Pierre Pontarotti *Editor*

Evolutionary Biology: Self/Nonself Evolution, Species and Complex Traits Evolution, Methods and Concepts

Springer

Evolutionary Biology: Self/Nonself Evolution, Species and Complex Traits Evolution, Methods and Concepts

Pierre Pontarotti
Editor

Evolutionary Biology: Self/Nonself Evolution, Species and Complex Traits Evolution, Methods and Concepts

Springer

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

ISBN 978-3-319-61568-4 ISBN 978-3-319-61569-1 (eBook)
DOI 10.1007/978-3-319-61569-1

Library of Congress Control Number: 2017945231

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

For the tenth year, we have published a book on evolutionary biology concept and application.

We tried catch the evolution and progress of this field, and to achieve this goal we were helped by the Evolutionary Biology Meeting in Marseilles. The goal of this annual meeting is to allow the scientists of different disciplines, who share a deep interest in evolutionary biology concepts, knowledge and applications, to meet, exchange and enhance the interdisciplinary collaborations. The Evolutionary Biology Meeting in Marseilles is now recognised internationally as an important exchange platform and a booster for the use of evolutionary-based approaches in biology and in other scientific areas.

The book chapters have been selected from the meeting presentations and from propositions, conceived by the interaction of the meeting participants.

The reader of the evolutionary biology books as well as the meeting participants would like us to witness regularly during the different meetings and book editions, a shift in the evolutionary biology concepts. The fact that the chapters of the book are selected from a meeting enables, the quick diffusion of the novelties.

Also, we would like to underline that the ten books are complementary to one another and should be considered as tomes.

The articles are organised in the following categories

Self/Nonself Evolution (Chapters "A New View of How MHC Class I Molecules Fight Disease: Generalists and Specialists"–"The Life History of Domesticated Genes Illuminates the Evolution of Novel Mammalian Genes").

Species Evolution and Evolution of Complex Traits (Chapters "Evolution of Complex Traits in Human Populations"–"Modelling the Evolution of Dynamic Regulatory Networks: Some Critical Insights").

Methods and Concepts (Chapters "Mechanistic Models of Protein Evolution"–"Case Studies of Seven Gene Families with Unusual High Retention Rate Since the Vertebrate and Teleost Whole-Genome Duplications").

Marseille, France
May 2017

Marie Hélène Rome
A.E.E.B Director
Pierre Pontarotti
A.E.E.B and CNRS

Acknowledgements

We would like to thank all the authors and the reviewers of the different chapters.

We thank the sponsors of the meeting: Aix Marseille Université, CNRS, ITMO, ECCOREV FEDERATION, Conseil Général 13, ITMO, Ville de Marseille.

We wish to thank the A.E.E.B team for the organisation of the meeting.

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

Marseille, France
May 2017

Marie Hélène Rome
A.E.E.B Director
Pierre Pontarotti
A.E.E.B and CNRS

Contents

Part I
Self/Nonself Evolution

A New View of How MHC Class I Molecules Fight Disease: Generalists and Specialists

Jim Kaufman

Abstract Animals respond to the enormous onslaught of potential pathogens by a huge variety of defences, ranging from cell-intrinsic mechanisms to highly sophisticated cellular and molecular immune systems (Murphy and Weaver 2016; Owen et al. 2013).

1 Introduction

Animals respond to the enormous onslaught of potential pathogens by a huge variety of defences, ranging from cell-intrinsic mechanisms to highly sophisticated cellular and molecular immune systems (Murphy and Weaver 2016; Owen et al. 2013). Most such defences require the immune systems at some point to tell the difference between self and non-self, which in turn requires a level of specific recognition. Such recognition works in a variety of ways, from a disruption of cellular homoeostasis resulting in the presence of stress signals (or danger-associated molecular patterns, DAMPs), to recognition of essential molecular structures (or pathogen-associated molecular patterns, PAMPs) broadly common to a group of pathogens but different from the host, to exquisitely precise recognition capable in principle of recognizing any molecule but constrained by immunological tolerance mechanisms to effective recognition of non-self-molecules.

Decades of investigation in mammals and chickens have identified several different lymphocytes that contribute to precise recognition of molecules (which become known as antigens, once they are recognized) in jawed vertebrates, from sharks to humans (Murphy and Weaver 2016; Owen et al. 2013). For instance,

J. Kaufman (✉)
Department of Pathology, University of Cambridge, Tennis Court Road, Cambridge CB2 1QP, UK
e-mail: jfk31@cam.ac.uk

J. Kaufman
Department of Veterinary Medicine, Madingley Road, Cambridge CB2 0ES, UK

P. Pontarotti (ed.), *Evolutionary Biology: Self/Nonself Evolution, Species and Complex Traits Evolution, Methods and Concepts*,
DOI 10.1007/978-3-319-61569-1_1

B cells produce antibodies, soluble effector molecules which recognize molecular shapes of antigens. The so-called γδ T cells use their cell surface γδ T cell receptors (TCRs) generally to recognize molecular shapes on other cell surfaces. The αβ T cells use their αβ TCRs to recognize antigen bound to the so-called MHC molecules, cell surface proteins most of which are encoded in the major histocompatibility complex (MHC) or similar regions. The antigen is usually in the form of a peptide derived from proteins by intracellular degradation, but in some cases can be a lipid. The antibody from B cells can be very effective against extracellular pathogens, but the T cells are constrained to bind antigen on the surface of cells, in general for the detection of intracellular pathogens.

The hallmark of these lymphocyte receptors (antibodies and TCRs) is that their genes are not present in their final forms within germline and most somatic cells, but are created in lymphocytes by various mechanisms of DNA modification, resulting in a vast repertoire of receptors with different recognition specificities (Murphy and Weaver 2016; Owen et al. 2013). Generally, there is a single such receptor expressed on each cell, so that the receptor repertoire is distributed clonally. These clones are subject to a variety of controlling mechanisms to avoid too much recognition of self. As the first control, αβ T cells develop in the thymus, where a vast number of thymocytes undergo positive selection to ensure that each selected thymocyte binds to the self-MHC molecules present (which are bound to self-peptides) with sufficient affinity to be useful later. Then, the positively selected thymocytes undergo negative selection to ensure that surviving thymocytes do not interact too strongly with those same MHC molecules bound to self-peptides. The processes in the thymus are intended to result in a TCR repertoire that can recognize self-MHC molecules but not when bound to self-peptides, but are ready to recognize self-MHC molecules bound to non-self-peptides. There are several other such tolerance mechanisms which operate once the T cells leave the thymus.

Jawless fish also appear to have an adaptive immune with many similarities to the familiar adaptive immune system of jawed vertebrates (Boehm et al. 2012a, b). Lampreys and hagfish lack genes with the characteristics of antibodies, TCRs and MHC molecules, but instead have variable lymphocyte receptors (VLRs) which are also diversified in somatic cells. Indeed, cells with transcriptomes much like B lymphocytes secrete VLR-B molecules, while cells that bear cell surface VLR-A and VLR-C molecules travel to the tips of the gill arches, spending time in the so-called thymoids which have some characteristics of the thymus of jawed vertebrates. Thus, it would appear that a cellular immune system that differentiated self- and non-self-molecules existed in the common ancestor of jawless fish and jawed vertebrates, although the presence of molecules analogous to MHC molecules has not been demonstrated.

The major thrust of this chapter is to discuss recent discoveries and speculative concepts for one kind of MHC molecule, the classical class I molecules. A full background on such molecules would be a chapter in itself, but it is important to understand that classical class I molecules present peptide antigen derived primarily from proteins in the cytoplasm and contiguous structures like the nucleus, where viruses (and a very few intracellular bacteria) replicate. Various self-proteins are

involved in the production of peptides, their transport to the lumen of the endoplasmic reticulum and binding to MHC class I molecules, processes known overall as antigen processing and peptide loading. Decades of research has detailed the biochemistry, cell biology and genetics of these processes (Blum et al. 2013; Trowsdale and Knight 2013), and one might be forgiven for believing that all the fundamental principles had been discovered. However, there remains much to be understood. This chapter sketches out the discovery of a suite of correlated properties of classical class I molecules that appear to be important in resistance to disease and may have important consequences for basic immunology, for human and veterinary medicine, and even for evolutionary biology, ecology and conservation. It is very early days yet, with much to learn and fit together, but this new view of class I molecules that function as generalists and specialists may ultimately require a reassessment of decades of research. The story starts with the discoveries over many decades of strong genetic associations of a chicken blood group with resistance and susceptibility to economically important infectious diseases.

2 The chicken MHC can determine life and death from infectious pathogens

Almost everything known about the MHC comes from research into humans and biomedical models such as mice and rats. The MHC was discovered as the major genetic locus determining the success or failure of tissue transplants and was found to be very polymorphic. It is now known that such transplant rejection is primarily determined by the classical class I and class II molecules encoded in the MHC and that they are highly polymorphic, with thousands of alleles in humans (Tiercy and Claas 2013).

It is generally accepted that the high polymorphism of such MHC molecules is driven, not by transplantation, but primarily by a molecular arms race with pathogens (Bernatchez and Landry 2003; Spurgin and Richardson 2010). MHC molecules within a cell bind peptides and present them on the cell surface to T lymphocytes of the immune system, and T cell recognition of peptides derived from pathogens can trigger an appropriate immune response for protection. Pathogens are selected to evade the immune response by changing the sequence of the peptides that would bind the MHC molecules present in the host population, and this in turn selects for changes in the peptide-binding specificities of MHC molecules. This ongoing molecular arms race drives both variation in the pathogens and polymorphism in the MHC molecules.

Since particular MHC molecules may or may not bind appropriate peptides to stimulate a protective T cell response, one might expect the MHC to determine resistance and susceptibility to particular infectious pathogens. In humans, the MHC is the genetic region with the most disease associations, but the vast majority and the strongest associations are with autoimmune diseases. By contrast, the associations with resistance to infectious disease generally are relatively weak

(Trowsdale and Knight 2013; Hill 1998) and in some cases due to natural killer (NK) rather than T cell recognition (Martin et al. 2007). The best of these weak associations have been with disease from certain viral infections, such as human immunodeficiency virus (HIV), hepatitis B and C viruses, and human T cell lymphotrophic virus (HTLV), and have taken the best immunogeneticists with most sophisticated tools many decades to establish convincingly (Bangham 2009; Goulder and Walker 2012; McLaren et al. 2015).

In contrast, poultry scientists were stumbling over strong associations of the B blood group with resistance and susceptibility to economically important pathogens from the start of such research (Dietert et al. 1990; Plachy et al. 1992). The haplotype of the B locus has been shown to affect the outcome of infection by a long list of poultry pathogens, mostly viruses but including bacteria and parasites (Cotter et al. 1992; Schou et al. 2007). Among the viruses examined in detail was Rous sarcoma virus (RSV), long an important model infectious agent, notably the first retrovirus discovered and the source of the first oncogene discovered (Schierman and Collins 1987; Taylor 2004). An oncogenic herpesvirus, Marek's disease virus (MDV), has been a major subject of research due to the devastating effects on the poultry industry and notable as the first example of successful vaccination against cancer (Plachy et al. 1992; Schierman and Collins 1987). Similarly, strong genetic association of MHC loci with economically important pathogens has been shown in another farmed animal species outside of mammals, the Atlantic salmon (Grimholt et al. 2003). Moreover, the best evidence for ongoing selection of MHC genes in wild animals has been with birds and fish, compared to mammals (Bernatchez and Landry 2003).

3 A single class I gene is expressed at a high level in vertebrates outside of mammals

This difference between humans and chickens in genetic association of the MHC with infectious disease can be explained by the architecture of the MHC, which in mammals allows a multigene family of class I molecules to be co-expressed but in most non-mammalian vertebrates allows only a single class I molecule to be expressed at a high level (Kaufman 2013, 2014, 2015a, b). In this view (Fig. 1), the properties of the single dominantly expressed class I molecule determine whether the individual non-mammalian vertebrate lives or dies after infection with certain pathogens, which reads out as strong genetic associations. In contrast, each class I molecule expressed by a typical mammal has a chance of presenting a protective peptide so that altogether each MHC haplotype confers more-or-less resistance to most pathogens, which reads out as weak genetic associations. Thus, the relative lack of strong genetic associations of the human MHC with infectious disease is due to the fact that most human MHC haplotypes confer a similar high level of protection, without that much difference between them.

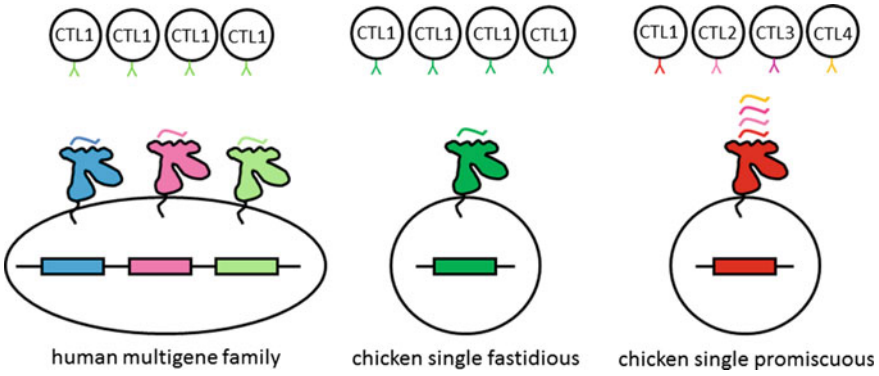


Fig. 1 The number of well-expressed class I loci may explain why human MHC haplotypes confer more-or-less resistance to most pathogens, while some chicken MHC haplotypes confer resistance and others confer susceptibility. Human MHC haplotypes express a multigene family of class I molecules (*light blue, pink, light green*), each one of which has a chance to find a peptide that confers protection to a given pathogen, so overall most haplotypes confer protection, which reads out as weak genetic associations with particular infectious pathogen. However, the eventual cytotoxic T cell (CTL) response narrows to a single clone (CTL1) recognizing a single immunodominant peptide on a single class I molecule. In contrast, each chicken MHC haplotype has a single dominantly expressed class I molecule, which may or may not find a peptide that confers protection to a given pathogen, so that there are big differences between haplotypes, which reads out as strong genetic associations. Some chicken class I molecules (*dark green*) have very fastidious peptide-binding motifs, so that they bind a very restricted variety of peptides, their chance of finding a peptide is low, and immunodominance means that only a single clone (CTL1) is likely to respond even if the peptide bound is protective. However, if both the peptide bound and the CTL clone responding are very effective, such a haplotype may confer protection to a particular pathogen in a specialized manner. Other chicken class I molecules (*dark red*) have a very promiscuous peptide-binding motif, so that they bind a wide variety of peptides and their chance of finding (several) protective peptides is high. In addition, it is proposed that a variety of clones (CTL1, CTL2, CTL3 and CTL4) would respond to these peptides, so that immunodominance is not so important, and overall, the chance of a protective response is high. However, if the pathogen is particularly virulent, a generalized response may not be enough to be protective. Not shown is the idea that the expression of each class I molecule at the cell surface is inversely correlated with the peptide repertoire. The idea that there is also a similar hierarchy of class I molecules in humans, but not to the same extent as in chickens is indicated by the colour (*pink human* vs. *red chicken molecules*, *light green human* vs. *dark green chicken molecules*.)

The organization of the human MHC is known in enormous detail (Beck and Trowsdale 2000; Horton et al. 2004) and outlined in every immunology textbook (Murphy and Weaver 2016; Owen et al. 2013). The mammalian MHC is large and divided by significant levels of recombination into regions (Dawkins et al. 1999) including a multigene family of class I molecules in the class I region, a multigene family of class II genes in the class II region (along with the genes involved in antigen processing and peptide loading) and in between a complex class III region with many different kinds of genes, some of which are involved in immune defence. In addition, some authorities consider the MHC to include the extended class II

region and the extended class I regions on the outside of the MHC, since they include some genes involved in immunity as well.

In humans, the highly polymorphic HLA-A, HLA-B and HLA-C genes are present in the class I region, but among the genes involved in antigen processing and peptide loading are those located far away in the class II and extended class II regions (Blum et al. 2013; Beck and Trowsdale 2000; Horton et al. 2004; Dawkins et al. 1999). The genes in the class I pathway include the transporter associated with antigen presentation (TAP) genes, TAP1 and TAP2, that together encode an ABC transporter that pumps peptides from the cytoplasm into the lumen of the endoplasmic reticulum to be loaded onto class I molecules. Nearby are some of the inducible proteasome (LMP) genes that encode proteases that adapt the cytoplasmic proteolytic complex to produce peptides suitable for binding to class I molecules. Finally, the tapasin gene encodes a dedicated chaperone and peptide editor to help certain class I molecules bind high-affinity peptides. Despite being located in the MHC, these antigen processing and peptide loading molecules are not functionally polymorphic and appear to work with all members and alleles of the class I multigene family.

In chickens, the BF-BL region of the B locus is the major determinant for graft rejection and so is the "major histocompatibility complex". The BF-BL region is very small and simple compared to the typical mammalian MHC (and thus was originally characterized as a "minimal essential MHC" and is still often called the "classical" or "core" MHC) and is also arranged differently than in mammals, with the class III region outside of the class I and class II regions (Kaufman 2014; Kaufman et al. 1999). There are two classical class I genes in chickens: the poorly expressed BF1 gene, which is crippled by mutations and deletions in the promoter or rendered a pseudogene by an insertion, and the dominantly expressed BF2 gene whose properties can determine the immune response. In between the two class I genes are the TAP1 and TAP2 genes, and nearby is the tapasin gene; the inducible proteasome components appear to have been deleted in chickens and other birds (Kaufman 2015).

In contrast to typical mammals, there is little recombination across the BF-BL region so that the MHC evolves as relatively stable haplotypes of polymorphic interacting genes. The chicken TAP1, TAP2 and tapasin genes are highly polymorphic and moderately diverse in sequence, with different alleles in every haplotype. In each haplotype, the peptide translocation specificity of the TAP is similar to the peptide-binding specificity of the dominantly expressed BF2 molecule, so that the minor BF1 molecule may not receive many peptides and therefore may fall into disuse over evolutionary time (Walker et al. 2011). A similar co-evolution appears to occur for tapasin (van Hateren et al. 2013).

The salient features of the MHC and class I system in chickens are found in many if not most non-mammalian vertebrates and are thought to be ancestral, with the mammalian MHC arising by an inversion (Kaufman 2014; Kaufman 1999). At this point in time, the evidence is scattered, fragmented and incomplete, but almost all examples can be fitted into the framework provided by the chicken. Among the many examples: there are five class I genes in ducks located next to the

polymorphic TAP genes, but only one is reported to be well-expressed (Mesa et al. 2004; Moon et al. 2005); there is a single classical class I gene in the frog *Xenopus* closely linked to polymorphic TAP and inducible proteasome genes (Ohta et al. 2006); and there is a single classical class I gene in the Atlantic salmon closely linked to the TAP2 gene (Lukacs et al. 2007). Taking the chicken as the general template for the ancestral MHC, the easiest mechanism to produce the typical mammalian MHC would be a simple inversion that swapped the class I gene(s) around with the class III region, but left the TAP, tapasin and inducible proteasome components behind to merge with the class II region. The particular alleles of antigen processing and peptide loading genes in the class II region could not be kept together with their partner class I allele, and so the co-evolutionary relationships would break down. In this view, the evolutionary solution was to select for monomorphic antigen processing and peptide loading genes that would function well with any class I allele that could appear by recombination, and once that happened, a multigene family could be supported.

4 MHC-determined resistance to Marek's disease and other infectious diseases correlates with cell surface expression levels of chicken class I molecules

Thus far, the model based on the chicken MHC can explain much of the existing data and resolve mysteries such as the weak genetic associations of the mammalian MHC with infectious pathogens, the presence of the class III region in between the class I and class II regions, and the presence of the antigen processing and peptide loading genes for class I molecules in the class II region. Moreover, the mechanism for selection of polymorphism in chickens is not terribly different from those worked out so clearly for humans and mice: there can be protective response only if a particular class I molecule presents a peptide from a pathogen to T cells. Indeed, the selective pressure on chickens should be much stronger than on mammals, since there is only one gene (so two chances in a heterozygote) to find such a protective peptide.

However, there were some uncomfortable questions that arose from this model. For instance, if a mutation in one chicken MHC gene requires a compensatory change in the other gene(s) to maintain fitness, then evolution would be slowed down enormously in chickens compared to mammals. And if there was an inversion in the lineage on the way to placental mammals, how did the first affected individual survive while the antigen processing and peptide loading genes accrued the changes necessary to become a monomorphic average best fit for all possible class I molecules? A potential answer eventually grew out of another uncomfortable issue, the strong genetic association of the chicken MHC with resistance to Marek's disease.

Marek's disease was originally described as fowl paralysis, due to peripheral neuropathy in association with lymphocytes, but became a major scourge of the poultry industry due to T cell tumours that arose in susceptible birds. Eventually,

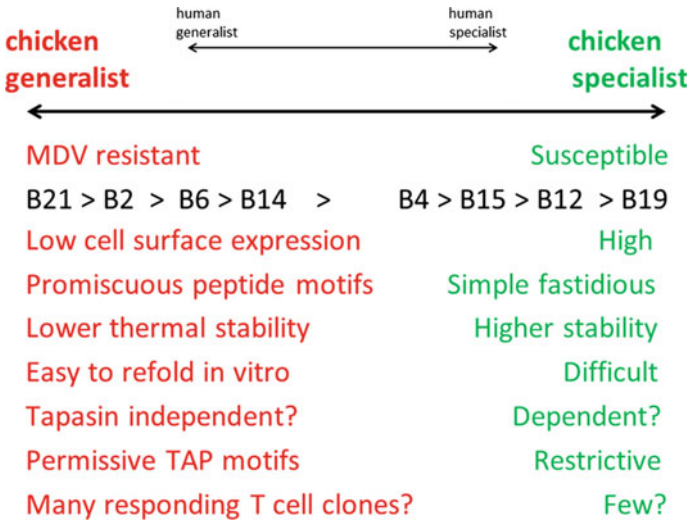


Fig. 2 There is a hierarchy of chicken class I molecules, which vary in a suite of properties, with some indication that a more limited hierarchy exists for human class I molecules. The rank order of chicken B haplotypes for historic resistance to Marek's disease reflects the rank order of certain properties of the dominantly expressed class I molecule, directly correlated with breadth of peptide repertoire and ease of in vitro refolding (as well as breadth of peptide translocation by TAPs), and inversely correlated with cell surface expression and thermal stability. The level of tapasin dependence and number of responding T cell clones are under current investigation, but it is proposed that tapasin dependence is inversely correlated and effective T cell clone number is directly correlated with peptide repertoire. The promiscuous class I molecules are proposed to be generally protective against a range of pathogens, with the fastidious class I molecules acting as specialists by presenting particularly protective peptides from certain pathogens. A narrower range of promiscuous and fastidious molecules has been proposed for human class I molecules, based on peptide repertoire and cell surface expression. The ease of in vitro refolding of the chicken molecules had been noted but not understood as a correlated feature until the report of the phenomenon for human class I molecules

it was found that Marek's disease was caused by MDV, an oncogenic herpesvirus with a complicated lifestyle (Calnek 1992; Osterrieder et al. 2006). Although the disease is mostly controlled in commercial flocks by vaccination with attenuated MDV strains (Witter 1998), it was found very early on that the B locus (later refined to the BF-BL region) confers resistance and susceptibility, with a hierarchy from the most resistant B21 haplotype to the most susceptible B19 haplotype (Plachy et al. 1992). The genetic association was very strong, explaining as much as 50% of the variance, but the basis for susceptibility was difficult to understand, since even a class I molecule with a very selective binding specificity would be expected to find a protective peptide among 100 MDV genes. Since the MHC association was with the response to tumours rather than infection by the virus, one possibility was that only a few viral proteins are expressed in tumour cells. Trying to answer this question eventually led to our new view of how MHC class I molecules prevent disease.

Many studies over decades established a rough hierarchy of B haplotypes for resistance to Marek's disease (Fig. 2) (Plachy et al. 1992). There was not universal agreement about the exact order for each haplotype, but given the wide range of viral strains, dose and route of infection, overall host genotype and many other factors, it is in fact surprising that any pattern could emerge. Long ago, it was found that the level of class I molecules on the surface of red blood cells varies as much as tenfold between MHC haplotypes and that the rank order of expression correlates inversely with the resistance of the B haplotype to Marek's disease, with the class I molecules from the most resistant B21 haplotype being expressed at the lowest level and the most susceptible B19 haplotype expressed at the highest level (Kaufman et al. 1995). Pulse-chase experiments showed that all haplotypes produce roughly the same amount of class I protein, but vary in the amount that moves to the cell surface (Tregaskes et al. 2016). Initially, the hypothesis was that this cell surface expression polymorphism would affect NK cell recognition of MDV-infected cells, but the ensuing work suggested that the important interaction is with T cells.

