012; Vinga and Almeida 2003) or which combine the construction of the multiple sequence alignment with tree reconstruction (Liu et al. 2009, 2012; Mirarab et al. 2012).

Sequence-based best-match approaches are the alternative, faster, more robust and easier way to deduce an orthologous relationship, given low sequence

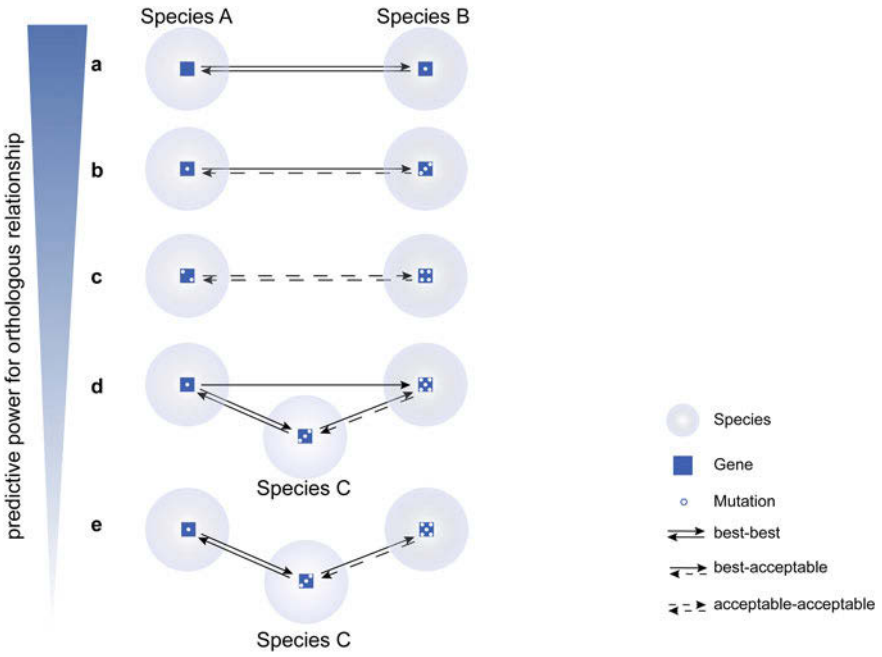


Fig. 22.3 Remote orthologs are often not best-match pairs. Other forms of relationships include best-acceptable, acceptable–acceptable relationships, as well as relationships via a protein from a linker species. However, the predictive power for an orthologous relationship is drastically reduced with the loss of a best-match pair and therefore needs to be experimentally verified. *Gray shaded* balls indicate species, *blue rectangles* represent a gene or protein, and *white balls* designate mutations. *Solid arrows* indicate a best hit, and *broken arrows* show acceptable (non-best) hits

conservation. They also tend to be less computationally expensive. In the best-match approach, two proteins should find each other as the best match (hit) in their respective species (Wolf and Koonin 2012) (Fig. 22.3a). This best match is also referred to as bidirectional best hit, reciprocal best hit, or symmetrical best hit. In plant and animal genomes, the theory of the bidirectional best hit is repeatedly challenged due to genome and gene duplication events (Dalquen and Dessimoz 2013; Gabaldón and Koonin 2013), even given high sequence conservation.

Orthologs in the twilight and midnight zones are also often not each other's bidirectional best hit, even if no gene duplication event has occurred. This can for instance be observed, when large evolutionary distances need to be crossed (Wagner et al. 2014). A number of alternatives exist next to the bidirectional best-hit relationship of two remote orthologs (Fig. 22.3): their relationship can be best-acceptable or acceptable–acceptable (Fig. 22.3b, c). Two orthologs in very distant species can even only be connected via a linker species (Fig. 22.3d, e), which formally raises the question, if these two proteins should be considered orthologous on sequence level. Yet, families with an experimentally verified

orthology-function conjecture exist, which have exactly this type of relationship (Schwickart et al. 2004; Wagner et al. 2014).

The fact that remote orthologs do not always have a best–best relationship is illustrated in Table 22.3 and Figs. 22.4 and 22.5 for the remotely conserved orthologs of the Apc13 family (Schwickart et al. 2004). The web-based search algorithms listed in Table 22.2 were chosen to find the remote ortholog of *Saccharomyces cerevisiae* (*S. cerevisiae, Sc*) Apc13 in *Schizosaccharomyces pombe* (*S. pombe, Sp*) and the human Apc13 ortholog using the *S. pombe* sequence (Table 22.3 and Fig. 22.4). Some algorithms, such as BLAST (Altschul et al. 1997) and HMMSEARCH (Finn et al. 2015), failed all together to detect the respective orthologs (Fig. 22.4a). For other search engines, like PSI-BLAST (Altschul et al. 1997), JackHMMER (Johnson et al. 2010), HHblits (Remmert et al. 2012), or morFeus (Wagner et al. 2014), linker species had to be chosen (Fig. 22.4a, b). Only HHpred searches (Söding et al. 2005) against the respective proteomes led to unambiguous identification of the remote Apc13 orthologs. A morFeus-based network of Apc13 orthologs for selected species is shown in Fig. 22.5. morFeus failed to detect orthologs of *S. pombe* Apc13 in metazoans, as well as the one from *S. cerevisiae*. While the human ortholog was found with *Schizosaccharomyces octosporus* Apc13, the one from *S. cerevisiae* was not detected with this sequence. Apc13 orthologs from the *Schizosaccharomycetes*, *S. cerevisiae*, as well as metazoans could only be united using the sequence from the fungus *Ogataea parapolymorpha*, since morFeus was able to identify the ortholog from the Californian sea hare (*Aplysia californica*) using this fungal Apc13 family member. When there is no best-match relationship in the absence of a paralog, a pattern of conserved amino acids will be the discriminative feature between a pairwise alignment of homologs and a random alignment. morFeus (Wagner et al. 2014) exploits this fact and clusters hits based on the similarity of their alignments. The predictive power for inferring orthology naturally decreases with the loss of a bidirectional best-hit relationship, which makes the experimental verification of a predicted remote ortholog essential.

There are many methods and resources available for prediction and storage of well-conserved orthologs (see Tables 22.2 and 22.4). However, fewer methods exist so far, which can detect remote orthologs automatically with high confidence, among which are Ortho-Profile (Szklarczyk et al. 2012) and morFeus (Wagner et al. 2014). Ortho-Profile successively uses sequence-to-sequence, sequence-to-profile, and profile-to-profile searches combined with the best-match approach to deduce orthology (Szklarczyk et al. 2012). morFeus relies on relaxed BLAST searches, alignment-based clustering of hits, followed by iterative reciprocal BLASTs to infer orthology by the best-match method (Wagner et al. 2014). Both perform comparably well, and both have their limitations with respect to unsupervised application. Ortho-Profile and morFeus together suffer from the lack of a suitable statistical framework to reliably score low sequence similarity, which makes automated assignment of remote orthologs for any method problematic.

Table 22.3 Detection of remote orthologs of Apc13/Swm11

Query or linker species	Detection method	E-value	Reciprocal best-hit E-value	Type of relationship
Detect Apc13 from *S. pombe* (*Sp*) with Apc13/Swm1 from *S. cerevisiae* (*Sc*)				
Sc	BLAST (standard)	–	–	–
Sc	BLAST relaxed (E-value 100, BLOSUM 45, RefSeq fungi)	–	–	–
Sc	BLAST relaxed (E-value 100, Schizosaccharomycetes only)	–	–	–
Sc	PSI-BLAST	–	–	–
Sc	PSI-BLAST relaxed (RefSeq Fungi)	–	–	–
	PSI-BLAST inclusion of *D. hansenii* Apc13 after conversion (4th iteration)	13 (5th iteration) 5e-07 (8th iteration)	23 (3rd iteration) 5e-07 (8th iteration) (RefSeq fungi)	Best–best
Sc	HHblits	–	–	–
Pichia angusta (Pa)	HHblits	*Sc* 3.3	2.2	Best–best
		Sp 0.00018	0.00011	Best–best
Sc	HHpred proteome	0.024	0.069	Best–best
Sc	HMMSEARCH (alignment downloaded from PHMMER)	–	–	–
Linker species	Several linker species were used; either they find back *Sc* or they find back *Sp* Apc13; none could find both			
Sc	JackHMMER	0.089 (4th iteration) 0.00013 (6th iteration)	–	–
D. hansenii (Dh)	JackHMMER	*Sc* 0.7 (3rd iteration)	0.062 (2nd iteration) 4.1e-05 (3rd iteration)	Best–best
		Sp 0.013 (1st iteration) 1.4e-07 (2nd iteration)	0.011 (1st iteration) 1.2e-15 (2nd iteration)	Best–best

(continued)

Table 22.3 (continued)

Query or linker species	Detection method	E-value	Reciprocal best-hit E-value	Type of relationship
Sc	morFeus	–	–	–
Dh	morFeus (E-value 1000, RefSeq Fungi)	*Sc* 0.05	609.492	Best–best
		Sp 0.026	0.0156	Best–best
Detect Apc13 from *H. sapiens (Hs)* with Apc13 from *Sp*				
Sp	BLAST standard	–	–	–
Sp	BLAST relaxed (E-value 100, BLOSUM 45, RefSeq Opisthokonta)	–	–	–
Sp	BLAST relaxed (E-value 100, BLOSUM 62, Refseq Primates)	–	–	–
Shizosaccharomyces octosporus (So)	BLAST relaxed (E-value 100, BLOSUM 62, Refseq Primates)	4.5	25	Best–best
Sp	PSI-BLAST	0.061 (2nd iteration) 2e-04 (3rd iteration)	3e-05 (2nd iteration)	Best–best
Sp	HHblits	–	–	–
Trichoplax adhaerens (Ta)	HHblits	*Sp* 0.12	1.7	Best–best
		Hs 1e-15	1e-10	Best–best
Sp	HHpred proteome	7.6e-15	9.8e-23	Best–best
Sp	HMMSEARCH (alignment downloaded from PHMMER)	0.11	0.005	Best–best
Sp	JackHMMER	0.15 (2nd iteration) 2.1e-05 (3rd iteration)	0.0012 (2nd iteration)	Best–best
Sp	morFeus	–	–	–
So	morFeus	*Sp* 3.76e-58	3.87e-58	Best–best
		Hs 83.027	46.9	Best–best

Not all methods are able to detect the remote orthologs of *S. cerevisiae* Swm1/Apc13 in *S. pombe* or the *Homo sapiens* (*H. sapiens*) Apc13 ortholog with the *S. pombe* Apc13 sequence. As is illustrated in Table 22.3, the use of a linker species can sometimes help to detect remotely conserved orthologs. See Fig. 22.4 for illustration of search results

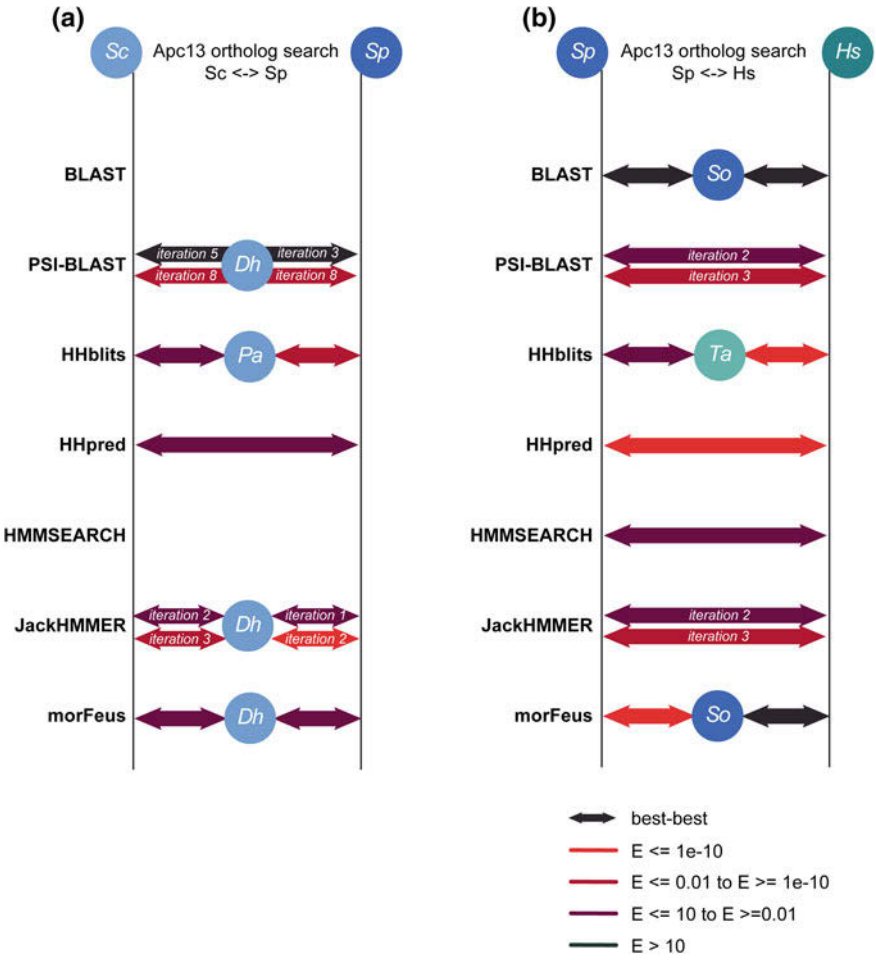


Fig. 22.4 Orthology searches using popular, web-based search engines. The task was to identify **a** *S. pombe (Sp)* Apc13 with *S. cerevisiae (Sc)* Swm1/Apc13 and **b** *H. sapiens (Hs)* Apc13 with *Sp* Apc13 (Table 22.3). The search algorithms used included BLAST, PSI-BLAST, HHblits, HHpred, HMMSEARCH, JackHMMER, and morFeus (indicated on the *left*). *Arrows* refer to best–best relationships and are colored according to E-value: *bright red* low, significant E-value, *dark red/gray* high, insignificant E-value. Linker species that had to be used are indicated between *arrows*. Iterations, where a hit was identified or where E-values changed to significance, are displayed as *two arrows* connecting the two species of interest. Abbreviations of species *Sc Saccharomyces cerevisiae*; *Sp Schizosaccharomyces pombe*; *Hs Homo sapiens*; *Dh Debaryomyces hansenii*; *Pa Pichia angusta*; *So Schizosaccharomyces octosporus*; *Ta Trichoplax adhaerens*