5 Cell surface expression levels of chicken class I molecules correlate with a suite of properties including diversity of peptide binding

The initial analyses of peptides bound to chicken class I molecules of the B4/B13, B12, B15 and B19 haplotypes gave clear and simple peptide-binding motifs with two or three anchor residues, each with only one (or two chemically similar) residues. These stringent (or "fastidious") motifs were much like those found in humans and mice and could be used to explain the responses to pathogens and vaccines (Kaufman et al. 1995; Wallny et al. 2006; Butter et al. 2013).

However, it was much harder to figure out the motifs from the B2, B14 and B21 haplotypes, both because the amount of peptide was much less (now known to be due to the fact that there are fewer class I molecules on the cell surface) and because there were no positions with only one (or two chemically similar) residues. Eventually, binding studies and X-ray crystallography showed that all three of these low-expressing class I molecules had "promiscuous" motifs, binding an astonishing variety of peptides. The dominantly expressed class I molecule from the B21 haplotype remodelled the binding site to accommodate peptides with completely different sequences, a mode of binding never reported for mammals. The dominantly expressed class I molecules from the B2 and B14 haplotypes bound peptides using anchor resides at position 2 and at the C-terminal position, much like many mammalian class I molecules, but the binding pockets were broad and would accommodate a wider range of hydrophobic amino acids than ever described for humans and mice (Koch et al. 2007; Chappell et al. 2015).

This inverse correlation between the diversity of peptides bound (or "peptide repertoire") and the cell surface expression level is only one of several with other

properties that vary between these chicken class I molecules (Kaufman 2015; Tregaskes et al. 2016) (Fig. 2). Compared to the peptide-binding specificity of the class I molecule, the peptide translocation specificity of the TAPs varies from wider in the low-expressing B21 haplotype to even more stringent in the high expressing B15 haplotype. The thermostability of the class I molecules correlates with the expression level, with the class I molecules from B21 cells less stable overall than from B19 cells. The ease of renaturing class I molecules with specific peptides in vitro to form sufficient complexes also varies, being much easier to accomplish for low-expressing class I molecules. It remains unclear to what extent all these properties vary along a hierarchy versus between two or more major groups of molecules.

One question that immediately arises is how the wider peptide repertoire of promiscuous class I molecules can confer a greater resistance to Marek's disease. We have proposed that the breadth of peptide presentation leads to a breadth of responding T cells, such that protection is conferred by the prolonged response of many T cell clones (Chappell et al. 2015). This idea flies in the face of an important concept for CD8 responses in mammals called immunodominance, in which CD8 responses start with many T cell clones but very rapidly narrow to a single T cell clone focused on an "immunodominant peptide". This concept is invoked to explain many experimental results in mammals, with multiple mechanisms by which it occurs (Yewdell and Bennink 1999). We predict that immunodominance is important for relatively fastidious class I molecules in mammals and in chickens, but that it does not occur for very promiscuous molecules. One possibility is that the extent of immunodominance is inversely correlated with peptide repertoire. Alternatively, promiscuous molecules may simply have more of a chance to bind and present an immunodominant protective peptide, compared to a fastidious molecule.

Another immediate question concerns the function of the cell surface expression-level polymorphism. One possibility is that the basis for the differences is mechanistic, with some biochemical trade-off in peptide loading, peptide editing or quality control in which many peptides binding at a lower affinity mean less transport to the cell surface. Another possibility could be a selective pressure to avoid more autoimmunity, in which fewer class I molecules on the cell surface balance the greater number of peptides that might result in autoimmune responses. However, our favourite proposal is based on the old idea that more MHC molecules mean fewer T cells survive negative selection (von Boehmer et al. 1989; Vidović and Matzinger 1988; Nowak et al. 1992). In this view (Fig. 3), an MHC molecule presenting a wide range of peptides might allow a much better T cell response in the periphery but would also result in a much greater deletion of T cells by negative selection in the thymus. Thus, we propose that the expression-level polymorphism comes about over evolutionary time, with the level of cell surface expression balancing the promiscuity of peptide presentation to give an optimal T cell repertoire (Chappell et al. 2015).

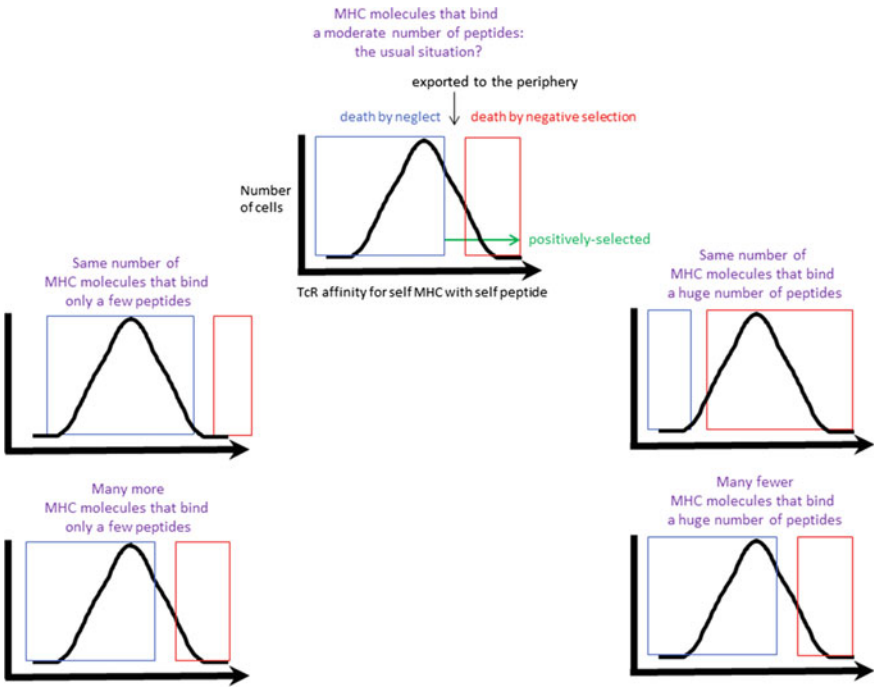


Fig. 3 The balance of peptide repertoire with cell surface expression level of class I molecules could result in an optimal T cell receptor repertoire after selection in the thymus. It is widely agreed that thymocytes with a particular affinity for self-MHC molecules bearing self-peptides are positively selected in the thymus (*green arrow* in the top panel), while those that do not bind with sufficient affinity do not receive survival signals and die by neglect (*blue boxes*). Those thymocytes that bind with a very high affinity to self-MHC molecules bearing self-peptides are negatively selected by apoptosis (*red boxes*). The top panel illustrates the situation considered in the literature, for which every MHC molecule is expressed at roughly the same level with the same peptide repertoire. In the panels on the middle level, the outcome of the same number of MHC molecules that bind either very few peptides or very many peptides is envisaged; in either case, a pauperized repertoire ensues. In the panels on the bottom level, the outcome of MHC molecules that bind very few peptides but are expressed at a higher level or MHC molecules that bind very many peptides but are expressed at a lower level is envisaged; in both cases, the average affinity of the T cell receptors on thymocytes after selection is restored to the normal level

6 Many of these properties are correlated among human class I molecules

Given the many important differences in MHC structure and function between placental mammals and non-mammalian vertebrates, it was a question whether the suite of properties described above is limited to chickens. A bioinformatics and modelling paper provided evidence that certain HLA-B alleles vary in the number of self-peptides predicted to bind and that the rank order correlates inversely with the odds ratio of acquired immunodeficiency syndrome (AIDS). In particular,

the HLA-B*57:01 and HLA-B*27:05 alleles were known to confer long-term non-progression from HIV infection to AIDS and were predicted to bind few peptides, whereas HLA-B*07:02 and particularly HLA-B*35:01 were known to lead to rapid progression and were predicted to bind many peptides. The authors created a model for the generation of cross-reactive T cell clones in the thymus to propose an explanation for the experimental data (Kosmrlj et al. 2010). Pertinent to our interest, we used several monoclonal antibodies to examine blood lymphocytes and monocytes by flow cytometry and found the same inverse correlation of cell surface expression level and peptide repertoire for these class I molecules as in chickens (Chappell et al. 2015). To be clear, this phenomenon seems to be completely different from the difference in expression between HLA-C alleles, which is determined at the level of RNA (Thomas et al. 2009; Kulkarni et al. 2011; Apps et al. 2013).

In another study (Paul et al. 2013), many human class I molecules were compared by prediction of peptide binding versus direct peptide binding, showing that the diversity of peptides bound by human HLA-A and–B alleles varies over a wide range, with HLA-B alleles generally (but certainly not always) having narrower peptide repertoires than HLA-A alleles. The highest diversity was found for HLA-A2 alleles, which are known to have motifs with hydrophobic amino acids at position 2 and at the C-terminal position of the peptide, much like chicken class I molecule from the B2 haplotype. Any motif that depends on hydrophobic amino acids will predict a lot of peptides, since most hydrophobic amino acids are very common in proteins. However, each of the HLA-A2 alleles is restricted to just a few such amino acids; for instance, HLA-A*02:01 binds only leucine or methionine at position 2 in a narrow pocket (Guo et al. 1993), while the chicken B2 molecule binds a wide variety of hydrophobic amino acids (Chappell et al. 2015). Overall, it appears that the range of peptide repertoire may be somewhat less in humans, with the most promiscuous chicken molecules being more promiscuous than the most promiscuous human molecules and the most fastidious chicken molecules being more fastidious than the most fastidious human molecules (Fig. 2).

In addition to the peptide-binding motif of a human class I molecule, some aspects of peptide loading appear to be important, just as in chickens. Differences in the transport of certain class I alleles to the cell surface were reported long ago (Neefjes and Ploegh 1988). More recently (Rizvi et al. 2014), a systematic study of 27 HLA-B alleles showed that, in the absence of tapasin, tapasin-independent alleles generally are better expressed on the cell surface than tapasin-dependent alleles (as assessed by flow cytometry after transfection in cell lines) and are more stable in the absence of peptide and more competent to assemble with peptide (as assessed by size exclusion chromatography after by renaturation). The tapasin-independent alleles are also more likely to lack the Bw4 epitope and to be associated with progression to AIDS. The properties that vary with tapasin dependence across these HLA-B alleles also correlate roughly with the peptide repertoire as predicted or measured (Fig. 4). In particular, HLA-B*57:01 and HLA-B*27:05 are tapasin-dependent and fastidious while HLA-B*07:02 and HLA-B*35:01 are tapasin-independent and promiscuous, which also fits the rank

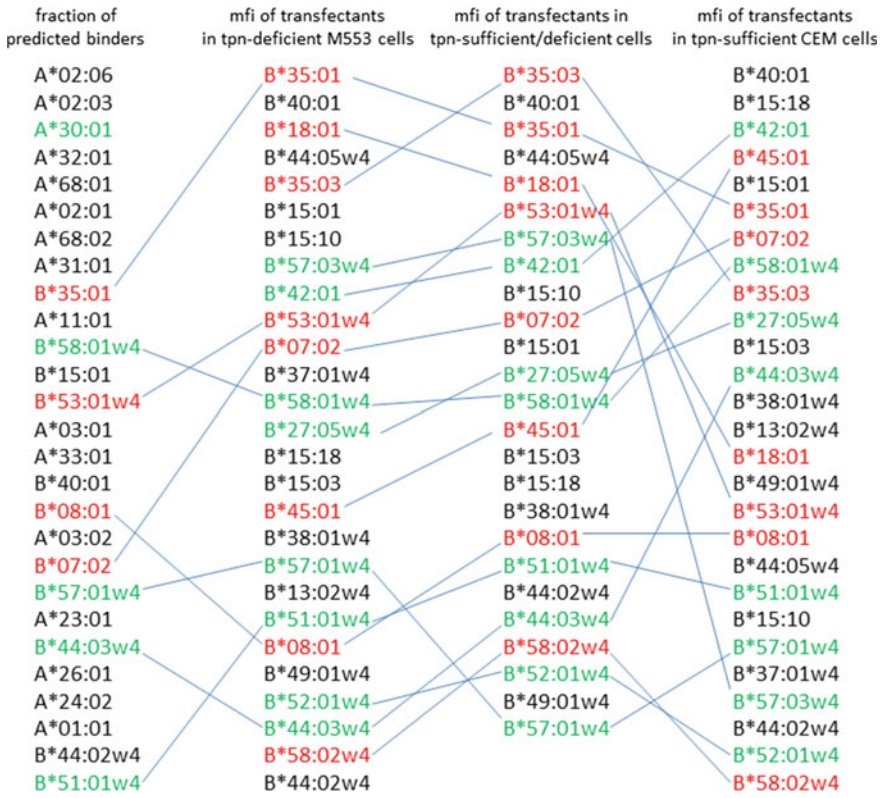


Fig. 4 Peptide repertoire, tapasin-independence, serological Bw4 epitope and progression to AIDS are (only) loosely correlated for certain HLA-A and HLA-B alleles. Columns give rank order of HLA alleles for fraction of predicted binding peptides (Nowak et al. 1992), mean fluorescence intensity (mfi) of cell surface expression of HLA alleles transfected into tapasin (tpn)-deficient M553 cells, ratio of the mfi for HLA alleles transfected into M553 cells with and without transduced tpn, and mfi of HLA alleles transfected into CEM cells that normally express tpn (Kulkarni et al. 2011). The rank orders are reasonably similar, given that many entries do not differ at a level of statistical significance (consult original references for further information). HLA-B alleles that bear the Bw4 epitope recognized by some NK cells are labelled "w4". HLA alleles that are associated with slow or no progression from HIV infection to AIDS (i.e. protective alleles) are coloured *green*, while those that are associated with faster progression (susceptibility alleles) are coloured *red*; note that HLA-B*51:01 is considered a protective allele for clade B virus in Caucasians, but a susceptibility allele for clade C virus in Africans (Goulder and Walker 2012). These disease-associated alleles are connected by lines in the figure, to help indicate the level of similarity between rank orders based on different properties

order for expression level in the presence of tapasin in ex vivo cells (Chappell et al. 2015). Thus, the hierarchy appears similar in nature to what was found with chickens. However, overall the mean fluorescence intensity in the transfectants in the presence of tapasin differed by less than two fold between the highest and the lowest expressed alleles.

7 Generalist and specialist class I molecules: a new paradigm?

The apparent similarities of the class I hierarchy between chicken and human class I molecules, and the correlations with resistance to infectious disease, suggest that these properties are fundamental and important in the function of class I molecules. However, the low-expressing promiscuous class I molecules confer resistance to Marek's disease in chickens, while the high expressing fastidious molecules confer resistance to AIDS in humans. How does this all fit together?

Looking through the literature, it became clear that the low-expressing promiscuous class I molecules can confer resistance, not just to Marek's disease, but to several other viruses. For instance, tumours arising from most strains of the retrovirus Rous sarcoma virus progressed in chickens with the fastidious B13 haplotype (with a class I identical to B4), but regressed in chickens with the promiscuous B6 haplotype (McBride et al. 1981). Chickens in Thai villages with the promiscuous B21 and B2 haplotypes survived natural influenza infection (as homozygotes and all but one heterozygote combination), while the chickens bearing only fastidious B12, B4/13, B15 and B19 haplotypes died (Boonyanuwat et al. 2006) (Fig. 5). After inoculation with infectious bronchitis virus, chickens with the promiscuous B2 haplotype had much less clinical respiratory illness than those with the fastidious B12 or B19 haplotypes (Banat et al. 2013). In fact,

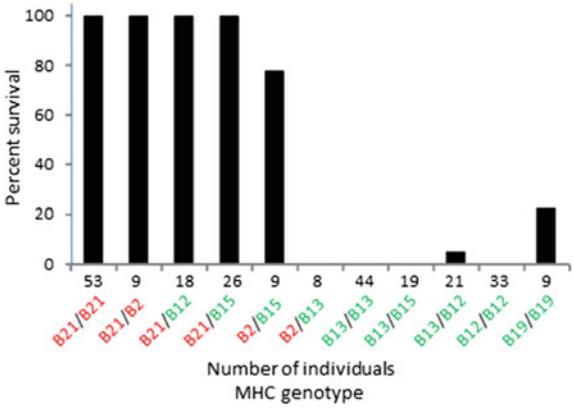


Fig. 5 The MHC haplotypes associated with promiscuous class I molecules can protect from mortality due to influenza infection in the field. Data abstracted from Table 1 of reference 51, in which Leung-Hahng-Kow indigenous chickens sourced from rural area of Thailand during an avian influenza outbreak were typed for MHC using single-strand conformational polymorphism analysis. From the text, B2/B2 and B6/B6 homozygotes also survived at 100%, but were present only at very low frequencies so were not included in the data. The susceptibility of B2/B13 birds is not explained by a model in which resistance is conferred by promiscuous class I molecules, and remains unexplained. Also, these are only correlations with MHC haplotype rather than proof that the class I molecules are responsible (rather than other genes within the haplotype or epistatic interactions with genes outside the haplotype)

we have not found good examples in which high expressing fastidious class I molecules confer resistance when compared to low-expressing promiscuous class I molecules. Based on these and other studies, We proposed (Chappell et al. 2015) that the low-expressing promiscuous class I molecules are generalists, providing resistance to a variety of pathogens. One might consider the peptide-binding specificity of a single promiscuous class I molecule as equivalent to several fastidious class I molecules, conceptually similar to our view of human MHC haplotypes, in which a multigene family of class I molecules confers more-or-less resistance to most pathogens.

In contrast to chickens, it is the high expressing fastidious class I molecules that are reported to confer resistance to certain infectious pathogens in humans. For instance, many groups have found that the low-expressing promiscuous class I molecules HLA-B*35:01 and HLA-B*07:02 are associated with progression to AIDS, while the high expressing fastidious HLA-B*57:01 and B*27:05 are associated with non-progression (Goulder and Walker 2012; Carrington et al. 1999; International HIV Controllers Study, Pereyra et al. 2010). Several mechanisms have been proposed for the protective response by the fastidious class I molecules, but the best evidence is that both HLA-B*57:01 and B*27:05 find particular peptides (from the gag protein) that elicit an effective cytotoxic T lymphocyte response, from which the virus cannot escape by mutation without a dramatic loss of fitness. These special peptides confer protection from which the virus has difficulty escaping, so that progression is dramatically slowed. We have proposed that these fastidious class I molecules act as specialists (Chappell et al. 2015), with particular alleles able to bind protective peptides for particular virus species or strains. In this view, the promiscuous HLA-B*35:01 and B*07:02 alleles function as generalists (usually dealing with most pathogens), but simply cannot find peptides for HIV that are protective in the long run.

An interesting consequence of this view is that populations with only promiscuous generalist alleles might be expected to be resistant to most common pathogens. Indeed, very few or even one promiscuous allele(s) might be enough to protect a population. There are examples of populations or species of wild animals that have few class I alleles, for example some moose, bison and deer species (Ellegren et al. 1996; Babik et al. 2012; Zhang et al. 2012). Interestingly, the idea of generalist and specialist MHC alleles has already been suggested on functional grounds for class II molecules of the striped mouse with regard to parasite infestation (Froeschke and Sommer 2012). Most studies suggest that class II molecules bind a much wider variety of peptides than class I molecules, and the existence of promiscuous (and perhaps ancestral) class I molecules has a satisfying resonance with the proposal that class II molecules were the ancestor of class I molecules (Kaufman et al. 1984; Kaufman 2011). In any case, conservation scientists often type MHC genes (usually class II B chain genes), arguing that low levels of allelic polymorphism (and/or low sequence diversity between alleles) signal potential inability to confer protection against diverse and evolving pathogens (as well as correlating with overall low diversity across the genome). However, a population with a few generalist MHC alleles (and with reasonable levels of diversity across

the genome) might not be in as much danger as a population with many specialist MHC alleles.

In fact, chickens and humans have a mixture of promiscuous and fastidious alleles, so how would this come about? We expect the promiscuous generalist alleles to be ancestral, so that an ancestral population might have generalist alleles at high frequency, with low frequencies of fastidious specialist alleles that arise in the normal course of mutation, drift and low levels of selection. However, the appearance of a new and particularly virulent pathogen for which the promiscuous generalists are not able to cope would strongly select for any fastidious alleles that could confer resistance, as specialists for that particular pathogen (Fig. 6). If the population did not have any relevant specialist alleles, it might be wiped out. If the population survives and then no new pathogens appear for a considerable period of time, then one might expect the promiscuous generalist alleles again come to dominate the population. However, if the population experiences new pathogens from time to time, eventually one might expect both promiscuous generalists and fastidious specialist alleles would be found at moderate gene frequencies. An interesting example is the chimpanzee, which is known to have two kinds of class I alleles (Sidney et al. 2006; de Groot et al. 2010), those with peptide motifs almost

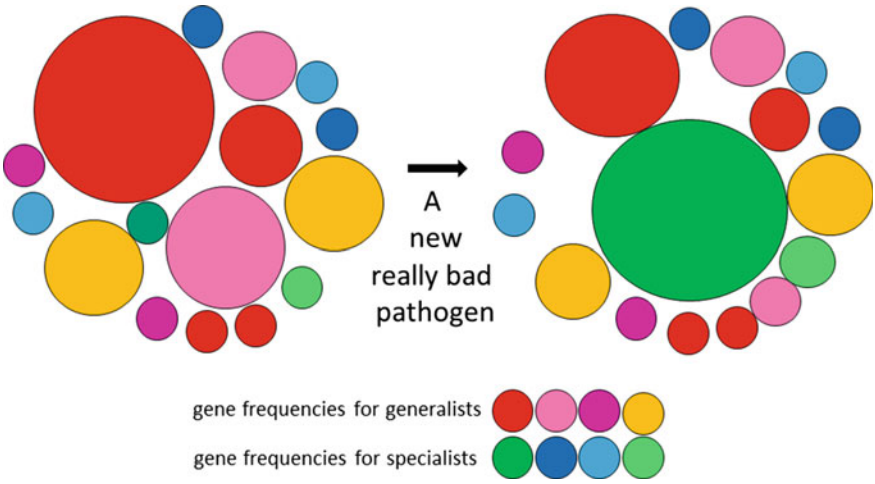


Fig. 6 Selection by a new and/or particularly virulent pathogen for which the generalists are unable to confer protection can change gene frequencies from predominantly a few generalist MHC molecules to a high level of a particular specialist molecule. The diameter of each circle indicates the frequency of a particular MHC gene in a population before and after selection by a pathogen, in this case a new and/or nasty pathogen. The warm colours indicate promiscuous molecules that act as generalists, conferring protection to most pathogens including those regularly found in the environment. The cooler colours indicate fastidious molecules encoded by genes that arise by mutation and are present at low frequency, but with the possibility of presenting a protective peptide from a particular pathogen, with the ensuing effective T cell response leading to survival of those individuals with that specialist MHC molecule, such that the particular specialist rises in gene frequency under strong selection

identical to HLA-B*57:01 and B*27:05 (and therefore almost certainly fastidious specialists which may confer long-term non-progression to AIDS) and those with peptide motifs much like the chicken class I molecule from the B2 haplotype (and therefore are likely to be promiscuous generalists that may confer resistance to many common pathogens).

Can we use the concept of generalists and specialists for prediction of resistance and susceptibility? Unfortunately, the situation is complex. For instance, there are many class I alleles associated with slow or with rapid progression from HIV infection to AIDS (Goulder and Walker 2012); only for HLA-B*57:01, B*27:05, B*07:02 and B*35:01 are surface expression levels on normal ex vivo cells known (Chappell et al. 2015). The cell surface expression of transfectants in the absence of tapasin and breadth of peptide binding is known for more alleles (Paul et al. 2013; Rizvi et al. 2014); the extremes are enriched in protective or non-protective alleles, but the middle of the hierarchies are a mixture of both (Fig. 4). We would certainly expect that many fastidious class I molecules would not find a protective peptide for a particular strain of HIV and therefore would not confer resistance to AIDS. Conversely, some generalists might find a protective peptide and thus confer some resistance to AIDS. Thus, the generalist–specialist paradigm may provide a useful framework for understanding resistance and susceptibility to pathogens, but is unlikely to be an absolute predictor of responses to particular pathogens.

On the other hand, this view of promiscuous generalists and fastidious specialists may help answer the uncomfortable questions about evolution of the class I molecules in the MHC. One problem was that in a co-evolving system as we envisage for the chicken MHC, a change in the dominantly expressed class I molecule might require a similar or compensatory change in the antigen processing and/or peptide loading molecules, thus slowing evolution down considerably. Another problem was the breaking of the co-evolutionary system in the lineage leading to placental mammals, once the first recombination event switched the class I allele. In both of these cases, it is easy to imagine a change in the binding specificity of the dominantly expressed class I molecule, as long as peptides transported by the promiscuous TAP would bind the new class I molecule; there would be ample time thereafter to refine the translocation specificity of the TAP in line with the new class I molecule. This scenario fits well with the idea that the promiscuous generalist haplotypes are ancestral, with the existence of promiscuous class I molecules in non-mammalian vertebrates bringing the binding strategy of class I and class II molecules closer together, in line with the proposal that they arose from a common ancestor.

8 The next steps?

So, what is left to do in trying to understand whether the idea of generalist and specialist MHC molecules is a good model to explain the wide variety of data described in this paper? These efforts might be divided into questions about the

molecular mechanisms, functional consequences at the cellular level and disease resistance at the population level.

One important molecular question is how tight the correlation between peptide repertoire and cell surface expression level might be, which requires a much more quantitative approach to both the diversity of peptides and the level of cell surface expression than has been utilized up to now. Another issue to resolve is the apparent mechanistic difference between what has been reported for humans and chickens. Is the hierarchy of tapasin dependence described for human class I molecules also true for chicken class I molecules, is it inversely correlated with peptide repertoire, and if so, which is causal? A most important question is to what extent these molecular findings are general to other vertebrates. Just as for chickens, there has been great interest in disease resistance in farmed animals such as cattle and salmon, and there has been much work on the mechanisms for class I molecules in mice, so information about the peptides found on MHC molecules is likely to be the first step in understanding the generality of these concepts.

Perhaps the most pressing question at the functional level is whether or not promiscuous class I molecules stimulate many T cell clones that are effective over the course of an immune response. In other words, is the prevailing concept of immunodominance described for (some) human and mouse alleles really true for all class I molecules or only for fastidious class I molecules? Under what situations is presenting one particularly protective peptide to a single effective T cell better than presenting many different peptides to a range of T cells? The second critical question is, whether or not the cell surface expression level is truly involved in creating some kind of optimal T cell repertoire by easing the intensity of negative selection. If it turns out to be difficult to confirm this hypothesis, then the true reason(s) for such differences need(s) to be determined.

Final (and perhaps the most important) step is to determine whether the concept of generalist and specialist MHC molecules can explain and predict the resistance and susceptibility to disease at the population level. There is an enormous scientific literature describing the response of humans and mice to infectious, inflammatory and autoimmune disease. A reassessment of this literature in the light of promiscuous generalists and fastidious specialists would require quantitative determination of peptide repertoire (and cell surface expression level) of many human and mouse class MHC alleles. Is the notion of fastidious and promiscuous MHC molecules best restricted to class I molecules or does it extend to class II molecules as well? Moreover, is it only able to explain responses to (certain) viral pathogens, or does it extend to MHC-determined responses to bacteria, fungi and eukaryotic parasites, and to autoimmune diseases, allergies and inflammatory diseases? To what extent can the survival of wild vertebrates, particularly in the context of endangered species, be explained (and even managed) by considering promiscuous versus fastidious MHC molecules? Finally, can we better understand the arms race between hosts and their pathogens using the generalist and specialist paradigm?

9 Conclusions

This chapter has come a long way from the initial paragraphs describing various levels of immunity and the importance of the recognition of self and non-self. The intended sketch of a new view of MHC classical class I molecules as promiscuous generalists and fastidious specialists turned out to be rather detailed, but the overall concepts may be of wider applicability. For instance, an evolutionary genetics investigation of gastrointestinal parasites in a wild mouse also invoked the concept of generalists and specialists (Froeschke and Sommer 2012), and it may be that class II molecules also have the properties of promiscuous generalists and fastidious specialists. Perhaps future analyses of the receptors for PAMPs and DAMPs, NK cell receptors and VLRs may uncover similar relationships for self and non-self recognition.

References

Apps R, Qi Y, Carlson JM, Chen H, Gao X, Thomas R, Yuki Y, Del Prete GQ, Goulder P, Brumme ZL, Brumme CJ, John M, Mallal S, Nelson G, Bosch R, Heckerman D, Stein JL, Soderberg KA, Moody MA, Denny TN, Zeng X, Fang J, Moffett A, Lifson JD, Goedert JJ, Buchbinder S, Kirk GD, Fellay J, McLaren P, Deeks SG, Pereyra F, Walker B, Michael NL, Weintrob A, Wolinsky S, Liao W, Carrington M (2013) Influence of HLA-C expression level on HIV control. Science 340(6128):87–91

Babik W, Kawalko A, Wójcik JM, Radwan J (2012) Low major histocompatibility complex class I (MHC I) variation in the European bison (Bison bonasus). J Hered. 103(3):349–359

Banat GR, Tkalcic S, Dzielawa JA, Jackwood MW, Saggese MD, Yates L, Kopulos R, Briles WE, Collisson EW (2013) Association of the chicken MHC B haplotypes with resistance to avian coronavirus. Dev Comp Immunol 39(4):430–437

Bangham CR (2009) CTL quality and the control of human retroviral infections. Eur J Immunol 39(7):1700–1712

Beck S, Trowsdale J (2000) The human major histocompatability complex: lessons from the DNA sequence. Annu Rev Genomics Hum Genet 1:117–137

Bernatchez L, Landry C (2003) MHC studies in nonmodel vertebrates: what have we learned about natural selection in 15 years? J Evol Biol 16(3):363–377

Blum JS, Wearsch PA, Cresswell P (2013) Pathways of antigen processing. Annu Rev Immunol 31:443–473

Boehm T, Iwanami N, Hess I (2012a) Evolution of the immune system in the lower vertebrates. Annu Rev Genomics Hum Genet 13:127–149

Boehm T, McCurley N, Sutoh Y, Schorpp M, Kasahara M, Cooper MD (2012b) VLR-based adaptive immunity. Annu Rev Immunol 30:203–220

Boonyanuwat K, Thummabutra S, Sookmanee N, Vatchavalkhu V, Siripholvat V (2006) Influences of major histocompatibility complex class I haplotypes on avian influenza virus disease traits in Thai indigenous chickens. Anim Sci J 77:285–289

Butter C, Staines K, van Hateren A, Davison TF, Kaufman J (2013) The peptide motif of the single dominantly expressed class I molecule of the chicken MHC can explain the response to a molecular defined vaccine of infectious bursal disease virus (IBDV). Immunogenetics 65 (8):609–618

Calnek BW (1992) Gordon memorial lecture. Chicken neoplasia–a model for cancer research. Br Poult Sci 33(1):3–16

Carrington M, Nelson GW, Martin MP, Kissner T, Vlahov D, Goedert JJ, Kaslow R, Buchbinder S, Hoots K, O'Brien SJ (1999) HLA and HIV-1: heterozygote advantage and B*35-Cw*04 disadvantage. Science 283(5408):1748–1752

Chappell P, el Meziane K, Harrison M, Magiera Ł, Hermann C, Mears L, Wrobel AG, Durant C, Nielsen LL, Buus S, Ternette N, Mwangi W, Butter C, Nair V, Ahyee T, Duggleby R, Madrigal A, Roversi P, Lea SM, Kaufman J (2015) Expression levels of MHC class I molecules are inversely correlated with promiscuity of peptide binding. Elife 10(4):e05345

Cotter PF, Taylor RL Jr, Abplanalp H (1992) Differential resistance to staphylococcus aureus challenge in major histocompatibility (B) complex congenic lines. Poult Sci 71(11):1873–1878

Dawkins R, Leelayuwat C, Gaudieri S, Tay G, Hui J, Cattley S, Martinez P, Kulski J (1999) Genomics of the major histocompatibility complex: haplotypes, duplication, retroviruses and disease. Immunol Rev 167:275–304

de Groot NG, Heijmans CM, Zoet YM, de Ru AH, Verreck FA, van Veelen PA, Drijfhout JW, Doxiadis GG, Remarque EJ, Doxiadis II, van Rood JJ, Koning F, Bontrop RE (2010) AIDS-protective HLA-B*27/B*57 and chimpanzee MHC class I molecules target analogous conserved areas of HIV-1/SIVcpz. Proc Natl Acad Sci USA 107(34):15175–15180

Dietert R, Taylor R, Dietert M (1990) The chicken major histocompatibility complex: structure and impact on immune function, disease resistance and productivity. In: Basta O (ed) MHC, differentiation antigens and cytokines in animals and birds. Bar-Lab Inc, Backsburg, pp 7–26

Ellegren H, Mikko S, Wallin K, Andersson L (1996) Limited polymorphism at major histocompatibility complex (MHC) loci in the Swedish moose A. alces. Mol Ecol 5(1):3–9

Froeschke G, Sommer S (2012) Insights into the complex associations between MHC class II DRB polymorphism and multiple gastrointestinal parasite infestations in the striped mouse. PLoS ONE 7(2):e31820

Goulder PJ, Walker BD (2012) HIV and HLA class I: an evolving relationship. Immunity 37 (3):426–440

Grimholt U, Larsen S, Nordmo R, Midtlyng P, Kjoeglum S, Storset A, Saebø S, Stet RJ (2003) MHC polymorphism and disease resistance in Atlantic salmon (Salmo salar); facing pathogens with single expressed major histocompatibility class I and class II loci. Immunogenetics 55 (4):210–219

Guo HC, Madden DR, Silver ML, Jardetzky TS, Gorga JC, Strominger JL, Wiley DC (1993) Comparison of the P2 specificity pocket in three human histocompatibility antigens: HLA-A*6801, HLA-A*0201, and HLA-B*2705. Proc Natl Acad Sci USA 90(17):8053–8057

Hill AV (1998) The immunogenetics of human infectious diseases. Annu Rev Immunol 16:593–617

Horton R, Wilming L, Rand V, Lovering RC, Bruford EA, Khodiyar VK, Lush MJ, Povey S, Talbot CC Jr, Wright MW, Wain HM, Trowsdale J, Ziegler A, Beck S (2004) Gene map of the extended human MHC. Nat Rev Genet 5(12):889–899

International HIV Controllers Study, Pereyra F et al (2010) The major genetic determinants of HIV-1 control affect HLA class I peptide presentation. Science 330(6010):1551–1557

Kaufman J (1999) Co-evolving genes in MHC haplotypes: the "rule" for nonmammalian vertebrates? Immunogenetics 50(3–4):228–236

Kaufman J (2011) The evolutionary origins of the adaptive immune system of jawed vertebrates (Chapter 3, pp 41-55). In: Kaufmann SHE, Rouse BT, Sachs DL (eds) The Immune Response to Infection. American Society of Microbiology Press, Washington DC

Kaufman J (2013) Antigen processing and presentation: evolution from a bird's eye view. Mol Immunol 55(2):159–161

Kaufman J (2014) The avian MHC. In: Schat KA, Kaiser P, Kaspers B (eds) Avian Immunology, edn 2. Elsevier, Ltd., Amsterdam, pp 149–167

Kaufman J (2015a) What chickens would tell you about the evolution of antigen processing and presentation. Curr Opin Immunol 34:35–42

Kaufman J (2015b) Co-evolution with chicken class I genes. Immunol Rev 267(1):56–71

Kaufman JF, Auffray C, Korman AJ, Shackelford DA, Strominger J (1984) The class II molecules of the human and murine major histocompatibility complex. Cell 36(1):1–13

Kaufman J, Völk H, Wallny HJ (1995) A "minimal essential Mhc" and an "unrecognized Mhc": two extremes in selection for polymorphism. Immunol Rev 143:63–88

Kaufman J, Milne S, Göbel TW, Walker BA, Jacob JP, Auffray C, Zoorob R, Beck S (1999) The chicken B locus is a minimal essential major histocompatibility complex. Nature 401 (6756):923–925

Koch M, Camp S, Collen T, Avila D, Salomonsen J, Wallny HJ, van Hateren A, Hunt L, Jacob JP, Johnston F, Marston DA, Shaw I, Dunbar PR, Cerundolo V, Jones EY, Kaufman J (2007) Structures of an MHC class I molecule from B21 chickens illustrate promiscuous peptide binding. Immunity 27(6):885–899

Kosmrlj A, Read EL, Qi Y, Allen TM, Altfeld M, Deeks SG, Pereyra F, Carrington M, Walker BD, Chakraborty AK (2010) Effects of thymic selection of the T-cell repertoire on HLA class I-associated control of HIV infection. Nature 465(7296):350–354

Kulkarni S, Savan R, Qi Y, Gao X, Yuki Y, Bass SE, Martin MP, Hunt P, Deeks SG, Telenti A, Pereyra F, Goldstein D, Wolinsky S, Walker B, Young HA, Carrington M (2011) Differential microRNA regulation of HLA-C expression and its association with HIV control. Nature 472 (7344):495–498

Lukacs MF, Harstad H, Grimholt U, Beetz-Sargent M, Cooper GA, Reid L, Bakke HG, Phillips RB, Miller KM, Davidson WS, Koop BF (2007) Genomic organization of duplicated major histocompatibility complex class I regions in Atlantic salmon (Salmo salar). BMC Genom 25(8):251

Martin MP, Qi Y, Gao X, Yamada E, Martin JN, Pereyra F, Colombo S, Brown EE, Shupert WL, Phair J, Goedert JJ, Buchbinder S, Kirk GD, Telenti A, Connors M, O'Brien SJ, Walker BD, Parham P, Deeks SG, McVicar DW, Carrington M (2007) Innate partnership of HLA-B and KIR3DL1 subtypes against HIV-1. Nat Genet 39(6):733–740

McBride RA, Cutting JA, Schierman LW, Strebel FR, Watanabe DH (1981) MHC gene control of growth of avian sarcoma virus-induced tumours in chickens: a study on the role of virus strains. J Immunogenet. 8(3):207–214

McLaren PJ, Coulonges C, Bartha I, Lenz TL, Deutsch AJ, Bashirova A, Buchbinder S, Carrington MN, Cossarizza A, Dalmau J, De Luca A, Goedert JJ, Gurdasani D, Haas DW, Herbeck JT, Johnson EO, Kirk GD, Lambotte O, Luo M, Mallal S, van Manen D, Martinez-Picado J, Meyer L, Miro JM, Mullins JI, Obel N, Poli G, Sandhu MS, Schuitemaker H, Shea PR, Theodorou I, Walker BD, Weintrob AC, Winkler CA, Wolinsky SM, Raychaudhuri S, Goldstein DB, Telenti A, de Bakker PI, Zagury JF, Fellay J (2015) Polymorphisms of large effect explain the majority of the host genetic contribution to variation of HIV-1 virus load. Proc Natl Acad Sci USA 112(47):14658–14663

Mesa CM, Thulien KJ, Moon DA, Veniamin SM, Magor KE (2004) The dominant MHC class I gene is adjacent to the polymorphic TAP2 gene in the duck, Anas platyrhynchos. Immunogenetics 56(3):192–203

Moon DA, Veniamin SM, Parks-Dely JA, Magor KE (2005) The MHC of the duck (Anas platyrhynchos) contains five differentially expressed class I genes. J Immunol 175(10): 6702–6712

Murphy K, Weaver C (2016) Janeway's immunobiology, 9th edn. Garland Press, New York

Neefjes JJ, Ploegh HL (1988) Allele and locus-specific differences in cell surface expression and the association of HLA class I heavy chain with beta2-microglobulin: differential effects of inhibition of glycosylation on class I subunit association. Eur J Immunol 18(5):801–810

Nowak MA, Tarczy-Hornoch K, Austyn JM (1992) The optimal number of major histocompatibility complex molecules in an individual. Proc Natl Acad Sci USA 89(22):10896–10899

Ohta Y, Goetz W, Hossain MZ, Nonaka M, Flajnik MF (2006) Ancestral organization of the MHC revealed in the amphibian Xenopus. J Immunol 176(6):3674–3685

Osterrieder N, Kamil JP, Schumacher D, Tischer BK, Trapp S (2006) Marek's disease virus: from miasma to model. Nat Rev Microbiol 4(4):283–294

Owen J, Punt J, Kuby Stranford S (2013) Immunology, 7th edn, WH Freeman and Company, New York

Paul S, Weiskopf D, Angelo MA, Sidney J, Peters B, Sette A (2013) HLA class I alleles are associated with peptide-binding repertoires of different size, affinity, and immunogenicity. J Immunol 191(12):5831–5839

Plachy J, Pink JR, Hála K (1992) Biology of the chicken MHC (B complex). Crit Rev Immunol 12 (1–2):47–79

Rizvi SM, Salam N, Geng J, Qi Y, Bream JH, Duggal P, Hussain SK, Martinson J, Wolinsky SM, Carrington M, Raghavan M (2014) Distinct assembly profiles of HLA-B molecules. J Immunol 192(11):4967–4976

Schierman LW, Collins WM (1987) Influence of the major histocompatibility complex on tumor regression and immunity in chickens. Poult Sci 66(5):812–818

Schou TW, Permin A, Juul-Madsen HR, Sørensen P, Labouriau R, Nguyên TL, Fink M, Pham SL (2007) Gastrointestinal helminths in indigenous and exotic chickens in Vietnam: association of the intensity of infection with the major histocompatibility complex. Parasitology 134(Pt 4):561–573

Sidney J, Asabe S, Peters B, Purton KA, Chung J, Pencille TJ, Purcell R, Walker CM, Chisari FV, Sette A (2006) Detailed characterization of the peptide binding specificity of five common Patr class I MHC molecules. Immunogenetics 58(7):559–570

Spurgin LG, Richardson DS (2010) How pathogens drive genetic diversity: MHC, mechanisms and misunderstandings. Proc Biol Sci 277(1684):979–988

Taylor RL Jr (2004) Major histocompatibility (B) complex control of responses against Rous sarcomas. Poult Sci 83(4):638–649

Thomas R, Apps R, Qi Y, Gao X, Male V (2009) O'hUigin C, O'Connor G, Ge D, Fellay J, Martin JN, Margolick J, Goedert JJ, Buchbinder S, Kirk GD, Martin MP, Telenti A, Deeks SG, Walker BD, Goldstein D, McVicar DW, Moffett A, Carrington M. HLA-C cell surface expression and control of HIV/AIDS correlate with a variant upstream of HLA-C. Nat Genet 41(12):1290–1294

Tiercy JM, Claas F (2013) Impact of HLA diversity on donor selection in organ and stem cell transplantation. Hum Hered 76(3–4):178–186

Tregaskes CA, Harrison M, Sowa AK, van Hateren A, Hunt LG, Vainio O, Kaufman J (2016) Surface expression, peptide repertoire, and thermostability of chicken class I molecules correlate with peptide transporter specificity. Proc Natl Acad Sci USA 113(3):692–697

Trowsdale J, Knight JC (2013) Major histocompatibility complex genomics and human disease. Annu Rev Genomics Hum Genet 14:301–323

van Hateren A, Carter R, Bailey A, Kontouli N, Williams AP, Kaufman J, Elliott T (2013) A mechanistic basis for the co-evolution of chicken tapasin and major histocompatibility complex class I (MHC I) proteins. J Biol Chem 288(45):32797–32808

Vidović D, Matzinger P (1988) Unresponsiveness to a foreign antigen can be caused by self-tolerance. Nature 336(6196):222–225

von Boehmer H, Teh HS, Kisielow P (1989) The thymus selects the useful, neglects the useless and destroys the harmful. Immunol Today 10(2):57–61

Walker BA, Hunt LG, Sowa AK, Skjødt K, Göbel TW, Lehner PJ, Kaufman J (2011) The dominantly expressed class I molecule of the chicken MHC is explained by coevolution with the polymorphic peptide transporter (TAP) genes. Proc Natl Acad Sci USA 108(20):8396–8401

Wallny HJ, Avila D, Hunt LG, Powell TJ, Riegert P, Salomonsen J, Skjødt K, Vainio O, Vilbois F, Wiles MV, Kaufman J (2006) Peptide motifs of the single dominantly expressed class I molecule explain the striking MHC-determined response to Rous sarcoma virus in chickens. Proc Natl Acad Sci USA 103(5):1434–1439

Witter RL (1998) Control strategies for Marek's disease: a perspective for the future. Poult Sci 77(8):1197–1203

Yewdell JW, Bennink JR (1999) Immunodominance in major histocompatibility complex class I-restricted T lymphocyte responses. Annu Rev Immunol 17:51–88

Zhang P, Kuang YY, Wu HL, Li L, Ge YF, Wan QH, Fang SG (2012) The Père David's deer MHC class I genes show unexpected diversity patterns, with monomorphic classical genes but polymorphic nonclassical genes and pseudogenes. J Exp Zool B Mol Dev Evol 318(4):294–307

Evolution and Diversity of Defensins in Vertebrates

Edward J. Hollox and Razan Abujaber

Abstract Defensins are a large family of genes that were first characterised as encoding antimicrobial peptides, with a broad range of activity against viruses, bacteria and fungi. It is clear, however, that at least in vertebrates, they have acquired a variety of other roles in addition to direct antimicrobial activity, including cell signalling, reproduction and mammalian coat colour. In this article, we review the evolutionary history of the three types of defensins found in vertebrates, namely α-, β- and θ-defensins. We consider evolution at a deep timescale, where a pattern of duplication and divergence emerges, consistent with birth-and-death evolution. At a more recent timescale, we consider the evolutionary genetics of defensins within species, particularly copy number variation which is observed for many defensins across several lineages. The different functions of at least some defensins in different evolutionary lineages raise some problems in inferring function based on identification of a homologous gene in a different species. However, defensins are also an excellent model for studying the evolution of new functions following duplication and divergence of genes.

1 The Big Picture of Defensin Evolution

Defensins are a family of genes that encode small proteins defined by a shared six-cysteine motif. These six cysteines form a distinct arrangement of three disulphide bridges in the mature tertiary structure and differ from α- and θ-defensins by the arrangement of these disulphide bridges (Fig. 1), which forms the basis for classification of defensins into α, β and θ. In β-defensins, Cys1 pairs with Cys5, Cys3 links to Cys6 and Cys2 links to Cys4, in contrast to α-defensins where Cys1 links to Cys6 and Cys3 links to Cys5. They have been characterised in a wide variety of vertebrates and were given the name defensins because of their antimi-

E.J. Hollox (✉) · R. Abujaber
Department of Genetics and Genome Biology, University of Leicester,
Adrian Building, University Road, Leicester LE1 7RH, UK
e-mail: Ejh33@le.ac.uk

P. Pontarotti (ed.), *Evolutionary Biology: Self/Nonself Evolution, Species and Complex Traits Evolution, Methods and Concepts*,
DOI 10.1007/978-3-319-61569-1_2

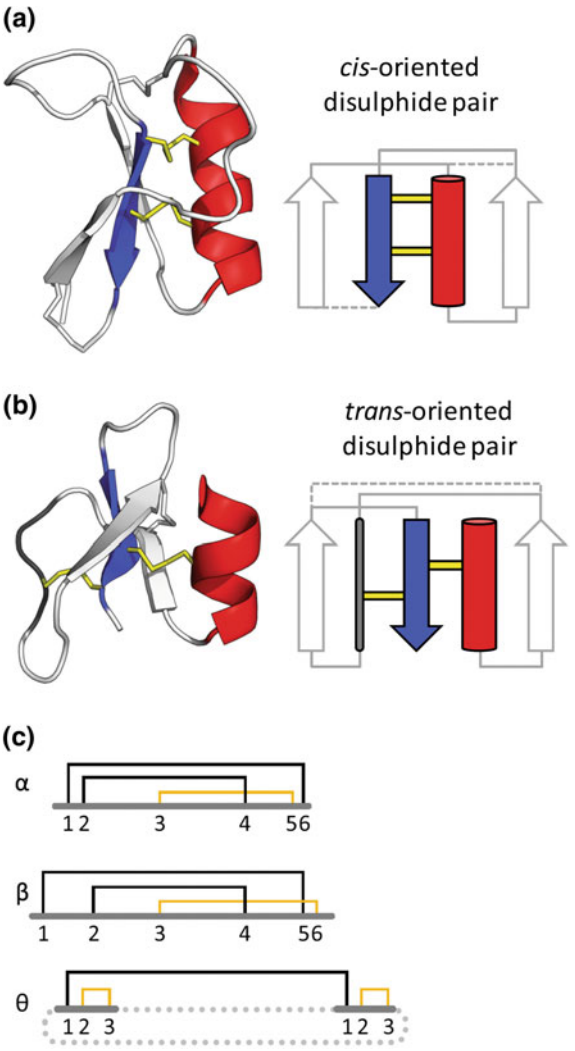


Fig. 1 Orientation of disulphide bonds in defensins. **a** and **b** Comparison of disulphide pair orientations, forming the highest-level differentiation between defensins. **c** Arrangement of disulphide bonds between cysteine residues in vertebrate α-, β- and θ-defensins. Figure reproduced, with modification, from Shafee et al. (2017) with permission

crobial activity. Indeed, defensins are an important part of the innate immune response, as they form part of the mucosal barrier against microbes. At the amino acid sequence level, different human β-defensins are very distinct (Fig. 2). The six-cysteine motif that defines a β-defensin is highly conserved, with only a glycine and aspartic acid within the β-defensin core region also showing extensive conservation. This is also the general case for α-defensins, although not for the more recently evolved θ-defensins, as only one member of this family exists. This amino acid diversity suggests that different β-defensins may have very diverse functions both within and outside the innate immune response.