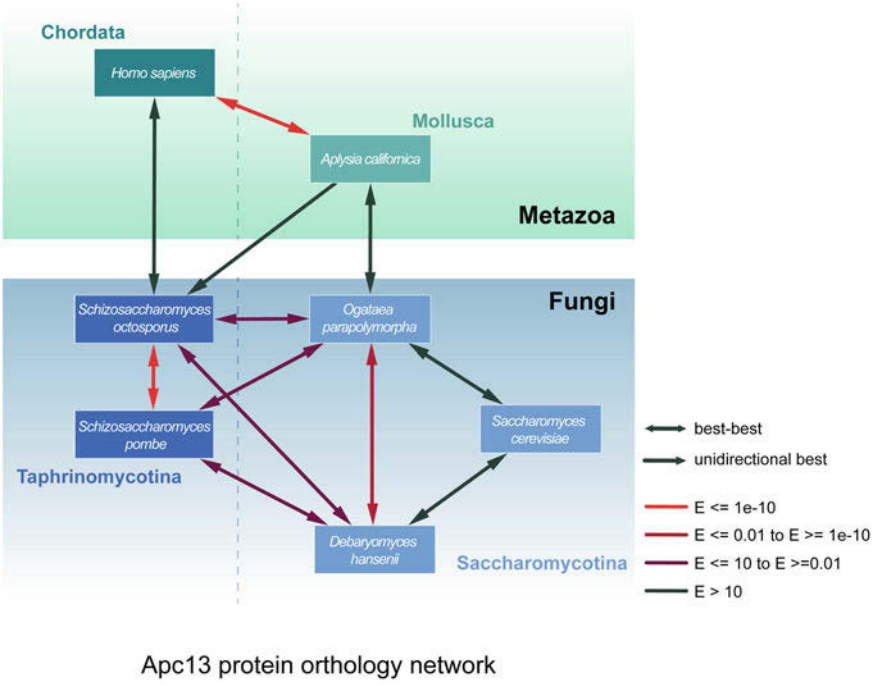


Fig. 22.5 morFeus search-based network of orthology of Apc13 orthologs from selected species. Not all orthologs have a best–best relationship or are detected by all orthologs. Only the fungus *Ogataea parapolymorpha* is able to unite in one search the Apc13 orthologs from *S. cerevisiae*, *S. pombe*, as well as one metazoan, the Californian sea hare *Aplysia californica*. *S. octosporus* is able to detect human Apc13; however, it fails to find the *S. cerevisiae* ortholog. Each of the shown species was used for a morFeus search. Lines with *arrows* at each end indicate best–best relationships; those with a *single arrow* indicate unidirectional best relationships. *Arrows* are colored according to E-value: *bright red* low, significant E-value, *dark red/gray* high, insignificant E-value

22.4 Challenges Arising When Inferring Orthology in the Twilight and Midnight Zones of Sequence Similarity

22.4.1 Proving Orthologous Relationships in Light of Lacking Statistical Power

The inference of remote orthologs is just like the detection of remote homologs a difficult problem, in that it carries all the difficulties of the latter and combines it with having to verify a specific type of relationship. The number of random alignments increases in the twilight and midnight zones of sequence similarity (Rost 1999), even if novel and more sensitive detection methods indicate more true

Table 22.4 Collection of orthology databases. These databases of orthologous proteins and genes do not hold information on orthologs from the twilight or midnight zones

Database Name	Web link	Seq-based search	Tree or alignments	Species	Meta-DB
HomoloGene (NCBI resource coordinators 2016)	http://www.ncbi.nlm.nih.gov/homologene	✓		Eukaryotes Limited to currently 20 species	
Inparanoid (Sonnhammer and Östlund 2015)	http://inparanoid.sbc.su.se/	✓		273 organisms from all kingdoms	
Ensembl Compara (Vilella et al. 2009) 1	http://www.ensembl.org/info/genome/compara/index.html	✓	✓	Data available for many species of different kingdoms in specialized Ensembl systems	
eggNOG (Powell et al. 2011)	http://eggnogdb.embl.de/#/app/home	✓	✓	2031 organisms from all kingdoms	
HCOP (Eyre et al. 2007)	http://www.genenames.org/cgi-bin/hcop			18 eukaryotic species available; human-centric database	✓
HOGENOM (Penel et al. 2009)	http://doua.prabi.fr/databases/hogenom/	✓	✓	1470 genomes from all kingdoms	
HOMOLENS (Penel et al. 2009)	http://pbil.univ-lyon1.fr/databases/homolens.php	✓	✓	Data for Ensembl animal genomes	
HOVERGEN (Dufayard et al. 2005)	http://pbil.univ-lyon1.fr/databases/hovergen.php	✓	✓	Vertebrate genomes only	
Kegg orthology (S. Kim et al. 2008)	http://www.genome.jp/kegg-bin/get_htext?ko00001.keg			Fully sequenced species from all kingdoms	
MetaPhOrs (Pryszcz et al. 2011)	http://orthology.phylomedb.org/	✓	✓	829 fully sequenced genomes from all kingdoms	✓

(continued)

Table 22.4 (continued)

Database Name	Web link	Seq-based search	Tree or alignments	Species	Meta-DB
OrthoDB (Kriventseva et al. 2008)	http://orthodb.org/	✓		2627 bacteria, 227 fungi, 173 metazoa	
OMA (Altenhoff et al. 2015)	http://omabrowser.org/oma/home/	✓		~2000 genomes from all kingdoms	
OrthoMCL DB (Li et al. 2003)	http://www.orthomcl.org/orthomcl/	✓	✓	149 species from all kingdoms	
PhylomeDB (Huerta-Cepas et al. 2007)	http://phylomedb.org/	✓	✓	191 species from all kingdoms	
PLAZA (Proost et al. 2015)	http://bioinformatics.psb.ugent.be/plaza/	✓	✓	Plant centric from algae to trees	
PhyloFACTS (Datta et al. 2009)	http://phylogenomics.berkeley.edu/data		✓	Small selection of species; alignments, HMMs and trees downloadable	
P-POD (Heinicke et al. 2007)	http://ppod.princeton.edu/		✓	12 model species from fungi and metazoa	✓
QuartetS-DB (Yu et al. 2012)	https://applications.bhsai.org/quartetsdb/		✓	1621 species from all kingdoms	
TreeFam (Ruan et al. 2008)	http://www.treefam.org/	✓	✓	109 species from Opisthokonta	
YOGY (Penkett et al. 2006)	http://128.40.79.33/YOGY/			10 eukaryotes, 1 bacteria	✓

positive hits in the zones of low sequence conservation than previously anticipated (Alva et al. 2010). The challenge therefore is to distinguish the true positives from the magnitude of false positives and find those cases, for which a true evolutionary relationship exists. A statistical framework for the reliable identification of non-random sequence similarity in the twilight and midnight zones does currently not exist. A score independent of the BLAST E-value is provided in morFeus, which is based on the reciprocal relationships of identified orthologs (Wagner et al. 2014). This orthology network score is, however, not robust enough to provide a dependable measure of orthology. In most cases, orthology inference in zones of

low sequence similarity will therefore be done manually by pushing the limits of existing sequence similarity search methods such as BLAST, PSI-BLAST, or HMM-based methods and combining it with much experience in observing the evolution of sequences—and thus interpreting the search results correctly. Still, even given manual inference of an orthologous relationship using several methods, a small uncertainty will remain due to the lack of statistical power in the twilight and midnight zones. Verifying an orthologous relationship using experimental methods is therefore highly recommended, if not essential. Following this rule, most of the reported remote orthologs in literature to date have been confirmed experimentally (Table 22.1).

22.4.2 *Influence of Database Content and Size on Detecting Remote Orthologs*

Success and reliability of a remote orthology search depend on the available sequence information. Several factors should be considered.

First, the premise for any reliable phylogenetic analysis is the completeness of genome information of the probed species. Working with remote orthologs means working with protein sequences; therefore, a high-quality and complete annotation of a species' proteome is compulsory.

Second, sequence information of species that can be used as linker species between highly divergent organisms is required for a successful identification of remote orthologs across kingdoms and phyla. A PSI-BLAST search, for instance, will greatly benefit from a larger sequence space covered for a protein family. To give a practical example: a particular protein from *Drosophila melanogaster* will have considerably higher conservation with its orthologs in other arthropods than with its orthologs from Fungi, Nematoda, or Chordata. This simple fact is based on the large difference in divergence time between species from the same phylum versus different phyla; however, it should be noted that depending on the protein family, there are exceptions to this rule (see Table 22.3 and Fig. 22.5). Within arthropods, its orthologs from other drosophilids will have higher conservation than the ones from beetles, lice, or spiders. When searching for remote orthologs, closely related species are unusable, as they do not add substantial new information on protein sequence variation. However, the availability of more divergent members of the same phylum or kingdom greatly expands the sequence space covered for a particular protein family. Thus, it is beneficial having as many divergent, but detectable orthologs present in the explored database. Mudgal et al. tried to simulate an extended sequence space by enriching sequence databases with protein-like sequences (Mudgal et al. 2014, 2015; Sandhya et al. 2012). They observed an improvement in remote homology detection. With the steadily and fast growing amount of available complete genome sequences, we can hope that we soon have a

substantially larger sequence space covered for different phyla, making remote ortholog identification easier.

Yet, the current growth in database size is both a blessing and a curse. For instance, many mammalian genomes have been sequenced recently. This flood of additional sequences provides very little information gain on the level of protein sequence variation. It rather hinders than helps the detection of remote orthologs: searches are getting slower and the majority of hits detected are nearly identical in sequence. Remote orthology detection would rather profit from sequencing more exotic members of different phyla with a higher rate of sequence variation. It would be necessary to fill up sequence space with potential linker species from other chordates, hemichordates, as well as more invertebrates such as nematodes, mollusks, or rotifers. In addition, sequencing more well-selected members of the huge universe of arthropods and fungi would be of great advantage to remote orthology detection. Given the current pace of genome sequencing, we may, however, just have to be patient and wait for more linker species available in sequence space.

22.4.3 Automated Detection of Remote Orthology

To automate remote orthology inference means to lose the well-educated user as an instance of control and guidance on the prediction process. Automated pipelines will therefore be more error-prone. As an example, in (Szklarczyk et al. 2012), co-orthologs in *S. cerevisiae* and *S. pombe* are in some cases mixed up with orthologs in metazoans. morFeus generally fails to identify orthologs in multi-branching protein families (Wagner et al. 2014). A very simple manual evaluation step can identify these wrong or missing orthology assignments. To avoid incorrect inference of orthology, it would be in the interest of the scientific community to maintain manual quality control in zones of low sequence conservation.

When developing algorithms for inferring orthology in the twilight and midnight zones, no reliable gold standard dataset exists for testing software performance. While the SCOP database (Murzin and Bateman 1997; Murzin et al. 1995) is often used to test the performance of algorithms for detecting remote homology on the basis of protein superfamilies, a test set of remote orthologs does not exist. Szklarczyk et al. (2012) solved this problem using an experimental strategy to verify a larger number of predicted orthologs. As mitochondrial proteins were chosen for this study, a localization screen was carried out. Though not unproblematic, as high-throughput methods are prone to errors and can lead to experimental false positives, it is to date the best strategy to assemble a reliable and large enough amount of remote orthologs for performance assessment. It would be highly desirable to gather information on experimentally verified remote orthologs from literature. As this type of work is cumbersome and has little scientific merit, it has not happened yet. One simple solution could be to create an open, community-driven repository, or database, where researchers could deposit published remote orthologs with an experimentally verified orthology-function conjecture.

22.4.4 Orthology-function Conjecture and Remote Orthology Prediction

To close this chapter, it is worth discussing the orthology-function conjecture and its experimental verification in light of remote orthology. Primarily, it is important to note that 'orthology' defines an evolutionary relationship of sequences, not of functions. The 'orthology-function conjecture' describes the assumption that orthologs have a similar biological and molecular function across different organisms [see for instance (Tatusov et al. 1997; Koonin 2005)]. Given complex evolutionary histories of genes (gene duplications, losses, or horizontal gene transfers), as well as the evolution of increasingly complex life forms, experimental verification of the orthology-function conjecture across distant phyla is a tremendous challenge, even given high sequence conservation and a one-to-one orthology situation.

Several fundamental problems have to be considered when probing the orthology-function conjecture of remote orthologs. First, remote orthologs typically cover large evolutionary distances and cross the boundaries of phyla or kingdoms. It is therefore possible that two orthologous proteins have evolved different functions in two evolutionary distant organisms [reviewed effectively in (Koonin 2005)]. Nevertheless, it should be noted that many of the reported remote orthologs, as well as highly conserved orthologs, share the same function even across large evolutionary distances (Table 22.1).

Second, the orthology-function conjecture can be challenged by paralogs. Conflicting results on the functional conservation of orthologs versus paralogs have already been reported in rather closely related species from the same class [(Altenhoff et al. 2012; Nehrt et al. 2011) see also (Studer and Robinson-Rechavi 2009) for a review]. Disentangling functional differences of recent and therefore closely related paralogs might turn out to be extremely difficult, as they could share the same function and thus be able to compensate for each others loss. Older paralogs had more time to evolve in sequence and consequently also in function, which is seen as one of the driving forces of evolution (Conant and Wolfe 2008; Sémon and Wolfe 2007). This makes their functional distinction easier.

Third, differentiating between divergent and convergent evolution given functional similarity of two proteins is not simple in the twilight or midnight zones of sequence similarity. Already in 1970, Fitch raised awareness of the problem that functional constraints could force two analogous proteins to obtain similar chemical and structural properties and be mistaken for homologs, even though they do not share a common ancestry (Fitch 1970). Back at that time, the limits of homology detection were by far not as refined as they are today. The genetic code could be used to discriminate homologs from analogs. If today we cannot observe any sequence homology between two proteins with methods that push the limits of detecting sequence similarity, it can be assumed but not ascertained that there is no common ancestry. Exceptions are some clear examples of convergent evolution of enzymatic function, where evidently dissimilar conserved sequence patterns, as well as three-dimensional structures of enzyme families with a similar molecular

function have been observed [see for instance (Bork et al. 1993)]. Other cases have, however, been reported where only a small domain or even motif is conserved between functionally homologous proteins [e.g., the polo-box domain binding site in mammalian Meikin, *S. pombe* Moa1, and *S. cerevisiae* Spo13 (Kim et al. 2015), all three of which are required for monopolar attachment and protection of centromeric cohesion in meiosis I; the inhibitory domain of the CDK inhibitors *S. cerevisiae* Sic1, *S. pombe* Rum1, and mammalian $p27^{Kip1}$ (Barberis et al. 2005; Sánchez-Díaz et al. 1998), which interacts with and inhibits the Cdk–cyclin complex]. Even in light of lacking sequence similarity, it is difficult to tell, whether there was convergent or divergent evolution at work: it is possible that the evolutionary constraint on a certain protein is limited to a very small region, which would allow the rest of the sequence to diverge beyond recognition. Therefore, reaching a better and more complete sampling of sequence space by sequencing more linker species could help resolve those cases and make evolutionary descent traceable.

Finally, the type of experiment chosen should lead to conclusive results for establishing functional homology between putative orthologs. Large-scale assays as used in (Nehrt et al. 2011) typically have neither the molecular resolution, nor the precision to ascertain homologous functions of putative orthologs between organisms. The question that should be addressed experimentally is whether two proteins perform the exact same biological function in two different organisms. The gold standard to answer this question is a complementation assay, as was for instance done in (Dong et al. 2007; Schwickart et al. 2004). It should be noted though that such genetic rescue experiments are not always successful as they strongly depend on coevolutionary processes, such as coevolution of binding partners.

22.5 Conclusions and Future Perspectives

Inferring remote orthology remains tedious and slow. The initial identification of remote orthologs in a large sequence space and their computational and experimental verification are demanding tasks. Thus, there are still many orthologs hidden in the twilight and midnight zones of sequence similarity. Though sequence databases have grown massively due to numerous genome sequencing studies using NGS technology, and search algorithms have become more sensitive, we still lack methods for deducing remote orthologs systematically in a fully automated way. This would require the development of a robust scoring system for searches in the twilight and midnight zones, which currently does not exist. To date, this is the major limitation in remote orthology detection.

The main improvement over the next years is likely to come from a growth of available sequence information from sensibly chosen linker species. This would also boost the performance of profile-based methods.

The computational verification of remote orthologs will further profit from improved alignment techniques as presented by (Meier and Söding 2015), as well as advances in tree reconstruction methods.

Experimental validation of remote orthologs will most likely remain laborious. However, advances in experimental techniques such as genome editing and experimental automation make a more systematic search for or verification of remote ortholog candidates by experimental assays imaginable.