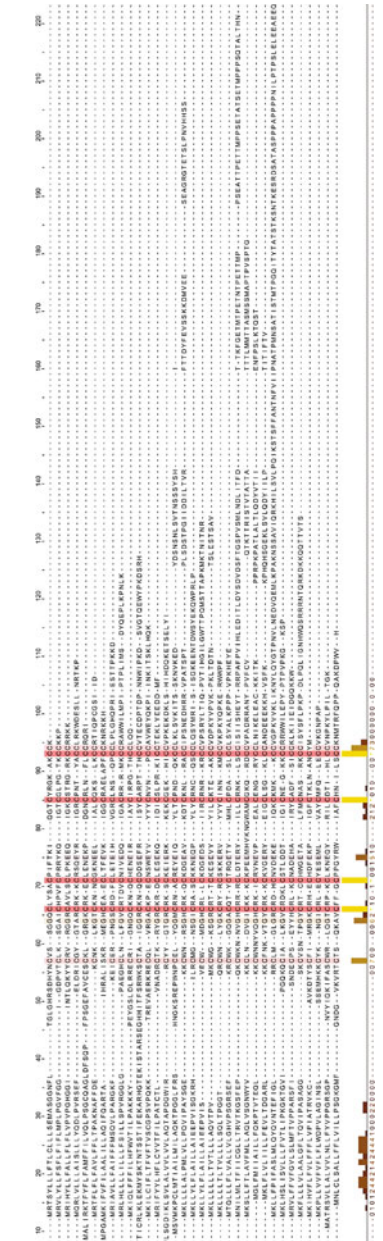


Fig. 2 Alignment of human β-defensin amino acid sequences. Known and predicted human β-defensin amino acid sequences aligned using Clustal Omega (Sievers et al. 2011) and plotted using JalView (Waterhouse et al. 2009). Signal sequence regions and core β-defensin regions are indicated. Conservation is indicated by the track at the bottom of the figure and by shading of amino acids, clearly showing the conserved six cysteines that define β-defensins. Note the extended C-terminal regions of several β-defensins, which contain potential glycosylation sites

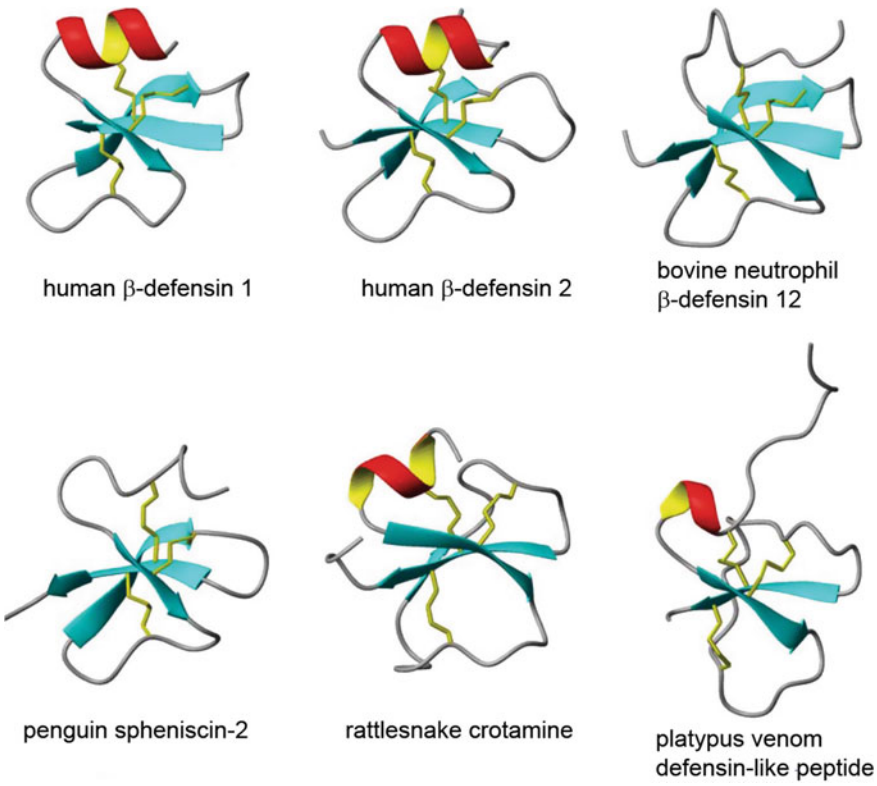


Fig. 3 Examples of vertebrate β-defensin structures. A variety of β-defensin protein structures, highlighting the β-defensin fold. The disulphide bonds are shown in *yellow*, β-strands in *cyan* and α-helices in *red/yellow*. Figure reproduced, with modification, from Torres and Kuchel (2004), with permission

The extensive variation at the amino acid level, combined with the small size of most defensins, limits our understanding of deep evolutionary relationships between β-defensins and other members of the defensin family. However, comparison of protein structures of defensins showed that all defensins, previously classified by cysteine-bridging patterns, can in fact be divided into just two main groups (cis- and trans-) based on their arrangement of the disulphide bridges in the three-dimensional protein structure (Fig. 1). These two groups have distinct evolutionary origins yet share a six-cysteine motif because of convergent evolution (Shafee et al. 2016). Vertebrate β-defensins are a type of trans-defensin that share a distinctive protein fold called the β-defensin fold (Fig. 3). Because α- and θ-defensins have arisen and diverged from β-defensins, they are also trans-defensins. Cis-defensins are present broadly across eukaryotes, but, because no cis-defensins have yet been identified in vertebrates, the origin and evolution of β-defensins in vertebrates may be a result of this loss of cis-defensins (Shafee et al. 2016).

In this review, we focus on defensins in vertebrates, aware that this is only part of the field of defensin evolution. However most is known about the defensins in vertebrates, particularly for the largest defensin family, the β-defensins, which have also been the focus of our research. Because of the size of the β-defensin family (27 members in humans, compared to four α-defensin genes and no θ-defensin genes), and because β-defensins are ancestral to α and θ, we also focus more on β-defensins than others. It is also the case that a single review could not encompass the whole field, and there are other excellent reviews elsewhere about other aspects of defensin biology, which we cite in this review.

2 Function of Defensins

Defensins were first isolated and characterised as small antimicrobial peptides expressed in neutrophils and at mucosal surfaces (Ganz et al. 1985; Diamond et al. 1991; Eisenhauer et al. 1990). In humans, α-defensins are expressed in the Paneth cells of the intestine and neutrophils, while both α- and β-defensins are expressed on a variety of mucosal surfaces. Mice lack α-defensins in neutrophils but express α-defensins (also known as cryptdins) in Paneth cells and, together with β-defensins, at mucosal surfaces. α-defensins play a key role in innate immune defence. This key role is emphasised by the fact that α-defensins 1–3 (encoded by *DEFA1A3*) comprise as much as 30–50% of human neutrophil granules (Rice et al. 1987), and α-defensin 5 (encoded by *DEFA5*) is active against *Salmonella typhimurium* in vivo (Salzman et al. 2003; Bevins 2013). Both α- and β-defensins have been shown to have broad antimicrobial spectrum activity against bacteria, fungi and viruses (Feng et al. 2005; Aerts et al. 2008; Chu et al. 2012; Raschig et al. 2017; Wilson et al. 2016; Wiens et al. 2014; Lehrer and Lu 2012; Taylor et al. 2008).

It soon was established that both α- and β-defensins had roles in immune signalling and at a concentration lower than that required for their antimicrobial effects (Lehrer and Lu 2012; Semple and Dorin 2012). For example, α-defensins 1–3 chemoattract naive CD4+ T cells and immature dendritic cells to the site of inflammation (Yang et al. 2000). Another example is human β-defensin 2, which interacts with the CCR6 and CCR2 receptors and chemoattracts CD4+ memory T cells and dendritic cells (Rohrl et al. 2010; Yang et al. 1999). It is clear that although defensins mediate these effects via receptors, they may in fact be promiscuous ligands that interact electrostatically with a wide variety of receptors involved in the immune response (Suarez-Carmona et al. 2015; Semple and Dorin 2012). In this way, the interactions of defensins with the immune system may have evolved early in vertebrate evolution as an effective way of co-opting an innate antimicrobial response to become a signal to the adaptive immune system.

Despite the well-established role of β-defensins at the mucosal surface, it is striking that most β-defensins are in fact expressed in the epididymis of the testis, and for most of these, their precise function is unknown (Fig. 4) (Zhou et al. 2004;

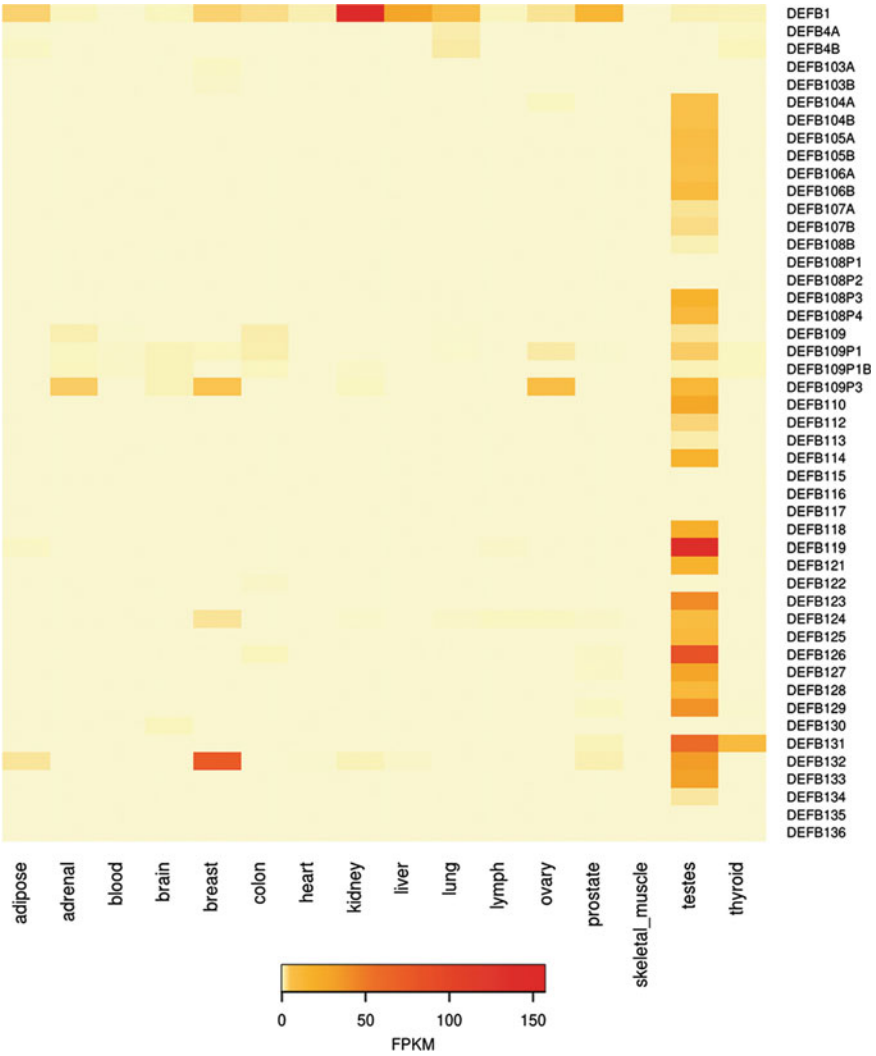


Fig. 4 β-defensin expression in humans across 16 tissues. Heatmap showing relative expression levels from RNASeq data generated by the Illumina BodyMap 2.0 project. Legend shows heat colour related to Fragments Per Kilobase of transcript per Million mapped reads (FPKM). Note the absence of skin tissue, and the predominance of testes expression of most human β-defensins

Dorin and Barratt 2014). All are annotated by databases as having direct antimicrobial activity, although for most this has not been directly demonstrated and is only an assumption from the fact that they share a predicted six-cysteine motif and are therefore identified by similarity to existing members of the β-defensin family.

In humans, β-defensin proteins are commonly referred to as hbds. The β-defensins hbd-1 (encoded by the gene *DEFB1*) and hbd-2 (encoded by the gene

DEFB4) were isolated first, and most research effort has focused on these. Other genes predicted to encode β-defensins were identified by genome sequence search strategies (Schutte et al. 2002). A total of 33 β-defensin genes that are transcribed and predicted to generate proteins have been identified in humans. They are comprised of a signal sequence (except DEFB112 which lacks a signal consensus cleavage site), a core β-defensin region, and some have an extended C-terminal sequence. Some β-defensins have been shown to undergo further proteolytic processing after signal sequence cleavage.

Mice carrying deletions of one or several β-defensin genes are now illuminating the function of these proteins beyond their direct antimicrobial activity. A key finding is that a knockout of nine β-defensins renders male mice infertile, supporting a key role of β-defensins in reproduction, previously suggested by the fact that many β-defensins are expressed solely in the epididymis (Dorin 2015; Zhou et al. 2013). The importance of β-defensins in fertility is underlined by work on DEFB126 in humans and rhesus macaques. This has shown that DEFB126 protein is highly glycosylated and is adsorbed on to the surface of sperm during movement through the epididymis (Tollner et al. 2008a; Yudin et al. 2005). DEFB126 may facilitate penetration of negative cervical mucus and protect the sperm against immune recognition in the female during transit (Tollner et al. 2008b). DEFB126 is subsequently shed in the oviduct allowing normal fertilisation to occur (Tollner et al. 2011, 2012). An important role for DEFB126 in sperm motility has also been shown in cattle (Fernandez-Fuertes et al. 2016).

Only one θ-defensin (also known as retrocyclin) is known, encoded by the *DEFT1* gene, identified in rhesus macaques but a non-functioning pseudogene in humans (Nguyen et al. 2003). The structure is very different from other defensins and involves head-to-tail ligation of two nine-amino acid peptides to form a circular molecule (Tang et al. 1999; Lehrer et al. 2012). The antimicrobial effects of this molecule are well characterised (Gallo et al. 2006; Wang et al. 2006; Beringer et al. 2016), but it is not known whether it has any other function, such as acting as a chemokine.

3 Rapid Evolution of β-Defensins

The human β-defensins are rather distinct from each other at the amino acid level, and clear orthologues of human β-defensins can be identified in primate genomes, arguing against very recent (i.e. within the primates) rapid duplication and divergence across the whole family but suggesting older origins for most of the β-defensins seen in humans today (Fig. 2). Most human β-defensins have clear orthologues not just in primates but in other mammals as well.

In primates, the strongest evidence for selection is on the β-defensins that are involved in reproduction. An early study showed some evidence in the vervet monkey (*Cercopithicus aethiops*) for positive selection of *DEFB107* and *DEFB108*, both genes expressed in the epididymis (Semple et al. 2003). A later

study identified four other β-defensin genes that are expressed in the epididymis showing evidence of positive selection by likelihood-based substitution rate analysis in catarrhine primates (*DEFB118*, *DEFB120*, *DEFB127*, *DEFB132*) (Hollox and Armour 2008). However, one antimicrobial β-defensin, *DEFB1*, showed evidence of positive selection in this study, and population genetic analysis in humans suggested that balancing selection is operating on this gene (Cagliani et al. 2008). This has also been suggested for *DEFB127*, raising the possibility of ongoing episodes of balancing selection and positive selection, depending on the selective environment (Hollox and Armour 2008).

In humans and other primates, the β-defensins on chromosome region 8p23.1 (with the exception of *DEFB1*) are in a complex repeated region that is polymorphically duplicated and show extensive copy number variation (see section below). This can potentially limit comparative analyses as the region tends to be poorly represented in genome assemblies and recombination between paralogues can affect potential signals of positive selection. This complex region includes some β-defensin genes that have expanded in copy number in the orangutan lineage only (*DEFB130*, *DEFB134*, *DEFB135*, *DEFB136*) (Mohajeri et al. 2016).

Some particular β-defensins have undergone repeated rounds of duplication and divergence, particularly in rodents (Morrison et al. 2003). Analysis of rodent genomes showed a number of genes that were very similar to each other, suggesting recent duplication and divergence. For example, the mouse *Defb4* gene has repeatedly duplicated to generate five paralogues (*Defb3*, *Defb5*, *Defb6*, *Defb7* and *Defb8*) clustered together in the genome. The rodent-specific clades show evidence of positive selection using likelihood-based models identifying increased non-synonymous substitution rates at particular amino acid residues. The selected residues occur throughout the protein, and also, surprisingly, within the preproprotein region which is usually cleaved intracellularly before export of the mature β-defensin from the cell (Morrison et al. 2003; Maxwell et al. 2003).

This rapid duplication and divergence of defensins in rodents initially led to some uncertainty in identifying the true orthologue of some human genes and because of this uncertainty, most defensin genes in humans and mice were named independently and orthologous relationships established afterwards. Mouse defensins are named *Defb*x, where x is a number that usually reflects the order of discovery in mice, and most human β-defensin genes are named *DEFB*x where x either reflects the order of discovery in humans or is a number starting at 103. The analysis of complete genomes of mouse and humans has established orthologous pairs by using synteny as well as sequence similarity (Table 1; Patil et al. 2005).

A large study of avian defensins from 53 species of birds showed particular amino acid residues under positive selection. The degree of positive selection varies across the different β-defensin genes, and the position of the selected residues is difficult to interpret, being spread across the mature peptide and preproprotein, although there was a suggestion that residues flanking the conserved cysteine residues were more likely to be subject to positive selection (Cheng et al. 2015).

Population genetic analysis of a single species can give evolutionary insights of a more recent timescale compared to comparative analysis across different species.

Table 1 Known mouse orthologues of human β-defensin genes

Human chromosomal region	Human gene	Known Mouse orthologue(s)	Mouse chromosomal region
8p23.1	*DEFB1*	*Defb1*	8qA1.3-A2
8p23.1	*DEFB4*	*Defb4 family*[a]	8qA1.3-A2
8p23.1	*DEFB103*	*Defb14*	8qA1.3-A2
8p23.1	*DEFB105*	*Defb12/Defb35*	8qA1.3-A2
8p23.1	*DEFB106*	*Defb15/Defb34*	8qA1.3-A2
8p23.1	*DEFB107*	*Defb13*	8qA1.3-A2
8p23.1	*DEFB109*	*Defb42*	14qC3
6p12.3	*DEFB110*	*Defb16*	1qA3
6p12.3	*DEFB112*	*Defb17*	1qA3
6p12.3	*DEFB113*	*Defb18*	1qA3
20q11.21	*DEFB115*	*Defb28*	2qH1
20q11.21	*DEFB116*	*Defb29*	2qH1
20q11.21	*DEFB117*	*Defb19*	2qH1
20q11.21	*DEFB118*	*Defb21*	2qH1
20q11.21	*DEFB119*	*Defb24*	2qH1
20q11.21	*DEFB122*	*Defb27*	2qH1
20q11.21	*DEFB123*	*Defb36*	2qH1
20q11.21	*DEFB124*	*Defb25*	2qH1
20p13	*DEFB125*	*Defb26*	2qH1
20p13	*DEFB126*	*Defb22*	2qH1
20p13	*DEFB128*	*Defb20*	2qH1
20p13	*DEFB129*	*Defb23*	2qH1
8p23.1	*DEFB130*	*Defb41*	14qC3
8p23.1[b]	*DEFB131*	*Defb43*	14qC3
6p12.3	*DEFB133*	*Defb49*	1qA3
8p23.1	*DEFB135*	*Defb30*	14qC3
8p23.1	*DEFB136*	*Defb44*	14qC3

Based on Zhou et al. (2013) and Patil et al. (2005)
[a]*Defb4, Defb3*, *Defb5*, *Defb6*, *Defb7* and *Defb8*, see text
[b]Annotated only on a duplication on chr4

A study of wild mallards (*Anas platyrhynchos*) showed strong evidence for negative selection, with some evidence of balancing selection at certain genes. This emphasises the fact that by looking at different timescales of evolution, different patterns emerge—because of the changing environment, a gene that was subject to positive selection in the past may not be subject to positive selection now and vice versa (Chapman et al. 2016). In contrast, population genetic analysis of the domestic dog *DEFB103* variant encoding the coat colour allele dominant black (Candille et al. 2007) indicates recent positive selection where it has been introduced into wild wolves by hybridisation (Anderson et al. 2009). The melanism

variant has risen to high frequency in forested areas, where it has a camouflage advantage for the predator in pursuit of prey. Alternatively, this polymorphism may be maintained by negative assortative mating (Hedrick et al. 2016).

Analysis of the platypus (*Ornithoryhnchus anatinus*) genome has identified a β-defensin family (Ornithorhynchus venom defensin-like peptides, OvDLPs) that has been subject to rapid duplication and divergence (Whittington et al. 2008a, b). This duplication and divergence process started ~190 million years ago, probably from a common ancestor with mouse *Defb33*. OvDLPs have a role in the venom of the male platypus, which is produced by a hollow spur on the hind leg of males and is thought to be involved in asserting dominance over other males in the breeding season. Other venomous non-mammalian vertebrates have β-defensin-derived peptides in their venom. For example, crotamines and venom crotamine-like peptides (vCLPs) have arisen from β-defensins (Yount et al. 2009). Snake venom crotamines have arisen by duplication and divergence from an ancestor of mouse *Defb51* (Whittington et al. 2008a). This evidence shows that vCLPs have arisen independently from the platypus OvDLPs, showing evidence of convergent evolution of function.

The example of defensin-like peptides in venom illustrates a couple of important points in defensin evolution. Firstly, rapid sequence changes are a signature of adaptive evolution, and the adaptive evolution results in a change of function. For defensins, the change of function was often interpreted to reflect a change in microbial specificity, reflecting a host–pathogen co-evolutionary arms race. However, it is clear that defensins can evolve to have different functions and may often have two physiological roles at the same time. Therefore, bursts of adaptive evolution may reflect dramatic changes in function, and that a β-defensin in one organism may not necessarily be performing the same role as a defensin in another organism (i.e. be homologous) even if the gene is orthologous. Secondly, mouse *Defb33* shares the most recent common ancestor with OvDLPs, and mouse *Defb51* shares the most recent common ancestor with vCLPs, but neither *Defb33* nor *Defb51* have an orthologue in humans. This shows that β-defensins are lost by pseudogenisation or deletion in lineages, as well as gained by duplication and divergence, in a process known as birth-and-death evolution (Nei and Rooney 2005). However, the full extent of this is unknown, as absence of particular β-defensins from non-humans or non-mouse genomes may be due to incomplete genome assembly of complex repeated regions rich in defensin genes, rather than a true loss of a gene in a lineage.

4 Rapid Evolution of α- and θ-Defensins

α-defensins are unique to mammals, as no examples have yet been found in non-mammalian vertebrates, and have rapidly duplicated and diverged in different mammalian lineages leading to different α-defensin repertoires in different mammalian clades. There is evidence of gene loss—for example, in mice, in contrast to

rats, there appear to be no neutrophil α-defensins (Eisenhauer and Lehrer 1992). In humans, there are six functional α-defensins and six α-defensin pseudogenes. The functional α-defensins are enteric (*DEFA5*, *DEFA6*) or neutrophil-specific (*DEFA4* and *DEFA1A3* encoding α defensins 1–4). *DEFA1A3* shows extensive polymorphic copy number variation (CNV) as it is a coding gene entirely within a tandem repeat with a 19 kb repeat size with diploid copy numbers ranging from 4 to 10. Next to the 19 kb tandem repeat is a partial repeat which also carries a copy of the DEFA1A3 gene (Aldred et al. 2005; Khan et al. 2013). Different copies of the repeat encode either *DEFA1* or *DEFA3*, which differ only by a single nucleotide base and encoded amino acid. DEFA2 is thought to derive from the *DEFA1* gene by proteolytic processing of the peptide removing an extra N-terminal amino acid. The human pseudogenes are named *DEFA7P-DEFA11P* (Li et al. 2014).

There is a similar ratio of genes to pseudogenes across other catarrhine primates, but in the marmoset, there appears to be fewer pseudogenes, although this could be an artefact of poor genome assembly. A comparative analysis of α-defensin sequences strongly suggests extensive positive selection throughout the mature peptide (Lynn et al. 2004; Patil et al. 2004; Das et al. 2010), and the high number of pseudogenes suggests rapid birth-and-death evolution. Expression patterns of α-defensins can also evolve, as rabbits appear to have two kidney-specific α-defensins in a clade. θ-defensin, encoded by the *DEFT1* gene, is related to α-defensins (Tang et al. 1999). It is catarrhine-primate specific, having been initially identified in *Macaca mulatta* (rhesus macaque), but the *DEFT1* gene has become a pseudogene in the hominid lineage, including humans (Nguyen et al. 2003).