Box 1—Different Web-based Ways to Find Remote Orthologs

1. morFeus (Wagner et al. 2014). morFeus is a web-based tool. It was developed to find remote orthologs. The output will be a list of orthologs based on a best-match approach. The detailed output, which includes BLAST and reciprocal BLAST search results, as well as pairwise alignments between the query and the hits makes interpretation of search results easy and convenient. The network view of the results enables users to quickly estimate, if orthologs in different phyla have been identified. As morFeus is not unfailing, ortholog predictions with high E-values should be verified by a different method.
2. Reciprocal PSI- (Altschul et al. 1997) or PHI-BLAST (Zhang et al. 1998). PSI-BLAST and the BLAST family of search algorithms at the NCBI are still one of the best and easiest-to-use methods to find remote sequence homology. In combination with reciprocal searches and the best-match approach, it is a good way to find remote orthologs. Look at the alignment rather than the E-value; though it might be high, a family pattern of conserved amino acids should be visible. The up-to-date databases ensure the presence of potential linker species, however, also becomes problematic, as the size of the database influences the E-value. Searching against subsets of the nr or RefSeq database can help.
3. HMM–HMM comparisons. Hidden Markov model comparisons are more sensitive than profile to sequence comparisons and thus represent the high end of sensitive database searches. They are typically superior in finding homologs with very low percentages sequence identity. Even though the influence of potentially missing linker species is not as large, choose a server that has an up-to-date database. Look at the alignment, as E-values can be high for remote homologs/orthologs. Reciprocal searches can be carried out using the same method. Available web-based methods include the MPI Bioinformatics toolkit (Biegert et al. 2006): HHblits (Remmert et al. 2012), HHsenser (Söding et al. 2006), hmmer3 (Eddy 2009), or HHpred (Söding et al. 2005) run against proteomes. Another convenient HMM-based web servers is for instance JackHMMER (Johnson et al. 2010).

General comment: use the ortholog from more than one species as a query for remote homology and orthology searches, if available. It is not uncommon that a sequence of one particular species is superior to its orthologs in finding remote family members in other phyla, supporting the idea of linker species.

See also Table 22.2 for ready-to-use and popular search engines to find remote orthologs. Table 22.3 and Fig. 22.4 illustrate their performance for the remotely conserved Apc13 orthologs.

Acknowledgements I would like to thank Frank Schnorrer and Friedhelm Pfeiffer for critical reading of the manuscript. This work was supported by the Max Planck Society and by the CNRS.

References

Abagyan RA, Batalov S (1997) Do aligned sequences share the same fold? J Mol Biol 273 (1):355–368. doi:10.1006/jmbi.1997.1287

Afrasiabi C, Samad B, Dineen D, Meacham C, Sjölander K (2013) The PhyloFacts FAT-CAT web server: ortholog identification and function prediction using fast approximate tree classification. Nucleic Acids Res 41(Web Server issue), W242–8. doi:10.1093/nar/gkt399

Alexeyenko A, Tamas I, Liu G, Sonnhammer ELL (2006) Automatic clustering of orthologs and inparalogs shared by multiple proteomes. Bioinformatics (Oxford, England), 22(14), e9–15. doi:10.1093/bioinformatics/btl213

Altenhoff AM, Dessimoz C (2009) Phylogenetic and functional assessment of orthologs inference projects and methods. PLoS Comput Biol 5(1):e1000262. doi:10.1371/journal.pcbi.1000262

Altenhoff AM, Studer RA, Robinson-Rechavi M, Dessimoz C (2012) Resolving the ortholog conjecture: orthologs tend to be weakly, but significantly, more similar in function than paralogs. PLoS Comput Biol 8(5):e1002514. doi:10.1371/journal.pcbi.1002514

Altenhoff AM, Škunca N, Glover N, Train C-M, Sueki A, Piližota I et al (2015) The OMA orthology database in 2015: function predictions, better plant support, synteny view and other improvements. Nucleic Acids Res 43(Database issue), D240–9. doi:10.1093/nar/gku1158

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25(17):3389–3402

Alva V, Remmert M, Biegert A, Lupas AN, Söding J (2010) A galaxy of folds. Protein Sci: A Publ Protein Soc 19(1):124–130. doi:10.1002/pro.297

Banumathy G, Somaiah N, Zhang R, Tang Y, Hoffmann J, Andrake M et al (2009) Human UBN1 is an ortholog of yeast Hpc2p and has an essential role in the HIRA/ASF1a chromatin-remodeling pathway in senescent cells. Mol Cell Biol 29(3):758–770. doi:10.1128/MCB.01047-08

Barberis M, De Gioia L, Ruzzene M, Sarno S, Coccetti P, Fantucci P et al (2005) The yeast cyclin-dependent kinase inhibitor Sic1 and mammalian p27Kip1 are functional homologues with a structurally conserved inhibitory domain. Biochem J 387(Pt 3):639–647. doi:10.1042/BJ20041299

Bedoya O, Tischer I (2014) Remote homology detection incorporating the context of physicochemical properties. Comput Biol Med 45:43–50. doi:10.1016/j.compbiomed.2013.11.012

Bedoya O, Tischer I (2015) Reducing dimensionality in remote homology detection using predicted contact maps. Comput Biol Med 59:64–72. doi:10.1016/j.compbiomed.2015.01.020

Bernardes JS, Dávila AMR, Costa VS, Zaverucha G (2007) Improving model construction of profile HMMs for remote homology detection through structural alignment. BMC Bioinform 8 (1):435. doi:10.1186/1471-2105-8-435

Bernardes JS, Carbone A, Zaverucha G (2011) A discriminative method for family-based protein remote homology detection that combines inductive logic programming and propositional models. BMC Bioinform 12(1):83. doi:10.1186/1471-2105-12-83

Bhadra R, Sandhya S, Abhinandan KR, Chakrabarti S, Sowdhamini R, Srinivasan N (2006) Cascade PSI-BLAST web server: a remote homology search tool for relating protein domains. Nucleic Acids Res 34(Web Server issue), W143–6. doi:10.1093/nar/gkl157

Bhardwaj G, Ko KD, Hong Y, Zhang Z, Ho NL, Chintapalli SV et al (2012) PHYRN: a robust method for phylogenetic analysis of highly divergent sequences. PLoS ONE 7(4):e34261. doi:10.1371/journal.pone.0034261

Biegert A, Mayer C, Remmert M, Söding J, Lupas AN (2006) The MPI bioinformatics toolkit for protein sequence analysis. Nucleic Acids Res 34(Web Server issue), W335–9. doi:10.1093/nar/gkl217

Blake JD, Cohen FE (2001) Pairwise sequence alignment below the twilight zone. J Mol Biol 307 (2):721–735. doi:10.1006/jmbi.2001.4495

Bork P, Sander C, Valencia A (1993) Convergent evolution of similar enzymatic function on different protein folds: the hexokinase, ribokinase, and galactokinase families of sugar kinases. Protein Sci: A Publ Protein Soc 2(1):31–40. doi:10.1002/pro.5560020104

Burmester T, Hankeln T (2014) Function and evolution of vertebrate globins. Acta Physiol *(Oxford, England)*, 211(3): 501–514. doi:10.1111/apha.12312

Chang GS, Hong Y, Ko KD, Bhardwaj G, Holmes EC, Patterson RL, van Rossum DB (2008) Phylogenetic profiles reveal evolutionary relationships within the "twilight zone" of sequence similarity. Proc Natl Acad Sci USA 105(36):13474–13479. doi:10.1073/pnas.0803860105

Comin M, Verzotto D (2011) The irredundant class method for remote homology detection of protein sequences. J Computat Biol: J Computat Mol Cell Biol 18(12):1819–1829. doi:10.1089/cmb.2010.0171

Conant GC, Wolfe KH (2008) Turning a hobby into a job: how duplicated genes find new functions. Nat Rev Genet 9(12):938–950. doi:10.1038/nrg2482

Dalquen DA, Dessimoz C (2013) Bidirectional best hits miss many orthologs in duplication-rich clades such as plants and animals. Genome Biol Evol 5(10):1800–1806. doi:10.1093/gbe/evt132

Darzentas N, Rigoutsos I, Ouzounis CA (2005) Sensitive detection of sequence similarity using combinatorial pattern discovery: a challenging study of two distantly related protein families. Proteins 61(4):926–937. doi:10.1002/prot.20608