In summary, α-defensins evolved from one, perhaps two unidentified ancestral β-defensins in the mammalian lineage, an expansion that appears to have been triggered by an alteration in the disulphide bridge formation pattern and a consequent change in structure (Patil et al. 2004). Subsequently, in catarrhine primates, an α-defensin was truncated and became *DEFT1*, which encodes a small peptide which is self-ligated into a circular structure forming a θ-defensin called retrocyclin. In hominids, *DEFT1* acquired an inactivating mutation becoming the pseudogene *DEFT1P* (Nguyen et al. 2003; Li et al. 2014; Cheng et al. 2014), illustrating the process of birth-and-death evolution across a ∼25-million-year time span from the divergence of platyrrhine and catarrhine primates to the divergence of human and gorilla lineages.

5 Copy Number Variation of α-Defensins

In humans, *DEFA1A3* and *DEFT1P* are on a 19 kb tandem repeat that is copy number variable, as described in the previous section. This CNV is shared with chimpanzees, bonobos and orangutans, but not with gorillas (Sudmant et al. 2013). It is unclear whether this pattern is due to loss of CNV in the gorilla lineage or independent evolution of CNV in the human–chimpanzee ancestor and in the orangutan lineage.

There is evidence of non-allelic homologous recombination events causing copy number changes at the *DEFA1A3* locus, but the high linkage disequilibrium of SNP alleles flanking the CNV suggests that alternative mechanisms, like gene conversion, account for the majority of copy number mutation events. Gene conversion events homogenise sequence repeats, which will prevent sequence divergence of different copies of the *DEFA1A3* gene. Indeed, the sequence variant which specifies the DEFA3 protein (in contrast to the DEFA1 protein) can exist at either the distal or proximal end of the repeat, suggesting extensive shuffling of sequence between the repeat units by gene conversion (Black et al. 2014).

6 Copy Number Variation of β-Defensins

A notable feature of β-defensin gene clusters is that they often show extensive genome structural variation, particularly CNV, within a species. Analyses of CNV have identified variable regions containing β-defensins in humans (Conrad et al. 2009; Sudmant et al. 2015), cattle (Liu et al. 2010; Bickhart et al. 2012), dogs (Leonard et al. 2012), pigs (Wang et al. 2013), rhesus macaque (Lee et al. 2008; Gokcumen et al. 2011) and chickens (Lee et al. 2016).

The human CNV is the most studied of all the CNVs involving β-defensins. It involves a repeat unit of 322 kb in length (called DEFB), with six β-defensin genes (*DEFB4*, *DEFB103*, *DEFB104*, *DEFB105*, *DEFB106* and *DEFB107*) and *SPAG11*, a β-defensin-related gene (Ottolini et al. 2014; Forni et al. 2015). In the latest genome assembly, two copies of DEFB are embedded within a complex repeated region called REPD at chromosomal region 8p23.1. However, genetic mapping has shown that DEFB can also be present, polymorphically, at a related complex repeat region called REPP, ~4 Mb proximal to REPD (Abu Bakar et al. 2009; Mohajeri et al. 2016). Total diploid copy number can range from 1 copy per diploid genome to 12, with copy number between 2 and 7 frequent in the population, and a diploid copy number of 4 being modal. High copy numbers due to tandemly arranged DEFB repeats on one homologous chromosome are visible directly using G-band staining of metaphase chromosomes, are called 8p23.1 euchromatic variants, and can be mistaken for pathological duplications of the entire region between REPP and REPD (Hollox et al. 2003; Barber et al. 2005). Copy number variation of DEFB is not pathological, but increased copy number of DEFB is associated with an increased risk of the inflammatory skin disease psoriasis (Hollox et al. 2008; Stuart et al. 2012).

Because of the unusual arrangement of DEFB repeats on chromosome 8, allelic recombination anywhere between REPP and REPD can potentially change the copy number of a particular haplotype. For example, if a meiotic crossover happened between a 1–1 chromosome (1 copy at REPD and 1 copy at REPP) and a 2–0 chromosome, then the resulting gametes would be 2–1 and 1–0. Measuring the copy number changes in human pedigrees established the copy number mutation rate to be around 0.7% per gamete per generation, which is between 5 and 6 orders of

magnitude faster than single nucleotide substitution rates (Abu Bakar et al. 2009), and is comparable with mutation rates at tandemly repeated minisatellite loci.

DEFB is also variable in copy number in chimpanzees (*Pan troglodytes*) and bonobos (*Pan paniscus*) but not in gorilla or orangutan, suggesting that this CNV arose 7–10 million years ago after the divergence of the human lineage with gorillas, but prior to the divergence of humans and chimpanzees (Sudmant et al. 2013; Pala 2012). In rhesus macaques, the genes present on the DEFB repeat in humans are all single copy and do not show CNV, with the exception of the *DEFB4* gene (termed *DEFB2L* in macaques). This gene is on a tandemly repeated 20 kb repeat unit in rhesus macaques which varies between 3 and 6 copies per diploid genome (Fig. 5). The duplication that has been maintained as a CNV arose at least 3MYa and shows a signature of positive selection following that duplication event when the substitution pattern between *DEFB2L* copies is analysed by a McDonald–Kreitman test (Ottolini et al. 2014).

7 Is Copy Number Variation Adaptive?

Mutation rate clearly evolves to a particular value, as evolved genomes must have had a mutation rate fast enough to generate that particular evolved genome yet slow enough to prevent a fatal accumulation of deleterious mutations. It is also possible that certain loci (sometimes called "contingency loci") may have higher mutation rates—be more "evolvable"—because genomes carrying these high mutation rate loci are more likely to be carrying a beneficial variant (Sniegowski et al. 2000). This would be an example of second-order selection, where the rapidly mutating CNV does not affect the fitness of its carriers but affects the fitness of its descendants (Yona et al. 2015). Such a high CNV mutation rate locus might be more likely to evolve if any deleterious effects of CNV mutation at that locus are low—a low-risk high-gain strategy.

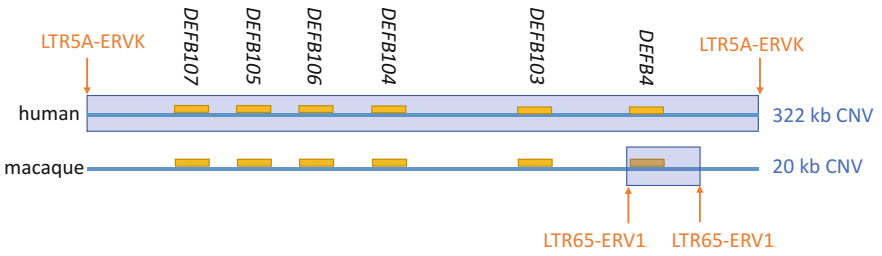


Fig. 5 Comparison between macaque and human β-defensin CNV. A cartoon showing the relative extents of the copy number variable region (*shaded in blue*) in humans and in the rhesus macaque. Genes are shown as *yellow boxes*, and the retroviral repeat elements at the boundaries of the CNV regions are also *highlighted*

Most modern population genetic methods to test for selection cannot easily be applied to complex CNV regions, like the β-defensin region in humans. Simple population genetic models, such as those using the stepwise mutation model, often do not fully integrate sequence variation between copies into the evolutionary model. Some attempts have been made to model CNV using coalescent approaches (Teshima and Innan 2012; Thornton 2007), but they often require an oversimplification of reality, particularly for multiallelic variants. Forward-in-time population genetic simulations are more flexible and may provide a more appropriate framework for examining the population genetics of complex CNV, but are computationally intensive. Almost all approaches require that the copy number and sequence variation of an individual are phased into individual haplotypes, which are still technically challenging. At present, this can be done most reliably by observing segregation of individual alleles in a pedigree (Palta et al. 2015) or by PCR methods designed to phase individual variants across tens of kb (Tyson and Armour 2017). When application of long-read sequencing technology, such as that provided by Pacific Biosystems or Oxford Nanopore, becomes routine for vertebrate genomes, phasing of complex CNV should become more straightforward (Buermans et al. 2017).

Nevertheless, CNV-aware comparative approaches across species and population genetic approaches within species can allow us to infer some aspects of the evolution of β-defensin CNV. The observation that β-defensin CNV has originated independently in both the macaque lineage and the human lineage is evidence of convergent evolution at the molecular level (Fig. 6). This argues that CNV itself has been favoured, at least for *DEFB4*, the gene that is copy number variable in humans and macaques. In humans, other defensin genes are on the CNV block, and this could either be adaptive or they could be bystanders in the CNV, with neutral or mildly deleterious consequences at high copy number, for example.

Could there be deleterious effects of high copy number at the β-defensin locus? In humans, the CNV repeat units can sponsor rare 3.6 Mb deletions in 8p23.1 which cause developmental delay (Mohajeri et al. 2016). We might predict that since larger regions of sequence identity are more prone to pathogenic NAHR mutations, there would be a positive relationship between β-defensin copy number and likelihood of a de novo pathogenic deletion involving these repeats, but at present, there is no evidence to support this. Furthermore, individuals carrying chromosomes with high β-defensin copy numbers (10–11 on one chromosome, diploid copy number of 12 or 13) are 8p23.1 euchromatic variant carriers and show no clinical pathology (Barber et al. 1998; Hollox et al. 2003). It is possible that an upper limit is placed on the DEFB copy number because high copy number chromosomes are more susceptible to genomic rearrangements, but this has not been shown.

At the lower end of the copy number distribution, deletion of the entire CNV region (0 copy allele) has been observed in heterozygous form (individuals with a diploid copy number of 1), but never in homozygous form. An estimate of the frequency of such a complete deletion allele is less than 1% in Europeans (Table 2), with a predicted homozygote frequency of 0.01%. At this low frequency, the expected number of 0 copy individuals in a sample size of over 20,000 northern

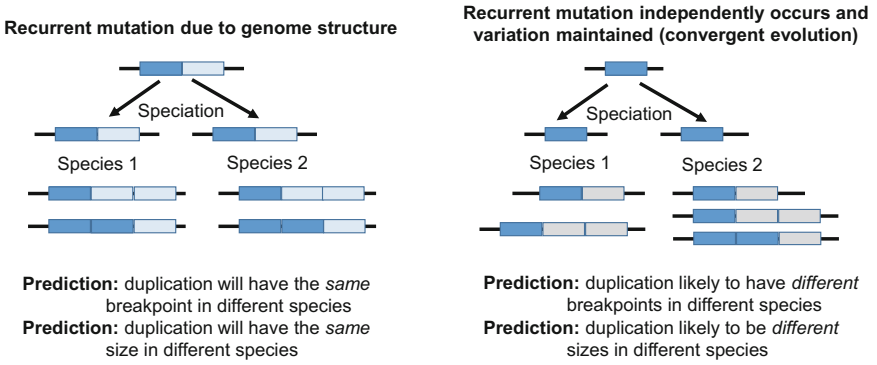


Fig. 6 Two models of CNV evolution across species. The first model shows recurrent generation of CNV due to a shared genomic structure that predisposes to formation of a particular CNV within the region (Fawcett and Innan model, Fawcett and Innan 2013). The second model shows convergent evolution: CNV occurring independently across a genomic region between different lineages and maintained

European individuals is two, so the observed absence of these individuals is consistent with sampling effects (Fisher's exact test, $p = 0.25$). Therefore, we have no evidence at present for deleterious effects of the 0 copy allele shown by selection against zero copy number individuals.

If the CNV really is adaptive in humans, could this be an example of duplication and divergence of coding sequences of β-defensins between copies? Evidence against this comes from firstly from comparative analyses across primates of the CNV genes, which show that negative selection has conserved the coding sequence of these genes, and there is no strong evidence for positive selection (Hollox and Armour 2008). Secondly, analysis of the coding sequences within the human population from exome sequencing data generated by the 1000 Genomes project shows that non-synonymous substitutions are rare, but enriched at low frequencies compared to synonymous substitutions, a hallmark of ongoing negative selection at the coding sequence level (Forni et al. 2015). However, analysis of non-coding variation shows that there is divergence between copies upstream of *DEFB103*, which has been shown to result in functional differences in expression level and differences in response to interferon-gamma (Hardwick et al. 2011).

Nevertheless, the evidence supports the fact that across all copies of the β-defensin genes in the CNV, the coding sequences are the same, and variation in the copy number of the gene could potentially alter the expression levels of the same gene. There is good evidence that β-defensin gene CNV alters levels of the mRNA (Hollox et al. 2003; Janssens et al. 2010) and also of the protein, at least for *DEFB4* and its protein product hbd2. This relationship between gene dosage and protein expression has been shown in the serum of 70 healthy volunteers from the Netherlands (Jansen et al. 2009), and 91 healthy volunteers and 136 volunteers with chronic periodontitis from Germany (Jaradat et al. 2013). The genetic association

Table 2 β-defensin diploid copy number and allele frequency counts in northern Europeans

Copy number	Observed diploid copy number counts	Allele frequency
0	0	0.009
1	45	0.146
2	691	0.568
3	3477	0.233
4	8171	0.041
5	5710	0.001
6	2115	0.003
7	433	0
8	106	0
9+	36	0
total	20,784	1.001

Data from Abujaber et al. (2017), Wain et al. (2014), Aldhous et al. (2010), Fode et al. (2011), Hardwick et al. (2011), Stuart et al. (2012) and unpublished data from our laboratory. Allele frequency and estimated counts using the software CNVice, implemented in the statistical language R (Zuccherato et al. 2017). The CNVice analysis used 1000 repetitions, and 90% of repetitions supported the frequencies shown

between β-defensin copy number and psoriasis suggests a functional link between β-defensin copy number and inflammation, perhaps through variation in epidermal signalling to T cells.

We have recently shown that variation of the CNV and resulting hbd-2 variation affects antimicrobial killing activity, at least in the mucosa of the vagina and possibly elsewhere (James et al. 2017). This provides a direct link between *DEFB4* CNV and a phenotype, antimicrobial killing, that is or has been potentially under natural selection. It seems most likely, therefore, that variation in antimicrobial killing activity and/or immune signalling activity provides the phenotypic variation for selection at this CNV.

8 Summary: From Antimicrobial Peptide to "Jack of All Trades"?

Defensins were first characterised as antimicrobial proteins, and this function continues to interest evolutionary biologists examining the evolution of the family for signatures of natural selection. Indeed, particular defensin clades show strong evidence of duplication and rapid divergence characteristic of natural selection acting on the gene sequences. Any signatures observed are often interpreted in the context of the protein evolving to adapt to a changing microbiota, and that this divergence and duplication is the result of an evolutionary arms race—sometimes characterised as a "red queen" model.

However, functional analysis of defensins in vertebrates, particularly β-defensins, has highlighted that these proteins can have many different functions. For some functions, such as signalling, a model of co-option makes most sense—a β-defensin whose expression is triggered by an infection can be co-opted as a signal to other cells that an infection is occurring, if an interaction with an appropriate receptor can evolve. Several functions, such as the role of β-defensins in hair colour and reproduction, are more difficult to explain in this way and instead point to complete changes in function. It is still the case that the majority of defensins have no demonstrated function and are annotated as antimicrobial in databases only because they are identified as a defensin, and it is likely that, in certain species, certain defensins will have new, unexpected, functions.

Together with the fact that selection signals in defensins seem to vary between genes and organisms, a simple unifying model of host–pathogen evolutionary arms race may not be appropriate, and different evolutionary pressures at different times are likely to explain the diversity of defensins seen in vertebrates today. It is not possible to distinguish, for example, duplication and divergence driven by an evolutionary arms race against bacteria with a similar pattern of natural selection due to a defensin acquiring a new function. Reconstruction of ancestral defensins is an approach that could dissect the evolution of a defensin's interaction with a receptor, but this approach is problematic when analysing evolution against bacteria, as the species and diversity of bacteria at distant points in the evolutionary past are not known.

Carefully determining the variation of defensins within species, with the awareness that defensins can be in regions that are poorly assembled and may be missing particular genes, is useful for evolutionary inference. Comparative analysis of this variation, for example the nature and extent of CNV across species, suggests that some variation is not present simply as a transitory phase of gene duplication and divergence but has itself been subject to natural selection.

Taken together, we believe there is much more to discover in the field of defensins. In particular, we urge evolutionary thinking for functional studies of defensins and functional thinking for evolutionary studies of defensins. A more interdisciplinary approach will yield important insights for defensin function and evolution alike.

Acknowledgements We would like to thank John Armour (University of Nottingham) and Phil Stuart (University of Michigan) for access to raw β-defensin copy number data. We would also like to thank past members of the laboratory who have worked on defensin problems over the past ten years, particularly Rob Hardwick, Lee Machado, Angelica Vittori, Barbara Ottolini, Luciana Zuccherato, Linda Odenthal-Hesse, Nuria Blanco Nevada and our numerous national and international collaborators.

References

Abu Bakar S, Hollox EJ, Armour JA (2009) Allelic recombination between distinct genomic locations generates copy number diversity in human beta-defensins. Proc Natl Acad Sci USA 106(3):853–858. doi:10.1073/pnas.0809073106

Abujaber R, Shea PR, McLaren PJ, Lakhi S, Gilmour J, Allen S, Fellay J, Hollox EJ (2017) No evidence for association of beta-defensin genomic copy number with HIV susceptibility, HIV load during clinical latency, or progression to AIDS. Ann Hum Genet 81(1):27–34. doi:10.1111/ahg.12182

Aerts AM, Francois IE, Cammue BP, Thevissen K (2008) The mode of antifungal action of plant, insect and human defensins. Cell Mol Life Sci 65(13):2069–2079. doi:10.1007/s00018-008-8035-0

Aldhous MC, Bakar SA, Prescott NJ, Palla R, Soo K, Mansfield JC, Mathew CG, Satsangi J, Armour JA (2010) Measurement methods and accuracy in copy number variation: failure to replicate associations of beta-defensin copy number with Crohn's disease. Hum Mol Genet 19 (24):4930–4938

Aldred PM, Hollox EJ, Armour JA (2005) Copy number polymorphism and expression level variation of the human alpha-defensin genes DEFA1 and DEFA3. Hum Mol Genet 14 (14):2045–2052. doi:10.1093/hmg/ddi209

Anderson TM, vonHoldt BM, Candille SI, Musiani M, Greco C, Stahler DR, Smith DW, Padhukasahasram B, Randi E, Leonard JA, Bustamante CD, Ostrander EA, Tang H, Wayne RK, Barsh GS (2009) Molecular and evolutionary history of melanism in North American gray wolves. Science 323(5919):1339–1343. doi:10.1126/science.1165448

Barber JC, Joyce CA, Collinson MN, Nicholson JC, Willatt LR, Dyson HM, Bateman MS, Green AJ, Yates JR, Dennis NR (1998) Duplication of 8p23.1: a cytogenetic anomaly with no established clinical significance. J Med Genet 35(6):491–496

Barber JC, Maloney V, Hollox EJ, Stuke-Sontheimer A, du Bois G, Daumiller E, Klein-Vogler U, Dufke A, Armour JA, Liehr T (2005) Duplications and copy number variants of 8p23.1 are cytogenetically indistinguishable but distinct at the molecular level. Eur J Hum Genet 13 (10):1131–1136. doi:10.1038/sj.ejhg.5201475

Beringer PM, Bensman TJ, Ho H, Agnello M, Denovel N, Nguyen A, Wong-Beringer A, She R, Tran DQ, Moskowitz SM, Selsted ME (2016) Rhesus theta-defensin-1 (RTD-1) exhibits in vitro and in vivo activity against cystic fibrosis strains of Pseudomonas aeruginosa. J Antimicrob Chemother 71(1):181–188. doi:10.1093/jac/dkv301

Bevins CL (2013) Innate immune functions of alpha-defensins in the small intestine. Dig Dis 31 (3–4):299–304. doi:10.1159/000354681

Bickhart DM, Hou Y, Schroeder SG, Alkan C, Cardone MF, Matukumalli LK, Song J, Schnabel RD, Ventura M, Taylor JF, Garcia JF, Van Tassell CP, Sonstegard TS, Eichler EE, Liu GE (2012) Copy number variation of individual cattle genomes using next-generation sequencing. Genome Res 22(4):778–790. doi:10.1101/gr.133967.111

Black HA, Khan FF, Tyson J, Armour J (2014) Inferring mechanisms of copy number change from haplotype structures at the human DEFA1A3 locus. BMC Genom 15:614. doi:10.1186/1471-2164-15-614

Buermans HP, Vossen RH, Anvar SY, Allard WG, Guchelaar HJ, White SJ, den Dunnen JT, Swen JJ, van der Straaten T (2017) Flexible and scalable full-length CYP2D6 long Amplicon PacBio sequencing. Hum Mutat 38(3):310–316. doi:10.1002/humu.23166

Cagliani R, Fumagalli M, Riva S, Pozzoli U, Comi GP, Menozzi G, Bresolin N, Sironi M (2008) The signature of long-standing balancing selection at the human defensin beta-1 promoter. Genome Biol 9(9):R143. doi:10.1186/gb-2008-9-9-r143

Candille SI, Kaelin CB, Cattanach BM, Yu B, Thompson DA, Nix MA, Kerns JA, Schmutz SM, Millhauser GL, Barsh GS (2007) A -defensin mutation causes black coat color in domestic dogs. Science 318(5855):1418–1423. doi:10.1126/science.1147880

Chapman JR, Hellgren O, Helin AS, Kraus RH, Cromie RL, Waldenstrom J (2016) The evolution of innate immune genes: purifying and balancing selection on beta-defensins in waterfowl. Mol Biol Evol 33(12):3075–3087. doi:10.1093/molbev/msw167

Cheng DQ, Li Y, Huang JF (2014) Molecular evolution of the primate alpha-/theta-defensin multigene family. PLoS ONE 9(5):e97425. doi:10.1371/journal.pone.0097425

Cheng Y, Prickett MD, Gutowska W, Kuo R, Belov K, Burt DW (2015) Evolution of the avian beta-defensin and cathelicidin genes. BMC Evol Biol 15:188. doi:10.1186/s12862-015-0465-3

Chu H, Pazgier M, Jung G, Nuccio SP, Castillo PA, de Jong MF, Winter MG, Winter SE, Wehkamp J, Shen B, Salzman NH, Underwood MA, Tsolis RM, Young GM, Lu W, Lehrer RI, Baumler AJ, Bevins CL (2012) Human alpha-defensin 6 promotes mucosal innate immunity through self-assembled peptide nanonets. Science 337(6093):477–481. doi:10.1126/science.1218831

Conrad DF, Pinto D, Redon R, Feuk L, Gokcumen O, Zhang Y, Aerts J, Andrews TD, Barnes C, Campbell P, Fitzgerald T, Hu M, Ihm C, Kristiansson K, Macarthur D, Macdonald J, Onyiah I, Pang A, Robson S, Stirrups K, Valsesia A, Walter K, Wei J, Consortium" WTCC, Tyler-Smith C, Carter N, Lee C, Scherer S, Hurles M (2009) Origins and functional impact of copy number variation in the human genome. Nature 464(7289):704–712

Das S, Nikolaidis N, Goto H, McCallister C, Li J, Hirano M, Cooper MD (2010) Comparative genomics and evolution of the alpha-defensin multigene family in primates. Mol Biol Evol 27(10):2333–2343. doi:10.1093/molbev/msq118

Diamond G, Zasloff M, Eck H, Brasseur M, Maloy WL, Bevins CL (1991) Tracheal antimicrobial peptide, a cysteine-rich peptide from mammalian tracheal mucosa: peptide isolation and cloning of a cDNA. Proc Natl Acad Sci USA 88(9):3952–3956

Dorin JR (2015) Novel phenotype of mouse spermatozoa following deletion of nine beta-defensin genes. Asian J Androl 17(5):716–719. doi:10.4103/1008-682x.159712

Dorin JR, Barratt CL (2014) Importance of beta-defensins in sperm function. Mol Hum Reprod 20(9):821–826. doi:10.1093/molehr/gau050

Eisenhauer PB, Lehrer RI (1992) Mouse neutrophils lack defensins. Infect Immun 60(8):3446–3447

Eisenhauer P, Harwig SS, Szklarek D, Ganz T, Lehrer RI (1990) Polymorphic expression of defensins in neutrophils from outbred rats. Infect Immun 58(12):3899–3902

Fawcett JA, Innan H (2013) The role of gene conversion in preserving rearrangement hotspots in the human genome. Trends in genetics: TIG 29(10):561–568. doi:10.1016/j.tig.2013.07.002

Feng Z, Jiang B, Chandra J, Ghannoum M, Nelson S, Weinberg A (2005) Human beta-defensins: differential activity against candidal species and regulation by Candida albicans. J Dent Res 84(5):445–450. doi:10.1177/154405910508400509

Fernandez-Fuertes B, Narciandi F, O'Farrelly C, Kelly AK, Fair S, Meade KG, Lonergan P (2016) Cauda epididymis-specific beta-defensin 126 promotes sperm motility but not fertilizing ability in cattle. Biol Reprod 95(6):122. doi:10.1095/biolreprod.116.138792

Fode P, Jespersgaard C, Hardwick RJ, Bogle H, Theisen M, Dodoo D, Lenicek M, Vitek L, Vieira A, Freitas J (2011) Determination of beta-defensin genomic copy number in different populations: a comparison of three methods. PLoS ONE 6(2):e16768

Forni D, Martin D, Abujaber R, Sharp AJ, Sironi M, Hollox EJ (2015) Determining multiallelic complex copy number and sequence variation from high coverage exome sequencing data. BMC Genom 16:891. doi:10.1186/s12864-015-2123-y

Gallo SA, Wang W, Rawat SS, Jung G, Waring AJ, Cole AM, Lu H, Yan X, Daly NL, Craik DJ, Jiang S, Lehrer RI, Blumenthal R (2006) Theta-defensins prevent HIV-1 Env-mediated fusion by binding gp41 and blocking 6-helix bundle formation. J Biol Chem 281(27):18787–18792. doi:10.1074/jbc.M602422200

Ganz T, Selsted ME, Szklarek D, Harwig SS, Daher K, Bainton DF, Lehrer RI (1985) Defensins. Natural peptide antibiotics of human neutrophils. J Clin Invest 76(4):1427–1435. doi:10.1172/jci112120

Gokcumen O, Babb PL, Iskow RC, Zhu Q, Shi X, Mills RE, Ionita-Laza I, Vallender EJ, Clark AG, Johnson WE, Lee C (2011) Refinement of primate copy number variation hotspots identifies candidate genomic regions evolving under positive selection. Genome Biol 12(5): R52. doi:10.1186/gb-2011-12-5-r52

Hardwick RJ, Machado LR, Zuccherato LW, Antolinos S, Xue Y, Shawa N, Gilman RH, Cabrera L, Berg DE, Tyler-Smith C (2011) A worldwide analysis of beta-defensin copy number variation suggests recent selection of a high-expressing DEFB103 gene copy in East Asia. Hum Mutat 32(7):743–750

Hedrick PW, Smith DW, Stahler DR (2016) Negative-assortative mating for color in wolves. Evolution 70(4):757–766. doi:10.1111/evo.12906

Hollox EJ, Armour JA (2008) Directional and balancing selection in human beta-defensins. BMC Evol Biol 8(1):113

Hollox EJ, Armour JA, Barber JC (2003) Extensive normal copy number variation of a β-defensin antimicrobial-gene cluster. Am J Hum Genet 73(3):591–600

Hollox EJ, Huffmeier U, Zeeuwen PL, Palla R, Lascorz J, Rodijk-Olthuis D, van de Kerkhof PC, Traupe H, de Jongh G, den Heijer M (2008) Psoriasis is associated with increased beta-defensin genomic copy number. Nat Genet 40(1):23

James C, Bajaj-Elliott M, Abujaber R, Forya F, Klein N, David A, Hollox EJ, Peebles D (2017) Beta-defensin copy number variation affects hbd2 levels and bacterial killing activity of cervical mucus during pregnancy (paper under review)

Jansen PA, Rodijk-Olthuis D, Hollox EJ, Kamsteeg M, Tjabringa GS, de Jongh GJ, van Vlijmen-Willems IM, Bergboer JG, van Rossum MM, de Jong EM (2009) β-Defensin-2 protein is a serum biomarker for disease activity in psoriasis and reaches biologically relevant concentrations in lesional skin. PLoS ONE 4(3):e4725

Janssens W, Nuytten H, Dupont LJ, Van Eldere J, Vermeire S, Lambrechts D, Nackaerts K, Decramer M, Cassiman JJ, Cuppens H (2010) Genomic copy number determines functional expression of β-defensin 2 in airway epithelial cells and associates with chronic obstructive pulmonary disease. Am J Respir Crit Care Med 182(2):163–169. doi:10.1164/rccm.200905-0767OC

Jaradat SW, Hoder-Przyrembel C, Cubillos S, Krieg N, Lehmann K, Piehler S, Sigusch BW, Norgauer J (2013) Beta-defensin-2 genomic copy number variation and chronic periodontitis. J Dent Res 92(11):1035–1040. doi:10.1177/0022034513504217

Khan FF, Carpenter D, Mitchell L, Mansouri O, Black HA, Tyson J, Armour JA (2013) Accurate measurement of gene copy number for human alpha-defensin DEFA1A3. BMC Genom 14:719. doi:10.1186/1471-2164-14-719

Lee AS, Gutierrez-Arcelus M, Perry GH, Vallender EJ, Johnson WE, Miller GM, Korbel JO, Lee C (2008) Analysis of copy number variation in the rhesus macaque genome identifies candidate loci for evolutionary and human disease studies. Hum Mol Genet 17(8):1127–1136. doi:10.1093/hmg/ddn002

Lee MO, Bornelov S, Andersson L, Lamont SJ, Chen J, Womack JE (2016) Duplication of chicken defensin7 gene generated by gene conversion and homologous recombination. Proc Natl Acad Sci USA 113(48):13815–13820. doi:10.1073/pnas.1616948113

Lehrer RI, Lu W (2012) Alpha-defensins in human innate immunity. Immunol Rev 245(1):84–112. doi:10.1111/j.1600-065X.2011.01082.x

Lehrer RI, Cole AM, Selsted ME (2012) Theta-defensins: cyclic peptides with endless potential. J Biol Chem 287(32):27014–27019. doi:10.1074/jbc.R112.346098

Leonard BC, Marks SL, Outerbridge CA, Affolter VK, Kananurak A, Young A, Moore PF, Bannasch DL, Bevins CL (2012) Activity, expression and genetic variation of canine beta-defensin 103: a multifunctional antimicrobial peptide in the skin of domestic dogs. J Innate Immun 4(3):248–259. doi:10.1159/000334566

Li D, Zhang L, Yin H, Xu H, Satkoski Trask J, Smith DG, Li Y, Yang M, Zhu Q (2014) Evolution of primate alpha and theta defensins revealed by analysis of genomes. Mol Biol Rep 41 (6):3859–3866. doi:10.1007/s11033-014-3253-z

Liu GE, Hou Y, Zhu B, Cardone MF, Jiang L, Cellamare A, Mitra A, Alexander LJ, Coutinho LL, Dell'Aquila ME, Gasbarre LC, Lacalandra G, Li RW, Matukumalli LK, Nonneman D, Regitano LC, Smith TP, Song J, Sonstegard TS, Van Tassell CP, Ventura M, Eichler EE, McDaneld TG, Keele JW (2010) Analysis of copy number variations among diverse cattle breeds. Genome Res 20(5):693–703. doi:10.1101/gr.105403.110

Lynn DJ, Lloyd AT, Fares MA, O'Farrelly C (2004) Evidence of positively selected sites in mammalian alpha-defensins. Mol Biol Evol 21(5):819–827. doi:10.1093/molbev/msh084

Maxwell AI, Morrison GM, Dorin JR (2003) Rapid sequence divergence in mammalian beta-defensins by adaptive evolution. Mol Immunol 40(7):413–421

Mohajeri K, Cantsilieris S, Huddleston J, Nelson BJ, Coe BP, Campbell CD, Baker C, Harshman L, Munson KM, Kronenberg ZN, Kremitzki M, Raja A, Catacchio CR, Graves TA, Wilson RK, Ventura M, Eichler EE (2016) Interchromosomal core duplicons drive both evolutionary instability and disease susceptibility of the Chromosome 8p23.1 region. Genome Res 26(11):1453–1467. doi:10.1101/gr.211284.116

Morrison GM, Semple CA, Kilanowski FM, Hill RE, Dorin JR (2003) Signal sequence conservation and mature peptide divergence within subgroups of the murine beta-defensin gene family. Mol Biol Evol 20(3):460–470

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

Nguyen TX, Cole AM, Lehrer RI (2003) Evolution of primate theta-defensins: a serpentine path to a sweet tooth. Peptides 24(11):1647–1654. doi:10.1016/j.peptides.2003.07.023

Ottolini B, Hornsby MJ, Abujaber R, MacArthur JA, Badge RM, Schwarzacher T, Albertson DG, Bevins CL, Solnick JV, Hollox EJ (2014) Evidence of convergent evolution in humans and macaques supports an adaptive role for copy number variation of the beta-defensin-2 gene. Genome Biol Evol 6(11):3025–3038. doi:10.1093/gbe/evu236

Pala RR (2012) Human beta-defensin gene copy number variation and consequences in disease and evolution. University of Nottingham

Palta P, Kaplinski L, Nagirnaja L, Veidenberg A, Mols M, Nelis M, Esko T, Metspalu A, Laan M, Remm M (2015) Haplotype phasing and inheritance of copy number variants in nuclear families. PLoS ONE 10(4):e0122713. doi:10.1371/journal.pone.0122713

Patil A, Hughes AL, Zhang G (2004) Rapid evolution and diversification of mammalian alpha-defensins as revealed by comparative analysis of rodent and primate genes. Physiol Genom 20(1):1–11. doi:10.1152/physiolgenomics.00150.2004

Patil AA, Cai Y, Sang Y, Blecha F, Zhang G (2005) Cross-species analysis of the mammalian beta-defensin gene family: presence of syntenic gene clusters and preferential expression in the male reproductive tract. Physiol Genom 23(1):5–17. doi:10.1152/physiolgenomics.00104.2005

Raschig J, Mailander-Sanchez D, Berscheid A, Berger J, Stromstedt AA, Courth LF, Malek NP, Brotz-Oesterhelt H, Wehkamp J (2017) Ubiquitously expressed Human Beta Defensin 1 (hBD1) forms bacteria-entrapping nets in a redox dependent mode of action. PLoS Pathog 13 (3):e1006261. doi:10.1371/journal.ppat.1006261

Rice WG, Ganz T, Kinkade JM Jr, Selsted ME, Lehrer RI, Parmley RT (1987) Defensin-rich dense granules of human neutrophils. Blood 70(3):757–765

Rohrl J, Yang D, Oppenheim JJ, Hehlgans T (2010) Human beta-defensin 2 and 3 and their mouse orthologs induce chemotaxis through interaction with CCR2. J Immunol (Baltimore, Md: 1950) 184(12):6688–6694. doi:10.4049/jimmunol.0903984

Salzman NH, Ghosh D, Huttner KM, Paterson Y, Bevins CL (2003) Protection against enteric salmonellosis in transgenic mice expressing a human intestinal defensin. Nature 422 (6931):522–526. doi:10.1038/nature01520

Schutte BC, Mitros JP, Bartlett JA, Walters JD, Jia HP, Welsh MJ, Casavant TL, McCray PB Jr (2002) Discovery of five conserved beta-defensin gene clusters using a computational search strategy. Proc Natl Acad Sci USA 99(4):2129–2133. doi:10.1073/pnas.042692699

Semple F, Dorin JR (2012) Beta-defensins: multifunctional modulators of infection, inflammation and more? J Innate Immun 4(4):337–348. doi:10.1159/000336619

Semple CA, Rolfe M, Dorin JR (2003) Duplication and selection in the evolution of primate beta-defensin genes. Genome Biol 4(5):R31

Shafee TM, Lay FT, Hulett MD, Anderson MA (2016) The defensins consist of two independent, convergent protein superfamilies. Mol Biol Evol 33(9):2345–2356. doi:10.1093/molbev/msw106

Shafee TM, Lay FT, Phan TK, Anderson MA, Hulett MD (2017) Convergent evolution of defensin sequence, structure and function. Cell Mol Life Sci 74(4):663–682. doi:10.1007/s00018-016-2344-5

Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Söding J (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol 7(1):539

Sniegowski PD, Gerrish PJ, Johnson T, Shaver A (2000) The evolution of mutation rates: separating causes from consequences. BioEssays 22(12):1057–1066

Stuart PE, Hüffmeier U, Nair RP, Palla R, Tejasvi T, Schalkwijk J, Elder JT, Reis A, Armour JA (2012) Association of β-defensin copy number and psoriasis in three cohorts of European origin. J Investig Dermatol 132(10):2407–2413

Suarez-Carmona M, Hubert P, Delvenne P, Herfs M (2015) Defensins: "Simple" antimicrobial peptides or broad-spectrum molecules? Cytokine Growth Factor Rev 26(3):361–370. doi:10.1016/j.cytogfr.2014.12.005

Sudmant PH, Huddleston J, Catacchio CR, Malig M, Hillier LW, Baker C, Mohajeri K, Kondova I, Bontrop RE, Persengiev S, Antonacci F, Ventura M, Prado-Martinez J, Marques-Bonet T, Eichler EE (2013) Evolution and diversity of copy number variation in the great ape lineage. Genome Res 23(9):1373–1382. doi:10.1101/gr.158543.113

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(7571):75–81. doi:10.1038/nature15394

Tang YQ, Yuan J, Osapay G, Osapay K, Tran D, Miller CJ, Ouellette AJ, Selsted ME (1999) A cyclic antimicrobial peptide produced in primate leukocytes by the ligation of two truncated alpha-defensins. Science 286(5439):498–502

Taylor K, Clarke DJ, McCullough B, Chin W, Seo E, Yang D, Oppenheim J, Uhrin D, Govan JR, Campopiano DJ, MacMillan D, Barran P, Dorin JR (2008) Analysis and separation of residues important for the chemoattractant and antimicrobial activities of beta-defensin 3. J Biol Chem 283(11):6631–6639. doi:10.1074/jbc.M709238200

Teshima KM, Innan H (2012) The coalescent with selection on copy number variants. Genetics 190(3):1077–1086. doi:10.1534/genetics.111.135343

Thornton KR (2007) The neutral coalescent process for recent gene duplications and copy-number variants. Genetics 177(2):987–1000. doi:10.1534/genetics.107.074948

Tollner TL, Yudin AI, Tarantal AF, Treece CA, Overstreet JW, Cherr GN (2008a) Beta-defensin 126 on the surface of macaque sperm mediates attachment of sperm to oviductal epithelia. Biol Reprod 78(3):400–412. doi:10.1095/biolreprod.107.064071

Tollner TL, Yudin AI, Treece CA, Overstreet JW, Cherr GN (2008b) Macaque sperm coating protein DEFB126 facilitates sperm penetration of cervical mucus. Hum Reprod 23(11):2523–2534. doi:10.1093/humrep/den276

Tollner TL, Venners SA, Hollox EJ, Yudin AI, Liu X, Tang G, Xing H, Kays RJ, Lau T, Overstreet JW, Xu X, Bevins CL, Cherr GN (2011) A common mutation in the defensin DEFB126 causes impaired sperm function and subfertility. Sci Trans Med 3(92):92ra65. doi:10.1126/scitranslmed.3002289

Tollner TL, Bevins CL, Cherr GN (2012) Multifunctional glycoprotein DEFB126–a curious story of defensin-clad spermatozoa. Nat Rev Urol 9(7):365–375. doi:10.1038/nrurol.2012.109

Torres AM, Kuchel PW (2004) The beta-defensin-fold family of polypeptides. Toxicon 44 (6):581–588. doi:10.1016/j.toxicon.2004.07.011

Tyson J, Armour JA (2017) Analysis of multiallelic CNVs by emulsion haplotype fusion PCR. Methods Mol Biol (Clifton, NJ) 1492:155–165. doi:10.1007/978-1-4939-6442-0_10

Wain LV, Odenthal-Hesse L, Abujaber R, Sayers I, Beardsmore C, Gaillard EA, Chappell S, Dogaru CM, McKeever T, Guetta-Baranes T, Kalsheker N, Kuehni CE, Hall IP, Tobin MD, Hollox EJ (2014) Copy number variation of the beta-defensin genes in europeans: no supporting evidence for association with lung function, chronic obstructive pulmonary disease or asthma. PLoS ONE 9(1):e84192. doi:10.1371/journal.pone.0084192

Wang W, Mulakala C, Ward SC, Jung G, Luong H, Pham D, Waring AJ, Kaznessis Y, Lu W, Bradley KA, Lehrer RI (2006) Retrocyclins kill bacilli and germinating spores of Bacillus anthracis and inactivate anthrax lethal toxin. J Biol Chem 281(43):32755–32764. doi:10.1074/jbc.M603614200

Wang J, Wang H, Jiang J, Kang H, Feng X, Zhang Q, Liu JF (2013) Identification of genome-wide copy number variations among diverse pig breeds using SNP genotyping arrays. PLoS ONE 8 (7):e68683. doi:10.1371/journal.pone.0068683

Waterhouse AM, Procter JB, Martin DM, Clamp M, Barton GJ (2009) Jalview Version 2—a multiple sequence alignment editor and analysis workbench. Bioinformatics 25(9):1189–1191

Whittington CM, Papenfuss AT, Bansal P, Torres AM, Wong ES, Deakin JE, Graves T, Alsop A, Schatzkamer K, Kremitzki C, Ponting CP, Temple-Smith P, Warren WC, Kuchel PW, Belov K (2008a) Defensins and the convergent evolution of platypus and reptile venom genes. Genome Res 18(6):986–994. doi:10.1101/gr.7149808

Whittington CM, Papenfuss AT, Kuchel PW, Belov K (2008b) Expression patterns of platypus defensin and related venom genes across a range of tissue types reveal the possibility of broader functions for OvDLPs than previously suspected. Toxicon 52(4):559–565. doi:10.1016/j.toxicon.2008.07.002

Wiens ME, Wilson SS, Lucero CM, Smith JG (2014) Defensins and viral infection: dispelling common misconceptions. PLoS Pathog 10(7):e1004186. doi:10.1371/journal.ppat.1004186

Wilson SS, Wiens ME, Holly MK, Smith JG (2016) Defensins at the mucosal surface: latest insights into defensin-virus interactions. J Virol 90(11):5216–5218. doi:10.1128/jvi.00904-15

Yang D, Chertov O, Bykovskaia SN, Chen Q, Buffo MJ, Shogan J, Anderson M, Schroder JM, Wang JM, Howard OM, Oppenheim JJ (1999) Beta-defensins: linking innate and adaptive immunity through dendritic and T cell CCR6. Science 286(5439):525–528

Yang D, Chen Q, Chertov O, Oppenheim JJ (2000) Human neutrophil defensins selectively chemoattract naive T and immature dendritic cells. J Leukoc Biol 68(1):9–14

Yona AH, Frumkin I, Pilpel Y (2015) A relay race on the evolutionary adaptation spectrum. Cell 163(3):549–559. doi:10.1016/j.cell.2015.10.005

Yount NY, Kupferwasser D, Spisni A, Dutz SM, Ramjan ZH, Sharma S, Waring AJ, Yeaman MR (2009) Selective reciprocity in antimicrobial activity versus cytotoxicity of hBD-2 and crotamine. Proc Natl Acad Sci USA 106(35):14972–14977. doi:10.1073/pnas.0904465106

Yudin AI, Generao SE, Tollner TL, Treece CA, Overstreet JW, Cherr GN (2005) Beta-defensin 126 on the cell surface protects sperm from immunorecognition and binding of anti-sperm antibodies. Biol Reprod 73(6):1243–1252. doi:10.1095/biolreprod.105.042432

Zhou CX, Zhang YL, Xiao L, Zheng M, Leung KM, Chan MY, Lo PS, Tsang LL, Wong HY, Ho LS, Chung YW, Chan HC (2004) An epididymis-specific beta-defensin is important for the initiation of sperm maturation. Nat Cell Biol 6(5):458–464. doi:10.1038/ncb1127

Zhou YS, Webb S, Lettice L, Tardif S, Kilanowski F, Tyrrell C, Macpherson H, Semple F, Tennant P, Baker T, Hart A, Devenney P, Perry P, Davey T, Barran P, Barratt CL, Dorin JR (2013) Partial deletion of chromosome 8 beta-defensin cluster confers sperm dysfunction and infertility in male mice. PLoS Genet 9(10):e1003826. doi:10.1371/journal.pgen.1003826

Zuccherato LW, Schneider S, Tarazona-Santos E, Hardwick R, Berg DE, Bogle H, Gouveia M, Machado L, Machado M, Rodrigues-Soares F, Soares-Souza G, Togni D, Zamudio R, Gilman RH, Duarte D, Hollox E, Rodrigues M (2017) Population genetics of immune-related multilocus copy number variation in Native Americans. J R Soc Interface 14(128). doi:10.1098/rsif.2017.0057

Interdependencies Between the Adaptation and Interference Modules Guide Efficient CRISPR-Cas Immunity

Ekaterina Semenova and Konstantin Severinov

Abstract CRISPR-Cas is a common adaptive RNA-guided prokaryotic immunity mechanism that limits the spread of mobile genetic elements such as phages and plasmids. A CRISPR-Cas system is composed of two seemingly independent modules. Cas proteins from the adaptation module are responsible for recording prior encounters with mobile genetic elements by incorporating fragments of foreign DNA into CRISPR array. Small protective RNAs generated after CRISPR array transcription are used by the interference module Cas proteins to locate complementary nucleic acids and destroy them. Here, we discuss how the activities and substrate preferences of these two functional modules must be tightly coordinated to provide efficient defence against foreign DNA.

1 Prokaryotic Immunity Systems CRISPR-Cas Operate Through Two Functionally Independent Modules

CRISPR-Cas systems are defence mechanisms that provide prokaryotes with adaptive immunity against bacteriophages and other mobile genetic elements by targeting their DNA and/or RNA (Barrangou et al. 2007; Brouns et al. 2008; Hale et al. 2009; Marraffini and Sontheimer 2008). CRISPR-Cas systems comprise *C*lusters of *R*egularly *I*nterspaced *S*hort *P*alindromic *R*epeats (CRISPR) and *C*RISPR-*as*sociated (*cas*) genes. While highly diverse (Makarova et al. 2015), all CRISPR-Cas systems share a common logic of function (Mohanraju et al. 2016), which is schematically depicted in Fig. 1. In the course of infection, short fragments

E. Semenova (✉) · K. Severinov (✉)
Waksman Institute for Microbiology, Rutgers, The State University of New Jersey, Piscataway, NJ, USA
e-mail: semenova@waksman.rutgers.edu

K. Severinov
e-mail: severik@waksman.rutgers.edu

K. Severinov
Skolkovo Institute of Science and Technology, Skolkovo, Russia

P. Pontarotti (ed.), *Evolutionary Biology: Self/Nonself Evolution, Species and Complex Traits Evolution, Methods and Concepts*,
DOI 10.1007/978-3-319-61569-1_3

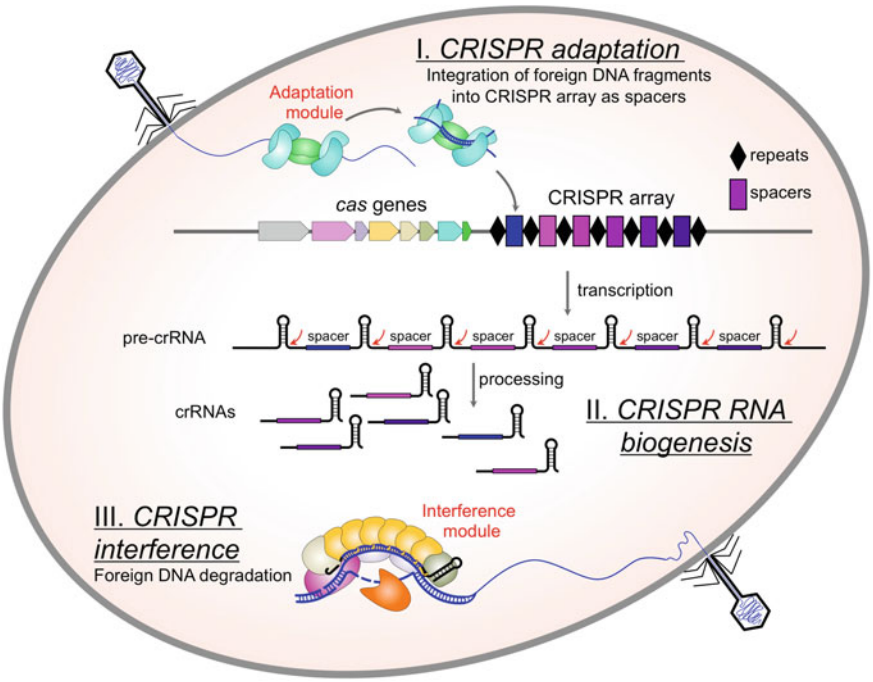


Fig. 1 CRISPR-Cas adaptive immunity. The three stages of CRISPR-Cas system function are schematically illustrated. During CRISPR adaptation, the injection of phage DNA into bacterial cell (illustrated at the *upper left*) activates the Cas1–Cas2 adaptation module proteins which excise spacer-sized fragments of phage DNA and channels then for incorporation into CRISPR array. During CRISPR RNA biogenesis, CRISPR array is transcribed and resulting pre-crRNA is processed at repeat sequences to generated crRNAs. Individual crRNAs are bound by Cas protein effectors. When phage DNA with sequences matching a CRISPR spacer appears in the cell (*lower right*), effectors programmed by appropriate crRNA bind to it and the resulting R-loop complex is destroyed by Cas executor nuclease

of viral DNA, termed protospacers, can be selected by the Cas1–Cas2 proteins complex and integrated into CRISPR array. As a result, the CRISPR array is extended; for every spacer acquired an additional copy of CRISPR repeat is generated. The process of spacer acquisition is referred to as "CRISPR adaptation". The CRISPR array is transcribed and the resulting non-coding transcript, termed "pre-crRNA", containing unique spacers separated by identical repeats is processed to generate individual short CRISPR RNAs (crRNAs). Each crRNA contains common flanking sequences comprising repeat fragments and a spacer of variable sequence. Individual crRNAs are bound to Cas proteins forming an "effector complex". A repertoire of different crRNAs present in a prokaryotic cell represents its potential to fight off infections by specific invaders. This potential is realized at the stage of "CRISPR interference", during which protospacers in foreign nucleic acids are recognized through complementary interactions with spacers of crRNA in

the effector complex. Upon the location of target protospacers, the foreign DNA is destroyed. While most CRISPR-Cas systems recognize DNA targets, systems with effectors that recognize protospacers in nascent transcripts have been described. In such systems, recognition of nascent RNA target stimulates degradation of DNA templates from which the RNA is transcribed (Kazlauskiene et al. 2016; Samai et al. 2015). In addition, RNA-recognizing effectors that upon target recognition activate non-specific "collateral" destruction of non-target RNA have been described (Abudayyeh et al. 2016). Such systems may not save the infected cell but can act for the benefit of clonal population be limiting the spread of viral infections (Koonin and Zhang 2017).