Datta RS, Meacham C, Samad B, Neyer C, Sjölander K (2009) Berkeley PHOG: PhyloFacts orthology group prediction web server. Nucleic Acids Res 37(Web Server issue), W84–9. doi:10.1093/nar/gkp373

Dietmann S, Fernandez-Fuentes N, Holm L (2002) Automated detection of remote homology. Curr Opin Struct Biol 12(3):362–367

Dong Y, Bogdanova A, Habermann B, Zachariae W, Ahringer J (2007) Identification of the *C. elegans* anaphase promoting complex subunit Cdc26 by phenotypic profiling and functional rescue in yeast. BMC Dev Biol 7(1):19. doi:10.1186/1471-213X-7-19

Doolittle RF (1986) Of Urfs and Orfs: a primer on how to analyze derived amino acid sequences. In: University Science Books, Herndon, VA vol 29, pp 1–103. doi:10.1002/jobm.3620290411

Dufayard J-F, Duret L, Penel S, Gouy M, Rechenmann F, Perrière G (2005) Tree pattern matching in phylogenetic trees: automatic search for orthologs or paralogs in homologous gene sequence databases. Bioinformatics (Oxford, England), 21(11): 2596–2603. doi:10.1093/bioinformatics/bti325

Eddy SR (2009) A new generation of homology search tools based on probabilistic inference. Genome Inform Int Conf Genome Inform 23(1): 205–211

Eyre TA, Wright MW, Lush MJ, Bruford EA (2007) HCOP: a searchable database of human orthology predictions. Briefings Bioinform 8(1):2–5. doi:10.1093/bib/bbl030

Fariselli P, Rossi I, Capriotti E, Casadio R (2007) The WWWH of remote homolog detection: the state of the art. Briefings Bioinform 8(2):78–87. doi:10.1093/bib/bbl032
Finn RD, Clements J, Arndt W, Miller BL, Wheeler TJ, Schreiber F et al (2015) HMMER web server: 2015 update. Nucleic Acids Res 43(W1):W30–W38. doi:10.1093/nar/gkv397
Fitch WM (1970) Distinguishing homologous from analogous proteins. Syst Zool 19(2):99–113
Gabaldón T, Koonin EV (2013) Functional and evolutionary implications of gene orthology. Nat Rev Genet 14(5):360–366. doi:10.1038/nrg3456
Galindo A, Hervás-Aguilar A, Rodríguez-Galán O, Vincent O, Arst HN, Tilburn J, Peñalva MA (2007) PalC, one of two Bro1 domain proteins in the fungal pH signalling pathway, localizes to cortical structures and binds Vps32. Traffic (Copenhagen, Denmark) 8(10): 1346–1364. doi:10.1111/j.1600-0854.2007.00620.x
Ginalski K (2003) ORFeus: detection of distant homology using sequence profiles and predicted secondary structure. Nucleic Acids Res 31(13):3804–3807. doi:10.1093/nar/gkg504
Gray GS, Fitch WM (1983) Evolution of antibiotic resistance genes: the DNA sequence of a kanamycin resistance gene from *Staphylococcus aureus*. Mol Biol Evol 1(1):57–66
Grossberger R, Gieffers C, Zachariae W, Podtelejnikov AV, Schleiffer A, Nasmyth K et al (1999) Characterization of the DOC1/APC10 subunit of the yeast and the human anaphase-promoting complex. J Biol Chem 274(20):14500–14507
Gupta MK, Niyogi R, Misra M (2013) An alignment-free method to find similarity among protein sequences via the general form of Chou's pseudo amino acid composition. SAR QSAR Environ Res 24(7):597–609. doi:10.1080/1062936X.2013.773378
Heinicke S, Livstone MS, Lu C, Oughtred R, Kang F, Angiuoli SV et al (2007) The Princeton protein orthology database (P-POD): a comparative genomics analysis tool for biologists. PLoS ONE 2(8):e766. doi:10.1371/journal.pone.0000766
Herrero J, Muffato M, Beal K, Fitzgerald S, Gordon L, Pignatelli M et al (2016) Ensemble comparative genomics resources. Database: J Biol Databases Curation 2016, bav096. doi:10.1093/database/bav096
Höhl M, Ragan MA (2007) Is multiple-sequence alignment required for accurate inference of phylogeny? Syst Biol 56(2):206–221. doi:10.1080/10635150701294741
Höhl M, Rigoutsos I, Ragan MA (2006) Pattern-based phylogenetic distance estimation and tree reconstruction. Evol Bioinform Online 2:359–375
Huerta-Cepas J, Bueno A, Dopazo J, Gabaldon T (2007) PhylomeDB: a database for genome-wide collections of gene phylogenies. Nucleic Acids Res 36(Database), D491–D496. doi:10.1093/nar/gkm899
Hutterer A, Berdnik D, Wirtz-Peitz F, Zigman M, Schleiffer A, Knoblich JA (2006) Mitotic activation of the kinase Aurora-A requires its binding partner Bora. Dev Cell 11(2):147–157. doi:10.1016/j.devcel.2006.06.002
Ivliev AE, Sergeeva MG (2008) OrthoFocus: program for identification of orthologs in multiple genomes in family-focused studies. Js Bioinform Comput Biol 6(4):811–824
Johnson LS, Eddy SR, Portugaly E (2010) Hidden Markov model speed heuristic and iterative HMM search procedure. BMC Bioinform 11(1):431. doi:10.1186/1471-2105-11-431
Karwath A, King RD (2002) Homology induction: the use of machine learning to improve sequence similarity searches. BMC Bioinform 3(1):11. doi:10.1186/1471-2105-3-11
Kim S, Kang J, Chung YJ, Li J, Ryu KH (2008) Clustering orthologous proteins across phylogenetically distant species. Proteins 71(3):1113–1122. doi:10.1002/prot.21792
Kim B-H, Cheng H, Grishin NV (2009) HorA web server to infer homology between proteins using sequence and structural similarity. Nucleic Acids Res 37(Web Server issue), W532–8. doi:10.1093/nar/gkp328
Kim J, Ishiguro K-I, Nambu A, Akiyoshi B, Yokobayashi S, Kagami A et al (2015) Meikin is a conserved regulator of meiosis-I-specific kinetochore function. Nature 517(7535):466–471. doi:10.1038/nature14097
Kitajima TS, Kawashima SA, Watanabe Y (2004) The conserved kinetochore protein shugoshin protects centromeric cohesion during meiosis. Nature 427(6974):510–517. doi:10.1038/nature02312

Koonin EV (2005) Orthologs, paralogs, and evolutionary genomics. Annu Rev Genet 39(1):309–338. doi:10.1146/annurev.genet.39.073003.114725

Kristensen DM, Wolf YI, Mushegian AR, Koonin EV (2011) Computational methods for Gene Orthology inference. Briefings Bioinform 12(5):379–391. doi:10.1093/bib/bbr030

Kriventseva EV, Rahman N, Espinosa O, Zdobnov EM (2008) OrthoDB: the hierarchical catalog of eukaryotic orthologs. Nucleic Acids Res 36(Database issue), D271–5. doi:10.1093/nar/gkm845

Kueng S, Hegemann B, Peters BH, Lipp JJ, Schleiffer A, Mechtler K, Peters J-M (2006) Wapl controls the dynamic association of cohesin with chromatin. Cell 127(5):955–967. doi:10.1016/j.cell.2006.09.040

Kumar S (2011) Remote homologue identification of Drosophila GAGA factor in mouse. Bioinformation 7(1):29–32

Kumar A, Cowen L (2009) Augmented training of hidden Markov models to recognize remote homologs via simulated evolution. Bioinformatics (Oxford, England) 25(13): 1602–1608. doi:10.1093/bioinformatics/btp265

Kuziemko A, Honig B, Petrey D (2011) Using structure to explore the sequence alignment space of remote homologs. PLoS Comput Biol 7(10):e1002175. doi:10.1371/journal.pcbi.1002175