As is apparent from the above description and from Fig. 1, CRISPR-Cas systems can be regarded as a composition of two functionally independent modules. The adaptation module consists of just two proteins, Cas1 and Cas2 that modify CRISPR array by insertion of new spacers. Cas1 is the most (in fact, the only) strictly conserved protein among known functional CRISPR-Cas systems. When provided with suitable substrates (spacer-sized DNA fragments), Cas1 and Cas2 are sufficient to perform the spacer acquisition reaction both in vivo and in vitro, provided that a DNA molecule with a cognate CRISPR array is present (Arslan et al. 2014; Nunez et al. 2015b; Yosef et al. 2012). The interference modules are much more diverse and their varieties form a basis of CRISPR-Cas system classification (Makarova et al. 2015). Effectors of Class 1 systems consist of multiple subunits. Target destruction can be carried out by dedicated subunits of the complex or through recruitment of additional Cas nuclease executors (Mohanraju et al. 2016). Class 2 effectors provide an "all in one" integrated solution to target recognition and destruction. They are composed of single large polypeptides that recognize targets through crRNA spacer and then destroy them through endonucleolytic activities that reside in the same protein (Shmakov et al. 2017). Effectors of both classes function without the adaptation module proteins both in vivo and in vitro.

Despite the functional independence of the two modules, which underlies the practical use of CRISPR-Cas system for genome editing, in their native biological context their activities and substrate preferences must be tightly coordinated at multiple levels to accomplish the proper function of the entire system. Conceptual and experimental analysis of these coordination requirements can provide insights into the evolution of CRISPR-Cas.

2 CRISPR Repeats Are the Structural Basis for Precise Spacer Integration and crRNA Processing

Both the adaptation and interference module components must be able to recognize CRISPR repeats, whose sequences and lengths can differ significantly even between closely related systems. The adaptation complex Cas1–Cas2, when loaded with spacer precursor, specifically binds CRISPR repeat (Moch et al. 2016) and catalyses

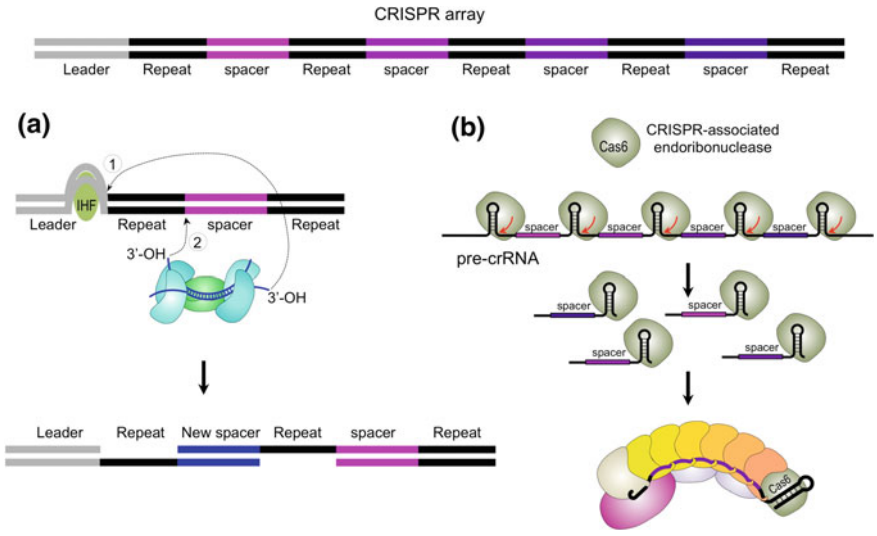


Fig. 2 CRISPR repeat recognition by the adaptation and interference module proteins. A CRISPR array is shown at the top. In **a**, spacer acquisition reaction catalysed by the Cas1–Cas2 adaptation complex is shown. The 3′ ends of foreign DNA fragment are used sequentially to attack the opposite strands at each end of the leader-proximal repeat. The first attack at the leader-repeat junction is stimulated by DNA bending be leader-bound IHF architectural protein. The partially double-stranded structure shown at the bottom is repaired by filling-in and ligation. As a result, an expanded array with new spacer (blue) and an extra copy of repeat is generated. In **b**, activity of Cas6 endonuclease that specifically recognizes repeat sequences in pre-crRNA and cleaves them to generate unit-sized crRNA is shown

two rounds of 3′-OH attack by spacer precursor DNA fragment at both ends (but different strands) of the repeat. The resulting partially double-stranded structure showed in Fig. 2, left-hand side, is filled in by cellular DNA polymerases to generate, after ligation, an expanded array. The presence of a single repeat is sufficient for spacer acquisition, though additional sequences, such as the leader sequence present at one end of the array, strongly stimulate the acquisition efficiency, directing a vast majority of new spacers to one end of the array (Yosef et al. 2012). It was shown that the first nucleophilic attack relies on sequence-specific recognition of leader-repeat junction by the Cas1–Cas2 complex that is strongly stimulated by bending of DNA by IHF (integration host factor), which binds in the leader (Nunez et al. 2016; Yoganand et al. 2017). The site of the second nucleophilic attack at the other end of the repeat is introduced through a ruler-like mechanism. The size of the Cas1–Cas2 complex fits precisely the size of the cognate repeat and the two staggered cuts introduced by Cas1–Cas2 account for constant size of new repeat copy generated during adaptation (Goren et al. 2016) and integration of spacers of defines size (see below). Mechanistically, the adaptation enzymes are similar to site-specific recombination and transposition proteins that generate direct repeats and the junction sites between inserted DNA and the

target sequence. Recent analysis identified stand-alone Cas1-like integrases or "casposases" that may be similar to precursors of CRISPR-Cas adaptation modules (Krupovic et al. 2014).

At the interference side, pre-crRNA processing during crRNA biogenesis requires specific recognition and cleavage at repeat sequences. This task is accomplished by dedicated endoribonucleases in the case of Class 1 systems that bind steam loop structure formed by palindromic repeat sequences (Hochstrasser and Doudna 2015). While often these proteins are part of the effector complex, they appear to be loosely bound. Moreover, when a source of crRNA is provided independently of the normal biogenesis pathway (i.e. through ribozyme cleavage or by RNA polymerase transcription of specially designed short transcription units) endonucleases that process pre-crRNA become dispensable (Maier et al. 2015; Semenova et al. 2015). In the case of some Class 2 systems, pre-crRNA processing requires an additional small tracRNA that is complementary to repeat sequence (repeats in such systems are asymmetric and such systems are therefore technically not "CRIS*P*R" for their repeats are not *P*alindromic). The duplex structure formed at the region of pre-crRNA repeat-tracRNA complementarity is recognized by cellular (non-Cas) RNases that generate crRNA. In some Class 1 and Class 2 systems, the 3′ ends of RNAs generated after initial pre-crRNA processing at repeat sequences are further trimmed—either by effector complex or by unidentified nucleases—to generate mature crRNAs whose spacer parts are shorter than spacers in the array (Deltcheva et al. 2011; Hatoum-Aslan et al. 2013; Rouillon et al. 2013; Zhang et al. 2012). Whatever the mechanism of mature crRNA generation, the effector complex must contain a specific site for recognition of repeat or its fragments sequences to specifically bind only crRNA and thus avoid destruction of unintended targets.

The above analysis seems to clearly point out that present-day CRISPR arrays must have originated from ancestral sequences that contained a single "repeat". Transcription of such a sequence should have generated an RNA with 5′-proximal "protorepeat" bound by the ancestral effector and a downstream sequence that after non-specific processing could have been used to direct target recognition, with or without its nucleolytic destruction. This hypothetical early stage CRISPR effector is functionally similar to abundant RNA binding proteins such as Hfq that mediate the antisense function of prokaryotic small RNAs (Vogel and Luisi 2011). Such a primitive system obviously did not require an adaptation module for it was not adaptive. The presence of an adaptation complex capable of recognizing a DNA sequence that was, as an RNA, bound by ancestral effector could have been the first step in generation of a primitive adaptive CRISPR-Cas system. Available evidence suggests that such events happened several times in evolution. Multiple solutions for generation of functional crRNA biogenesis from pre-crRNA indicate that this step, which becomes essential only when arrays become long, was a later addition.

Stand-alone ancestral interference module could have functioned as an innate immunity system and could resemble present-day prokaryotic Argonaute (pAgo) proteins. These recently discovered counterparts of eukaryotic Argonaute family provide defence against foreign DNA (Olovnikov et al. 2013; Swarts et al. 2014).

The pAgo proteins require guides to bring them to their targets. Some prokaryotic Argonautes use RNA guides that are produced by cellular nucleases. In contrast, pAgo from *Thermus thermophilus* generates its own guide DNA molecules through a "chopping" (Swarts et al. 2017).

3 Common Motifs Are Recognized by Adaptation and Interference Modules During Spacer Selection and Target Recognition/Destruction

The next point of functional interaction between the adaptation and interference modules is necessary first to avoid self-immunity and second to ensure that spacers brought in by the adaptation module have a protective function at the interference stage. Inspection of the general CRISPR-Cas system function scheme presented in Fig. 1 immediately poses a question: how is this that the effector complex loaded with crRNA does not recognize the CRISPR array spacer from which this crRNA was transcribed? The answer to this question is that target recognition by effectors requires, in addition to crRNA spacer-target DNA complementarity, a presence of a protospacer adjacent motif (PAM). Repeat sequences located at positions corresponding to the location of PAM in DNA targets do not function as PAM, thus solving the auto-immunity problem and allowing the self versus non-self discrimination (Fig. 3a). PAM sequences and their locations (upstream or downstream of the protospacer) vary for different CRISPR-Cas systems. Cognate PAMs are specifically recognized by effectors as double-stranded DNA (Hayes et al. 2016). PAM recognition initiates a complex and very poorly understood process of target interrogation. It requires gradual melting of target DNA and determination whether the melted sequence is complementary to crRNA spacer and should be considered as a protospacer destined for destruction. In the case of effectors recognizing RNA targets there is no problem of self versus non-self discrimination and single-stranded targets can be recognized by complementarity alone without the need of an energetically costly and slow melting process. Indeed, RNA-recognizing effectors appear to function without PAM or very simple rudimentary PAMs (Abudayyeh et al. 2016; Elmore et al. 2016; Marraffini and Sontheimer 2010).

The requirement for PAM introduces an important constraint on the adaptation module. To provide effective immunity, acquired spacers must originate from protospacers with a functional PAM. Indeed, in the cases where this has been studied, the Cas1–Cas2 adaptation complex recognizes PAM sequences and then proceeds to excise adjacent spacer-sized fragments channelled for incorporation into array. Though PAM itself is not incorporated in the array, its presence affects the orientation of the inserted spacer in vitro and in vivo, introducing a strong bias towards the orientation that is compatible with subsequent recognition of the target by crRNA in a way that promotes interference. PAM recognition by the adaptation and interference modules proceeds through unrelated mechanisms and must be a

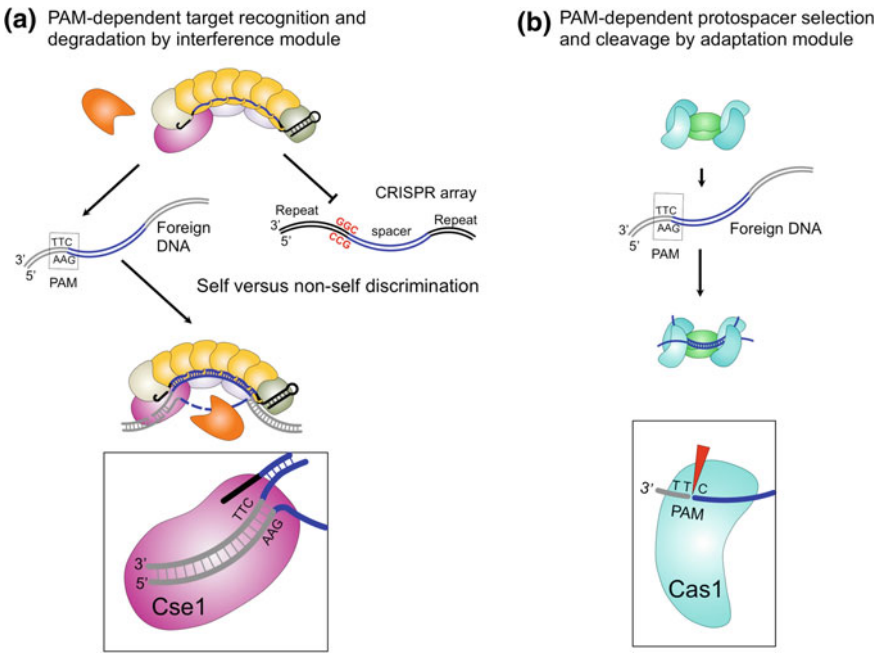


Fig. 3 A common PAM sequence is recognized by the adaptation and interference modules. In **a**, the role of PAM (protospacer adjacent motif) in distinguishing self and non-self during CRISPR interference is shown. Targets recognized by effector component Cse1 and destined for destruction contain a consensus PAM (5′-AAG-3′ in the case of *E. coli* CRISPR-Cas Type I–E system). The corresponding position in repeat is occupied by a 5′-CCG-3′ sequence which does not function as PAM. During CRISPR adaptation (**b**), the Cas1–Cas2 complex recognizes consensus 5′-AAG-3′ PAM to initiate excision of intermediates of space acquisition reaction

product of convergent evolution at conditions of strong positive selection. An analogous mechanism must have been operational during the evolution of DNA binding specificities of cognate restriction endonucleases and methyltransferases, whose DNA recognition domains are unrelated.

In an interesting twist showing the importance of coordination between the adaptation and interference pathways of the CRISPR-Cas response, it has been shown that spacers acquired by Class 1 systems at conditions of primed adaptation, when both interference and adaptation happen simultaneously in the same cell, almost universally originate from protospacers with optimal consensus interference-proficient PAMs. This is illustrated by Fig. 3b example. In *E. coli*, Class 1 Type I-E system requires an AAG PAM for target interference. Overproduction of Cas1 and Cas2 proteins in the absence of interference module proteins promotes efficient spacer acquisition but only 40% of newly acquired spacers originate from protospacers with AAG PAM. In contrast, when spacers are acquired from targets being destroyed by the interference machinery, almost 100% of spacers originate from AAG PAM protospacers, providing a positive feedback

loop that should stimulate more rapid destruction of invader DNA. The mechanism that accounts for increased specificity during primed adaptation is subject to debate and may involve, for example, the generation of interference intermediates enriched with PAMs or direct modulation of the adaptation complex by interaction with interference proteins.

4 The Length of Spacers Selected by the Adaptation Module and Processed into CrRNA Must Fit the Interference Module Effector

The third point of coordination between the adaptation and interference machinery concerns the length of spacers. Bioinformatics analysis indicates that spacer lengths can vary, both within an array and, more significantly, between arrays belonging to different systems (Kuznedelov et al. 2016). During adaptation, spacer length is determined by a ruler-like mechanism schematically presented in Fig. 4a. The adaptation complex binds to and processes DNA fragments producing the spacer-precursors of certain length. The best-characterized adaptation complex isolated from *E. coli* consists of two Cas1 dimers bound to Cas2 dimer. 23 bp-duplex DNA is located between the Cas1 dimers and the length of the duplex is determined by the Cas2 dimer bridge and bracket conserved tyrosine residues of Cas1. Unpaired or 3′-overhung DNA ends, one of them containing PAM-complementary sequence 5′-CTT-3′, are placed into Cas1 nuclease catalytic sites and undergo cleavage 5-nt away from both sides of the duplex. When the resulting intermediate is incorporated into CRISPR array a 33-bp spacer appears after the filling-in reaction (Nunez et al. 2015a; Wang et al. 2015). Some variation in the length of acquired spacers can arise because of exonucleolytic trimming

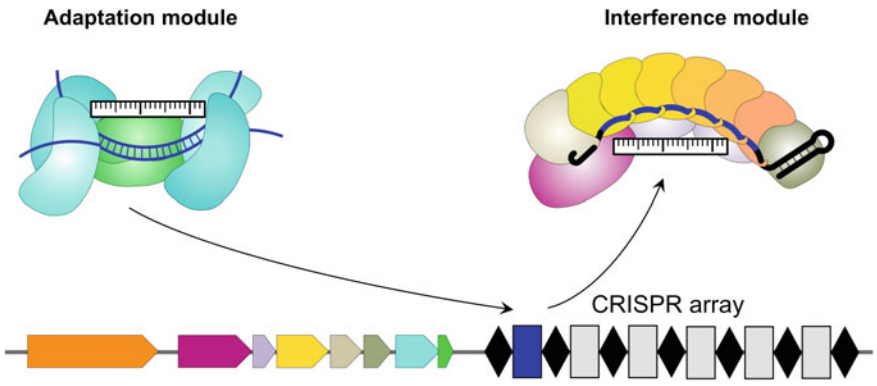


Fig. 4 The adaptation and effector complex use ruler-like mechanism to bind DNA and RNA substrates of matching lengths. See text for details

and/or nucleotide additions to intermediates of acquisition reaction shown in Fig. 2, but overall the length of acquired spacers is strongly constrained by the Cas1–Cas2 complex structure. From a functional perspective, spacer length must be sufficient to specifically recognize a foreign target and avoid the recognition of similar sequences in cellular genome. Additionally, the spacer must be long enough to allow the formation of a stable R-loop complex with target DNA needed to initiate its destruction (Fig. 1). For systems where mature crRNA contains flanking repeat sequences at both sides of the spacer segment (most Class 1 effectors), the length of spacer-binding site of the effector must also match the size of spacers acquired by the adaptation machinery. While diverse, the multisubunit Class 1 effectors are organized in a similar way. The effector complex has an extended and bent "seahorse" appearance with "head" and "tail" elements recognizing repeat sequences of crRNA located at opposing outside edges of the molecule. The backbone part of the seahorse is binding to crRNA spacer, exposing its nucleotides and making them available for interrogation of potential protospacers associated with a suitable PAM. Interestingly, the spacer-binding part of effector is an oligomer of identical RNA binding subunits. In most studied cases, the number of monomers is six, but recently an effector containing only three molecules has been reported. Three groups have recently reported that modification of crRNA spacer length leads to assembly of effector complexes with altered—increased or decreased—stoichiometry of spacer RNA binding subunits (Gleditzsch et al. 2016; Kuznedelov et al. 2016; Luo et al. 2016). In the case of the *E. coli* effector complex, decreasing the length of crRNA spacer from 32 to 26, 20 and 14 bases leads to formation of effector complexes with 5, 4 or 3 crRNA spacer-binding subunits instead of the hexamer found in the standard complex. Predictably, the ability of such shortened complexes to recognize targets and cause their destruction is decreased with decreasing size, due to decreased stability of shorter crRNA-protospacer DNA duplex. Yet, even the smallest complexes are capable of residual CRISPR interference. At conditions of primed adaptation, such miniaturized effectors stimulate the acquisition of standard 33-bp spacers from target DNA. From the evolutionary perspective, these results seem to suggest that the present-day CRISPR-Cas systems spacer lengths have evolved by mutual adjustments and growth of "rulers" used by the adaptation and interference modules until a length that is sufficient for specific recognition of foreign targets was achieved.

Acknowledgements This work is supported by an NIH grant GM10407.

References

Abudayyeh OO, Gootenberg JS, Konermann S, Joung J, Slaymaker IM, Cox DB, Shmakov S, Makarova KS, Semenova E, Minakhin L, Severinov K, Regev A, Lander ES, Koonin EV, Zhang F (2016) C2c2 is a single-component programmable RNA-guided RNA-targeting CRISPR effector. Science 353(6299):aaf5573. doi:10.1126/science.aaf5573

Arslan Z, Hermanns V, Wurm R, Wagner R, Pul U (2014) Detection and characterization of spacer integration intermediates in type I-E CRISPR-Cas system. Nucl Acids Res 42(12):7884–7893. doi:10.1093/nar/gku510

Barrangou R, Fremaux C, Deveau H, Richards M, Boyaval P, Moineau S, Romero DA, Horvath P (2007) CRISPR provides acquired resistance against viruses in prokaryotes. Science 315 (5819):1709–1712. doi:10.1126/science.1138140

Brouns SJ, Jore MM, Lundgren M, Westra ER, Slijkhuis RJ, Snijders AP, Dickman MJ, Makarova KS, Koonin EV, van der Oost J (2008) Small CRISPR RNAs guide antiviral defense in prokaryotes. Science 321(5891):960–964. doi:10.1126/science.1159689

Deltcheva E, Chylinski K, Sharma CM, Gonzales K, Chao Y, Pirzada ZA, Eckert MR, Vogel J, Charpentier E (2011) CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III. Nature 471(7340):602–607. doi:10.1038/nature09886

Elmore JR, Sheppard NF, Ramia N, Deighan T, Li H, Terns RM, Terns MP (2016) Bipartite recognition of target RNAs activates DNA cleavage by the Type III-B CRISPR-Cas system. Genes Dev 30(4):447–459. doi:10.1101/gad.272153.115

Gleditzsch D, Muller-Esparza H, Pausch P, Sharma K, Dwarakanath S, Urlaub H, Bange G, Randau L (2016) Modulating the cascade architecture of a minimal Type I-F CRISPR-Cas system. Nucl Acids Res 44(12):5872–5882. doi:10.1093/nar/gkw469

Goren MG, Doron S, Globus R, Amitai G, Sorek R, Qimron U (2016) Repeat size determination by two molecular rulers in the Type I-E CRISPR array. Cell Rep 16(11):2811–2818. doi:10.1016/j.celrep.2016.08.043

Hale CR, Zhao P, Olson S, Duff MO, Graveley BR, Wells L, Terns RM, Terns MP (2009) RNA-guided RNA cleavage by a CRISPR RNA-Cas protein complex. Cell 139(5):945–956. doi:10.1016/j.cell.2009.07.040

Hatoum-Aslan A, Samai P, Maniv I, Jiang W, Marraffini LA (2013) A ruler protein in a complex for antiviral defense determines the length of small interfering CRISPR RNAs. J Biol Chem 288(39):27888–27897. doi:10.1074/jbc.M113.499244

Hayes RP, Xiao Y, Ding F, van Erp PB, Rajashankar K, Bailey S, Wiedenheft B, Ke A (2016) Structural basis for promiscuous PAM recognition in type I-E Cascade from E. coli. Nature. doi:10.1038/nature16995

Hochstrasser ML, Doudna JA (2015) Cutting it close: CRISPR-associated endoribonuclease structure and function. Trends Biochem Sci 40(1):58–66. doi:10.1016/j.tibs.2014.10.007

Kazlauskiene M, Tamulaitis G, Kostiuk G, Venclovas C, Siksnys V (2016) Spatiotemporal control of Type III-A CRISPR-Cas immunity: coupling DNA degradation with the target RNA recognition. Mol Cell 62(2):295–306. doi:10.1016/j.molcel.2016.03.024

Koonin EV, Zhang F (2017) Coupling immunity and programmed cell suicide in prokaryotes: life-or-death choices. BioEssays 39(1):1–9. doi:10.1002/bies.201600186

Krupovic M, Makarova KS, Forterre P, Prangishvili D, Koonin EV (2014) Casposons: a new superfamily of self-synthesizing DNA transposons at the origin of prokaryotic CRISPR-Cas immunity. BMC Biol 12:36. doi:10.1186/1741-7007-12-36