Lawo S, Bashkurov M, Mullin M, Ferreria MG, Kittler R, Habermann B et al (2009) HAUS, the 8-subunit human Augmin complex, regulates centrosome and spindle integrity. Current Biol: CB 19(10):816–826. doi:10.1016/j.cub.2009.04.033

Lee MM, Bundschuh R, Chan MK (2008) Distant homology detection using a LEngth and STructure-based sequence alignment tool (LESTAT). Proteins 71(3):1409–1419. doi:10.1002/prot.21830

Li L, Stoeckert CJ, Roos DS (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res 13(9):2178–2189. doi:10.1101/gr.1224503

Liu K, Raghavan S, Nelesen S, Linder CR, Warnow T (2009) Rapid and accurate large-scale coestimation of sequence alignments and phylogenetic trees. Science (New York, N.Y.) 324 (5934):1561–1564. doi:10.1126/science.1171243

Liu K, Warnow TJ, Holder MT, Nelesen SM, Yu J, Stamatakis AP, Linder CR (2012) SATe-II: very fast and accurate simultaneous estimation of multiple sequence alignments and phylogenetic trees. Syst Biol 61(1):90–106. doi:10.1093/sysbio/syr095

Liu B, Zhang D, Xu R, Xu J, Wang X, Chen Q et al (2014) Combining evolutionary information extracted from frequency profiles with sequence-based kernels for protein remote homology detection. Bioinformatics (Oxford, England) 30(4): 472–479. doi:10.1093/bioinformatics/btt709

Liu B, Chen J, Wang X (2015) Protein remote homology detection by combining Chou's distance-pair pseudo amino acid composition and principal component analysis. Mol Genet Genomics: MGG 290(5):1919–1931. doi:10.1007/s00438-015-1044-4

Makarova KS, Koonin EV, Kelman Z (2012) The CMG (CDC45/RecJ, MCM, GINS) complex is a conserved component of the DNA replication system in all archaea and eukaryotes. Biol Direct 7(1):7. doi:10.1186/1745-6150-7-7

Maulik U, Sarkar A (2013) Searching remote homology with spectral clustering with symmetry in neighborhood cluster kernels. PLoS ONE 8(2):e46468. doi:10.1371/journal.pone.0046468

Meier A, Söding J (2015) Context similarity scoring improves protein sequence alignments in the midnight zone. Bioinformatics (Oxford, England) 31(5): 674–681. doi:10.1093/bioinformatics/btu697

Mina JG, Okada Y, Wansadhipathi-Kannangara NK, Pratt S, Shams-Eldin H, Schwarz RT et al (2010) Functional analyses of differentially expressed isoforms of the Arabidopsis inositol phosphorylceramide synthase. Plant Mol Biol 73(4–5):399–407. doi:10.1007/s11103-010-9626-3

Mirarab S, Nguyen N, Warnow T (2012) SEPP: SATé-enabled phylogenetic placement. In: Pacific symposium on biocomputing. Pacific symposium on biocomputing, pp. 247–258. doi:10.1142/9789814366496_0024

Muda HM, Saad P, Othman RM (2011) Remote protein homology detection and fold recognition using two-layer support vector machine classifiers. Comput Biol Med 41(8):687–699. doi:10.1016/j.compbiomed.2011.06.004

Mudgal R, Sowdhamini R, Chandra N, Srinivasan N, Sandhya S (2014) Filling-in void and sparse regions in protein sequence space by protein-like artificial sequences enables remarkable enhancement in remote homology detection capability. J Mol Biol 426(4):962–979. doi:10.1016/j.jmb.2013.11.026

Mudgal R, Sandhya S, Kumar G, Sowdhamini R, Chandra NR, Srinivasan N (2015) NrichD database: sequence databases enriched with computationally designed protein-like sequences aid in remote homology detection. Nucleic Acids Res 43(Database issue), D300–5. doi:10.1093/nar/gku888

Murzin AG, Bateman A (1997) Distant homology recognition using structural classification of proteins. Proteins Suppl 1:105–112

Murzin AG, Brenner SE, Hubbard T, Chothia C (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 247(4):536–540. doi:10.1006/jmbi.1995.0159

NCBI Resource Coordinators (2016) Database resources of the national center for biotechnology information. Nucleic Acids Res 44(D1):D7–D19. doi:10.1093/nar/gkv1290

Nehrt NL, Clark WT, Radivojac P, Hahn MW (2011) Testing the ortholog conjecture with comparative functional genomic data from mammals. PLoS Comput Biol 7(6):e1002073. doi:10.1371/journal.pcbi.1002073

Nelesen S, Liu K, Wang L-S, Linder CR, Warnow T (2012) DACTAL: divide-and-conquer trees (almost) without alignments. Bioinformatics (Oxford, England) 28(12): i274–82. doi:10.1093/bioinformatics/bts218

Nishiyama T, Ladurner R, Schmitz J, Kreidl E, Schleiffer A, Bhaskara V et al (2010) Sororin mediates sister chromatid cohesion by antagonizing Wapl. Cell 143(5):737–749. doi:10.1016/j.cell.2010.10.031

Östlund G, Schmitt T, Forslund K, Köstler T, Messina DN, Roopra S et al (2010) InParanoid 7: new algorithms and tools for eukaryotic orthology analysis. Nucleic Acids Res 38(Database issue), D196–203. doi:10.1093/nar/gkp931

Ozlü N, Srayko M, Kinoshita K, Habermann B, O'toole ET, Müller-Reichert T et al (2005) An essential function of the C. elegans ortholog of TPX2 is to localize activated aurora A kinase to mitotic spindles. Dev Cell 9(2): 237–248. doi:10.1016/j.devcel.2005.07.002

Pelletier L, Ozlü N, Hannak E, Cowan C, Habermann B, Ruer M et al (2004) The Caenorhabditis elegans centrosomal protein SPD-2 is required for both pericentriolar material recruitment and centriole duplication. Current Biol: CB 14(10):863–873. doi:10.1016/j.cub.2004.04.012

Penel S, Arigon A-M, Dufayard J-F, Sertier A-S, Daubin V, Duret L et al (2009) Databases of homologous gene families for comparative genomics. BMC Bioinform 10 Suppl 6(Suppl 6), S3. doi:10.1186/1471-2105-10-S6-S3

Penkett CJ, Morris JA, Wood V, Bähler J (2006) YOGY: a web-based, integrated database to retrieve protein orthologs and associated gene ontology terms. Nucleic Acids Res 34(Web Server issue), W330–4. doi:10.1093/nar/gkl311

Perutz MF, ROSSMANN MG, CULLIS AF, MUIRHEAD H, WILL G, NORTH AC (1960) Structure of haemoglobin: a three-dimensional Fourier synthesis at 5.5-A. resolution, obtained by X-ray analysis. Nature 185(4711), 416–422

Powell S, Szklarczyk D, Trachana K, Roth A, Kuhn M, Muller J et al (2011) eggNOG v3.0: orthologous groups covering 1133 organisms at 41 different taxonomic ranges. Nucleic Acids Res 40(D1):D284–D289. doi:10.1093/nar/gkr1060

Proost S, Van Bel M, Vaneechoutte D, Van de Peer Y, Inzé D, Mueller-Roeber B, Vandepoele K (2015) PLAZA 3.0: an access point for plant comparative genomics. Nucleic Acids Res 43(Database issue), D974–81. doi:10.1093/nar/gku986

Pryszcz LP, Huerta-Cepas J, Gabaldón T (2011) MetaPhOrs: orthology and paralogy predictions from multiple phylogenetic evidence using a consistency-based confidence score. Nucleic Acids Res 39(5):e32–e32. doi:10.1093/nar/gkq953

Rabitsch KP, Gregan J, Schleiffer A, Javerzat J-P, Eisenhaber F, Nasmyth K (2004) Two fission yeast homologs of Drosophila Mei-S332 are required for chromosome segregation during meiosis I and II. Current Biol: CB 14(4):287–301. doi:10.1016/j.cub.2004.01.051

Remmert M, Biegert A, Hauser A, Söding J (2012) HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. Nat Methods 9(2):173–175. doi:10.1038/nmeth.1818

Rost B (1999) Twilight zone of protein sequence alignments. Protein Eng 12(2):85–94

Ruan J, Li H, Chen Z, Coghlan A, Coin LJM, Guo Y et al (2008) TreeFam: 2008 Update. Nucleic Acids Res 36(Database issue), D735–40. doi:10.1093/nar/gkm1005

Sánchez-Díaz A, González I, Arellano M, Moreno S (1998) The Cdk inhibitors p25rum1 and p40SIC1 are functional homologues that play similar roles in the regulation of the cell cycle in fission and budding yeast. J Cell Sci 111(Pt 6):843–851

Sandhya S, Mudgal R, Jayadev C, Abhinandan KR, Sowdhamini R, Srinivasan N (2012) Cascaded walks in protein sequence space: use of artificial sequences in remote homology detection between natural proteins. Mol BioSyst 8(8):2076–2084. doi:10.1039/c2mb25113b

Schreiber F, Sonnhammer ELL (2013) Hieranoid: hierarchical orthology inference. J Mol Biol 425 (11):2072–2081. doi:10.1016/j.jmb.2013.02.018

Schwickart M, Havlis J, Habermann B, Bogdanova A, Camasses A, Oelschlaegel T et al (2004) Swm1/Apc13 is an evolutionarily conserved subunit of the anaphase-promoting complex stabilizing the association of Cdc16 and Cdc27. Mol Cell Biol 24(8):3562–3576. doi:10.1128/MCB.24.8.3562-3576.2004

Sémon M, Wolfe KH (2007) Consequences of genome duplication. Curr Opin Genet Dev 17(6):505–512. doi:10.1016/j.gde.2007.09.007

Shah AR, Oehmen CS, Webb-RobertsonB-J (2008) SVM-HUSTLE–an iterative semi-supervised machine learning approach for pairwise protein remote homology detection. Bioinformatics (Oxford, England) 24(6): 783–790. doi:10.1093/bioinformatics/btn028

Shevchenko A, Roguev A, Schaft D, Buchanan L, Habermann B, Sakalar C et al (2008) Chromatin Central: towards the comparative proteome by accurate mapping of the yeast proteomic environment. Genome Biol 9(11):R167. doi:10.1186/gb-2008-9-11-r167

Shi G, Zhang L, Jiang T (2010) MSOAR 2.0: Incorporating tandem duplications into ortholog assignment based on genome rearrangement. BMC Bioinform 11(1):10. doi:10.1186/1471-2105-11-10

Sinha S, Lynn AM (2014) HMM-ModE: implementation, benchmarking and validation with HMMER3. BMC Res Notes 7(1):483. doi:10.1186/1756-0500-7-483

Söding J, Biegert A, Lupas AN (2005) The HHpred interactive server for protein homology detection and structure prediction. Nucleic Acids Res 33(Web Server issue), W244–8. doi:10.1093/nar/gki408

Söding J, Remmert M, Biegert A, Lupas AN (2006) HHsenser: exhaustive transitive profile search using HMM-HMM comparison. Nucleic Acids Res 34(Web Server issue), W374–8. doi:10.1093/nar/gkl195

Sonnhammer ELL, Östlund G (2015) InParanoid 8: orthology analysis between 273 proteomes, mostly eukaryotic. Nucleic acids research 43(Database issue), D234–9. doi:10.1093/nar/gku1203

Stingele J, Habermann B, Jentsch S (2015) DNA-protein crosslink repair: proteases as DNA repair enzymes. Trends Biochem Sci 40(2):67–71. doi:10.1016/j.tibs.2014.10.012

Studer RA, Robinson-Rechavi M (2009) How confident can we be that orthologs are similar, but paralogs differ? Trends Genet: TIG 25(5):210–216. doi:10.1016/j.tig.2009.03.004

Szklarczyk R, Wanschers BF, Cuypers TD, Esseling JJ, Riemersma M, van den Brand MA et al (2012) Iterative orthology prediction uncovers new mitochondrial proteins and identifies C12orf62 as the human ortholog of COX14, a protein involved in the assembly of cytochrome c oxidase. Genome Biol 13(2):R12. doi:10.1186/gb-2012-13-2-r12

Szklarczyk R, Wanschers BFJ, Nijtmans LG, Rodenburg RJ, Zschocke J, Dikow N et al (2013) A mutation in the FAM36A gene, the human ortholog of COX20, impairs cytochrome c oxidase

assembly and is associated with ataxia and muscle hypotonia. Hum Mol Genet 22(4):656–667. doi:10.1093/hmg/dds473

Tatusov RL, Koonin EV, Lipman DJ (1997) A genomic perspective on protein families. Science (New York, N.Y.) 278(5338):631–637

Vilella AJ, Severin J, Ureta-Vidal A, Heng L, Durbin R, Birney E (2009) EnsemblCompara GeneTrees: complete, duplication-aware phylogenetic trees in vertebrates. Genome Res 19 (2):327–335. doi:10.1101/gr.073585.107

Vinga S, Almeida J (2003) Alignment-free sequence comparison-a review. Bioinformatics (Oxford, England) 19(4): 513–523

Vogt G, Etzold T, Argos P (1995) An assessment of amino acid exchange matrices in aligning protein sequences: the twilight zone revisited. J Mol Biol 249(4):816–831. doi:10.1006/jmbi.1995.0340

Wagner I, Volkmer M, Sharan M, Villaveces JM, Oswald F, Surendranath V, Habermann BH (2014) morFeus: a web-based program to detect remotely conserved orthologs using symmetrical best hits and orthology network scoring. BMC Bioinform 15(1):263. doi:10.1186/1471-2105-15-263

Wang Y, Levy DE (2006) C. elegans STAT: evolution of a regulatory switch. FASEB J: Official Publ Fed Am Soc Exp Biol 20(10):1641–1652. doi:10.1096/fj.06-6051com

Watson HC, Kendrew JC (1961) The amino-acid sequence of sperm whale myoglobin. Comparison between the amino-acid sequences of sperm whale myoglobin and of human hemoglobin. Nature 190:670–672

Wieser D, Niranjan M (2009) Remote homology detection using a kernel method that combines sequence and secondary-structure similarity scores. Silico Biol 9(3):89–103

Wolf YI, Koonin EV (2012) A tight link between orthologs and bidirectional best hits in bacterial and archaeal genomes. Genome Biol Evol 4(12):1286–1294. doi:10.1093/gbe/evs100

Wu S, Zhang Y (2008) MUSTER: Improving protein sequence profile-profile alignments by using multiple sources of structure information. Proteins 72(2):547–556. doi:10.1002/prot.21945

Yamada K, Tomii K (2014) Revisiting amino acid substitution matrices for identifying distantly related proteins. Bioinformatics (Oxford, England) 30(3): 317–325. doi:10.1093/bioinformatics/btt694

Yang Y, Tantoso E, Li K-B (2008) Remote protein homology detection using recurrence quantification analysis and amino acid physicochemical properties. J Theor Biol 252(1):145–154. doi:10.1016/j.jtbi.2008.01.028

Yona G, Levitt M (2002) Within the twilight zone: a sensitive profile-profile comparison tool based on information theory. J Mol Biol 315(5):1257–1275. doi:10.1006/jmbi.2001.5293

Yu C, Desai V, Cheng L, Reifman J (2012) QuartetS-DB: a large-scale orthology database for prokaryotes and eukaryotes inferred by evolutionary evidence. BMC Bioinform 13(1):143. doi:10.1186/1471-2105-13-143

Zhang Z, Schäffer AA, Miller W, Madden TL, Lipman DJ, Koonin EV, Altschul SF (1998) Protein sequence similarity searches using patterns as seeds. Nucleic Acids Res 26(17):3986–3990

Index

P. Pontarotti (ed.), *Evolutionary Biology*, DOI 10.1007/978-3-319-41324-2