Kuznedelov K, Mekler V, Lemak S, Tokmina-Lukaszewska M, Datsenko KA, Jain I, Savitskaya E, Mallon J, Shmakov S, Bothner B, Bailey S, Yakunin AF, Severinov K, Semenova E (2016) Altered stoichiometry Escherichia coli cascade complexes with shortened CRISPR RNA spacers are capable of interference and primed adaptation. Nucl Acids Res 44 (22):10849–10861. doi:10.1093/nar/gkw914

Luo ML, Jackson RN, Denny SR, Tokmina-Lukaszewska M, Maksimchuk KR, Lin W, Bothner B, Wiedenheft B, Beisel CL (2016) The CRISPR RNA-guided surveillance complex in Escherichia coli accommodates extended RNA spacers. Nucl Acids Res. doi:10.1093/nar/gkw421

Maier LK, Stachler AE, Saunders SJ, Backofen R, Marchfelder A (2015) An active immune defense with a minimal CRISPR (clustered regularly interspaced short palindromic repeats) RNA and without the Cas6 protein. J Biol Chem 290(7):4192–4201. doi:10.1074/jbc.M114.617506

Makarova KS, Wolf YI, Alkhnbashi OS, Costa F, Shah SA, Saunders SJ, Barrangou R, Brouns SJ, Charpentier E, Haft DH, Horvath P, Moineau S, Mojica FJ, Terns RM, Terns MP, White MF, Yakunin AF, Garrett RA, van der Oost J, Backofen R, Koonin EV (2015) An updated evolutionary classification of CRISPR-Cas systems. Nat Rev Microbiol 13(11):722–736. doi:10.1038/nrmicro3569

Marraffini LA, Sontheimer EJ (2008) CRISPR interference limits horizontal gene transfer in staphylococci by targeting DNA. Science 322(5909):1843–1845. doi:10.1126/science.1165771

Marraffini LA, Sontheimer EJ (2010) Self versus non-self discrimination during CRISPR RNA-directed immunity. Nature 463(7280):568–571. doi:10.1038/nature08703

Moch C, Fromant M, Blanquet S, Plateau P (2016) DNA binding specificities of Escherichia coli Cas1-Cas2 integrase drive its recruitment at the CRISPR locus. Nucl Acids Res. doi:10.1093/nar/gkw1309

Mohanraju P, Makarova KS, Zetsche B, Zhang F, Koonin EV, van der Oost J (2016) Diverse evolutionary roots and mechanistic variations of the CRISPR-Cas systems. Science 353(6299): aad5147. doi:10.1126/science.aad5147

Nunez JK, Harrington LB, Kranzusch PJ, Engelman AN, Doudna JA (2015a) Foreign DNA capture during CRISPR-Cas adaptive immunity. Nature 527(7579):535–538. doi:10.1038/nature15760

Nunez JK, Lee AS, Engelman A, Doudna JA (2015b) Integrase-mediated spacer acquisition during CRISPR-Cas adaptive immunity. Nature 519(7542):193–198. doi:10.1038/nature14237

Nunez JK, Bai L, Harrington LB, Hinder TL, Doudna JA (2016) CRISPR immunological memory requires a host factor for specificity. Mol Cell 62(6):824–833. doi:10.1016/j.molcel.2016.04.027

Olovnikov I, Chan K, Sachidanandam R, Newman DK, Aravin AA (2013) Bacterial argonaute samples the transcriptome to identify foreign DNA. Mol Cell 51(5):594–605. doi:10.1016/j.molcel.2013.08.014

Rouillon C, Zhou M, Zhang J, Politis A, Beilsten-Edmands V, Cannone G, Graham S, Robinson CV, Spagnolo L, White MF (2013) Structure of the CRISPR interference complex CSM reveals key similarities with cascade. Mol Cell 52(1):124–134. doi:10.1016/j.molcel.2013.08.020

Samai P, Pyenson N, Jiang W, Goldberg GW, Hatoum-Aslan A, Marraffini LA (2015) Co-transcriptional DNA and RNA cleavage during Type III CRISPR-Cas immunity. Cell 161(5):1164–1174. doi:10.1016/j.cell.2015.04.027

Semenova E, Kuznedelov K, Datsenko KA, Boudry PM, Savitskaya EE, Medvedeva S, Beloglazova N, Logacheva M, Yakunin AF, Severinov K (2015) The Cas6e ribonuclease is not required for interference and adaptation by the E. coli type I-E CRISPR-Cas system. Nucl Acids Res 43(12):6049–6061. doi:10.1093/nar/gkv546

Shmakov S, Smargon A, Scott D, Cox D, Pyzocha N, Yan W, Abudayyeh OO, Gootenberg JS, Makarova KS, Wolf YI, Severinov K, Zhang F, Koonin EV (2017) Diversity and evolution of class 2 CRISPR-Cas systems. Nat Rev Microbiol 15(3):169–182. doi:10.1038/nrmicro.2016.184

Swarts DC, Jore MM, Westra ER, Zhu Y, Janssen JH, Snijders AP, Wang Y, Patel DJ, Berenguer J, Brouns SJ, van der Oost J (2014) DNA-guided DNA interference by a prokaryotic Argonaute. Nature 507(7491):258–261. doi:10.1038/nature12971

Swarts DC, Szczepaniak M, Sheng G, Chandradoss SD, Zhu Y, Timmers EM, Zhang Y, Zhao H, Lou J, Wang Y, Joo C, van der Oost J (2017) Autonomous generation and loading of DNA guides by bacterial argonaute. Mol Cell 65(6):985–998. doi:10.1016/j.molcel.2017.01.033

Vogel J, Luisi BF (2011) Hfq and its constellation of RNA. Nat Rev Microbiol 9(8):578–589. doi:10.1038/nrmicro2615

Wang J, Li J, Zhao H, Sheng G, Wang M, Yin M, Wang Y (2015) Structural and mechanistic basis of PAM-dependent spacer acquisition in CRISPR-Cas systems. Cell 163(4):840–853. doi:10.1016/j.cell.2015.10.008

Yoganand KN, Sivathanu R, Nimkar S, Anand B (2017) Asymmetric positioning of Cas1-2 complex and integration host factor induced DNA bending guide the unidirectional homing of protospacer in CRISPR-Cas type I-E system. Nucl Acids Res 45(1):367–381. doi:10.1093/nar/gkw1151

Yosef I, Goren MG, Qimron U (2012) Proteins and DNA elements essential for the CRISPR adaptation process in Escherichia coli. Nucl Acids Res 40(12):5569–5576. doi:10.1093/nar/gks216

Zhang J, Rouillon C, Kerou M, Reeks J, Brugger K, Graham S, Reimann J, Cannone G, Liu H, Albers SV, Naismith JH, Spagnolo L, White MF (2012) Structure and mechanism of the CMR complex for CRISPR-mediated antiviral immunity. Mol Cell 45(3):303–313. doi:10.1016/j.molcel.2011.12.013

How the Other Half Lives: CRISPR-Cas's Influence on Bacteriophages

Melia E. Bonomo and Michael W. Deem

Abstract CRISPR-Cas is a genetic adaptive immune system unique to prokaryotic cells used to combat phage and plasmid threats. The host cell adapts by incorporating DNA sequences from invading phages or plasmids into its CRISPR locus as spacers. These spacers are expressed as mobile surveillance RNAs that direct CRISPR-associated (Cas) proteins to protect against subsequent attack by the same phages or plasmids. The threat from mobile genetic elements inevitably shapes the CRISPR loci of archaea and bacteria, and simultaneously the CRISPR-Cas immune system drives evolution of these invaders. Here, we highlight our recent work, as well as that of others, that seeks to understand phage mechanisms of CRISPR-Cas evasion and conditions for population coexistence of phages with CRISPR-protected prokaryotes.

1 Introduction

Uncovering the structure, function, and potential applications of the prokaryotic CRISPR-Cas locus has been a growing research interest over the past 30 years (Lander 2016). These loci contain a special family of **c**lustered **r**egularly **i**nterspaced **s**hort **p**alindromic **r**epeats (CRISPR) and a unique group of **C**RISPR-**as**sociated (*cas*) genes encoding Cas proteins. The 30-bp intervening sequences called "spacers" are of extrachromosomal origin and correspond to bacteriophage and plasmid genes, many of which are essential to infection or plasmid transference (Mojica et al. 2005; Pourcel et al. 2005). Early discoveries from genomic sequence analyses, including the negative correlation found between the number of CRISPR

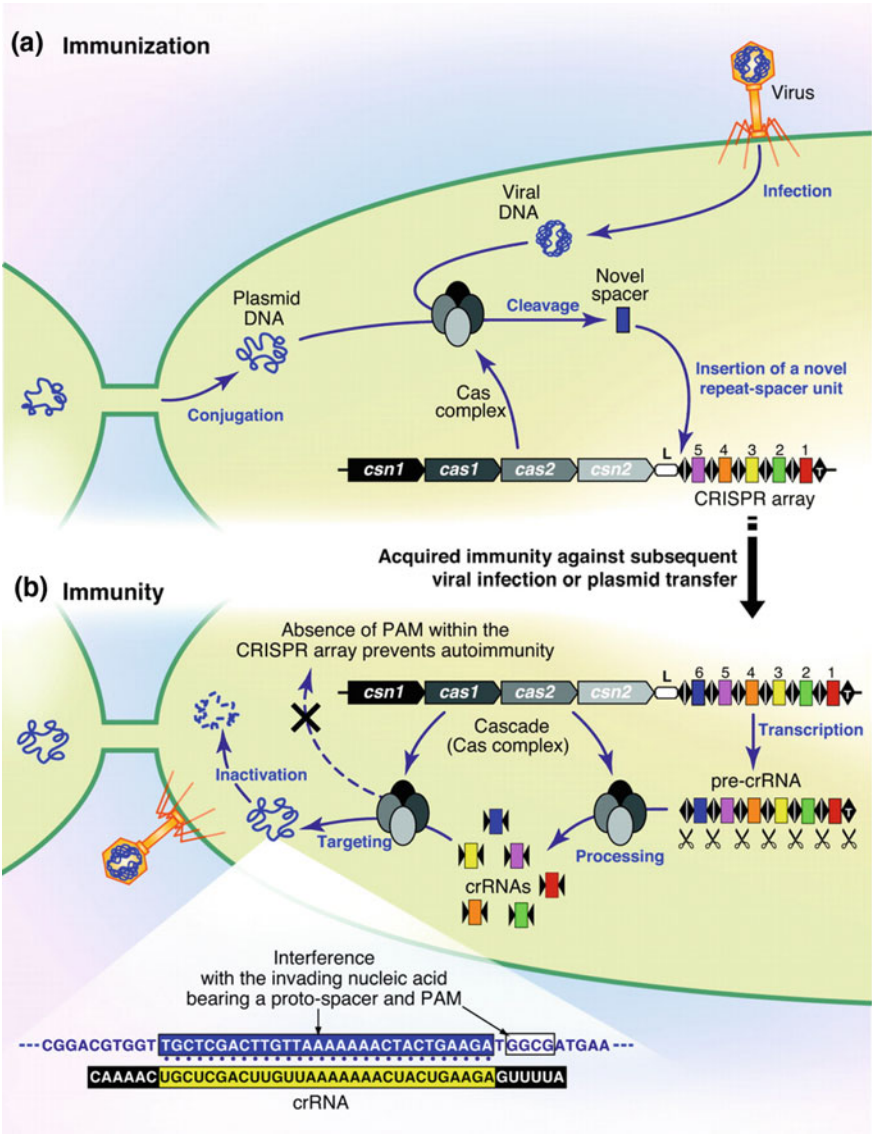
M.E. Bonomo · M.W. Deem (✉)
Department of Physics and Astronomy, Rice University, Houston, TX 77005, USA
e-mail: mwdeem@rice.edu

M.W. Deem
Department of Bioengineering, Rice University, Houston, TX 77005, USA

M.E. Bonomo · M.W. Deem
Center for Theoretical Biological Physics, Rice University, Houston, TX 77005, USA

P. Pontarotti (ed.), *Evolutionary Biology: Self/Nonself Evolution, Species and Complex Traits Evolution, Methods and Concepts*,
DOI 10.1007/978-3-319-61569-1_4

spacers in *Streptococcus thermophilus* and the strain's sensitivity to phage infection (Bolotin et al. 2005) and the lack of CRISPR loci in unthreatened laboratory strains, led researchers to postulate that these elements constituted a genetic adaptive immune system shaped by the host's immediate environment (Lillestøl et al. 2006). Soon after, CRISPR-mediated phage resistance by the integration of spacers, as well as the loss of resistance following the deletion of these crucial spacers, was experimentally demonstrated (Barrangou et al. 2007; Sorek et al. 2008).



◂**Fig. 1** Prokaryotic CRISPR-Cas defense cycles through three stages of adaptation, expression, and interference that are mediated by Cas proteins unique to the CRISPR locus (see also Fig. 2). **a** During adaptation, the CRISPR-Cas system incorporates protospacer sequences from previous invaders into its locus as spacers. A new repeat is copied as the spacer is inserted directly downstream from the leader sequence. **b** During expression, a crRNA guide is created. Depending on the type of CRISPR system, the crRNA is anchored either to one Cas protein or to a multi-component Cas protein complex. During interference, the Cas protein(s) surveil mobile genetic elements that enter the cell and specifically cut sequences that match the crRNA to inhibit infection and replication. There is experimental evidence of both DNA and RNA targeting, depending on the type of CRISPR-Cas system. Reprinted with permission from Horvath and Barrangou (2010)

Though there is a vast variety of these systems (Makarova et al. 2015), the general mechanisms of CRISPR-Cas can be divided into three stages of adaptation, expression, and interference, as seen in Fig. 1. Biochemical and structural analyses have investigated the molecular mechanisms and conformational changes of the Cas proteins associated with each of these stages (Sorek et al. 2013). The host cell combats phage and plasmid threats in its environment by encoding spacers into its genome from one or multiple DNA sequences of previous invaders, called protospacers. New spacers are incorporated directly downstream of an AT-rich "leader" sequence, which characteristically flanks the start of the locus, and older spacers may be deleted at random. The sequential ordering of spacer acquisition provides chronological information about the order in which a cell encountered each phage or plasmid. Each spacer is then expressed as a mobile surveillance **CR**ISPR **RNA** (guide crRNA) that contains a single spacer and a partial repeat sequence on one or both sides. In some CRISRP types, an additional **tra**ns-activating **CR**ISPR **RNA** (tracrRNA) is needed to anchor the crRNA to the Cas surveillance protein. The guide crRNA directs these Cas proteins to interfere with subsequent threats by targeting and specifically cleaving the invading DNA sequences, or in less common cases RNA sequences, that match those of the spacers. Specificity requirements for the recognition of targets vary among CRISPRs. Some require a perfect match between the guide crRNA and the target DNA sequences, while others can tolerate a certain number of mismatches if, for example, the invader has undergone a point mutation. CRISPR-Cas systems that utilize a **p**rotospacer **a**djacent **m**otif (PAM) to distinguish between self and target genomes generally cannot tolerate mutations in this motif region (Bhaya et al. 2011).

Most of the archaea and about half of the bacteria that have been sequenced contain functioning CRISPR-Cas systems. An evolving CRISPR-Cas classification system (Makarova et al. 2015) currently organizes these systems into two overarching classes, for those that utilize a single interference protein versus those that use multiple Cas protein units to survey and cut the target. The systems are further delineated into six major types based on their principle *cas* gene and more than 16 subtypes defined by their Cas protein content. A single organism could have multiple types of CRISPR loci. Additionally, there are still a number of rare, unclassified systems. Figure 2 shows a representative example of the Cas protein content within the Type I CRISPR.

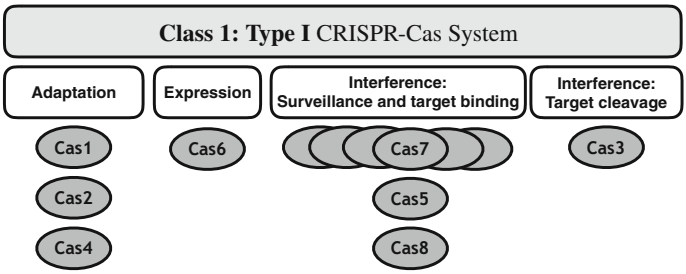


Fig. 2 Each CRISPR-Cas system has a diverse set of Cas protein machinery. For example, in the Class 1, Type I systems, Cas1, Cas2, and Cas4 are used to acquire spacers; Cas6 processes these spacers into crRNA; a complex of multiple Cas protein subunits is used to surveil the target sequence; and Cas3 is recruited for target cleavage. The seven subtypes for Type I are I-A, I-B, I-C, I-D, I-E, I-F, and I-U. See Makarova et al. 2015, for the complete Cas protein content of other CRISPR systems (Makarova et al. 2015)

An extensive genomic analysis of the CRISPR repeats, spacers, leader sequences, and *cas* genes in lactic acid bacteria genomes revealed the likelihood of CRISPR locus acquisition through horizontal gene transfer (HGT) between distant organisms (Horvath et al. 2009). Interestingly, further HGT in a CRISPR-Cas-protected genome appears to be blocked, explaining the lack of loci in antibiotic-resistant and lysogenic bacteria (Marraffini and Sontheimer 2010). Due to the polarized spacer acquisition that causes the ancestral end to contain phylogenetic anchors and the active end to contain recent encounters, the locus can be used to reconstruct the history of strain divergence (Briner and Barrangou 2016). Additionally, CRISPR immunity has been shown to facilitate speciation within the *Streptococcus*, *Staphylococcus*, *Lactobacillus*, and *Bifidobacterium* genera (Briner and Barrangou 2016). The loss of CRISPR-Cas in some strains within a given species allows those strains to acquire virulence via HGT, eventually leading to the emergence of a new pathogenic species.

In exploring how CRISPR-Cas systems could be manipulated, a locus was transferred from one organism into a distantly related one to confer protection against specific plasmids and phage infections (Sapranauskas et al. 2011). One of the more drastic transfers was an oral bacterium's RNA-targeting Type VI-A system that was successfully introduced into *Escherichia coli*, which naturally contains DNA-targeting CRISPRs, to defend the cell from an RNA bacteriophage (Abudayyeh et al. 2016). Following this initial success, immunization of dairy industry-relevant prokaryotes was carried out to establish resistance to anticipated phage attacks (Al-Attar et al. 2011; Mahony and van Sinderen 2015). A turning point came in the applications side of the field when researchers realized the possibility of harnessing the CRISPR-Cas system to make specific genomic modifications in both prokaryotic and eukaryotic cells (Sorek et al. 2013; Pennisi 2013). The Cas9 protein from Type II systems can be re-programed with a single, custom guide sequence to make specific genomic cuts for sequence insertions or deletions.

Catalytically **d**eactivated Cas9 (dCas9) can furthermore be fused to a promoter or represser to respectively **a**ctivate (CRISPRa) or **i**nterfere (CRISPRi) with targeted genes (Peters et al. 2015) and facilitate epigenetic studies (Hsu et al. 2014).

The coevolution of CRISPR-Cas-containing bacteria with plasmids and virulent phages creates what many call a coevolutionary "arms race." Mathematical models that have been developed to investigate this coevolution either take a mean-field approach to model the rate of change of population abundances, usually of wild-type and mutant phages and sensitive and immune bacteria, or look on a more detailed level at phage and bacterial strains represented as arrays of protospacers and spacers. Generally, the degree of immunity, heritability, and benefit of maintaining the CRISPR-Cas system are studied as functions of the number and content of spacers, the abundance and diversity of phages and hosts, and CRISPR-associated fitness costs, such as autoimmunity and the restriction of HGT. Koonin and Wolf (2015), provides a detailed review of phage-host evolution models (Koonin and Wolf 2015). While the threat from mobile genetic elements inevitably shapes the CRISPR loci of archaea and bacteria, it is equally interesting to focus on the evolution of these invaders as they respond to the CRISPR-Cas system. Here, we highlight our recent work, as well as that of others, that seeks to understand phage mechanisms of CRISPR-Cas evasion and conditions for population coexistence of phages with CRISPR-protected prokaryotes. We begin with a look at the nature of the host cell's defense that puts pressure on phages to diversify. Then, we describe the theoretical conditions and advantages of phage protospacer evolution, followed by experimental observations of this as well as observations of novel phage counterattack mechanisms. We end with a couple of representative clever applications that utilize the phage-CRISPR host interaction.

2 Targeting of Phages by CRISPR

2.1 CRISPR Spacer Content

The CRISPR spacer content provides a record of the phages to which bacteria have been exposed, as viewed through the lens of selection. Experiments with *S. thermophilus* (Deveau et al. 2008) and *Leptospirillum* (Tyson and Banfield 2008) have shown that the diversity of CRISPR spacers in a population of bacteria decreases with distance from the leader. Many subsequent studies confirmed these initial observations. However, some studies showed a more uniform dependence of diversity with distance from the leader.

In one of the first theoretical studies of the CRISPR system, we sought to explain these observations using a population dynamics model (He and Deem 2010). Each bacterium had a CRISPR locus of a finite length, with the oldest spacer dropped when the number of spacers exceeded 30 per locus. The CRISPR locus was copied to daughter cells after bacterial division. We found that the diversity of the spacers

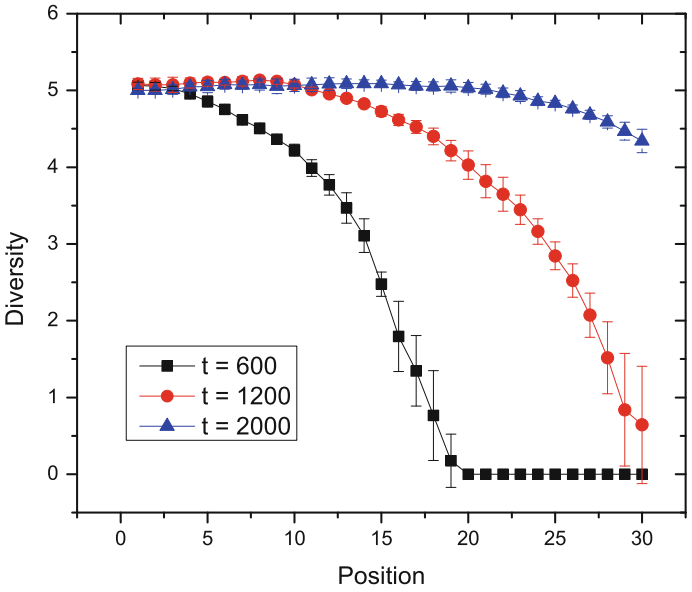


Fig. 3 Theoretical results for the diversity of spacers in the CRISPR locus as a function of distance from the leader sequence. The leader-proximal spacers are more diverse than the leader-distal spacers as the CRISPR samples a new environment. After a long time in a stable environment, the diversity of spacers becomes constant along the locus, a function of the relatively constant diversity of phages in the environment. Reprinted with permission from Han et al. (2013)

decreased with distance from the leader. Spacers leading to resistance against the dominant phage were especially selected for and accumulated in the CRISPR array.

In a second model, we sought to explain the time dependence of this decay of diversity with distance from the leader (Han et al. 2013). Again, we found that spacer diversity decreased toward the leader-distal end due to selection pressure on shorter timescales, as shown in Fig. 3. On longer timescales, we found that spacer diversity was nearly constant with distance from the leader. Thus, spacer diversity decays more rapidly when bacteria are exposed to new phages, either through bacterial migration or phage influx. These results offer one explanation for the two differing experimental observations of spacer diversity.

2.2 Gain or Loss of Immunity

Immunity to phages that CRISPR confers upon bacteria is not perpetual. Changes in the phage population lead to abrogation of the protection afforded by the CRISPR spacers. Defining the spacer effectiveness as the match between a spacer and the phage strains present in a population, we found that spacer effectiveness decreases toward the leader-distal end as well (Han et al. 2013).

While the mechanism by which protospacers from the phages are inserted as spacers into the bacterial CRISPR array adjacent to the leader is known, the mechanism by which spacers are deleted is less clear. We investigated whether the results for spacer diversity and immunity were persistent with changes to the mechanism of spacer deletion (Han et al. 2013). The results for spacer diversity and immunity were relatively insensitive to whether the oldest spacer was deleted, one of the older spacers was deleted with increasing probability toward the leader-distal end, or a random spacer was deleted from anywhere in the locus. This insensitivity to deletion mechanism results because selection provides a strong bias for successful deletion of the old spacers that no longer match actively infecting phage.

Loss of immunity can lead to oscillations in the population size of bacteria and phage. This phenomenon was investigated in a minimal, Lotka–Volterra type predator–prey model of a host with a heritable, adaptable immune system, *e.g.*, a CRISPR-Cas system (Berezovskaya et al. 2014). When the immunity decay rate is larger than the immunity acquisition rate, periodic oscillations of the populations of immune hosts, sensitive hosts, and phages become larger and lead to quasi-chaotic behavior. A similar behavior is also observed for the case in which the immunity acquisition rate is greater than the immunity decay rate; however, the fraction of immune hosts is larger here. There were critical values of the phage reproduction rate separating the phases of stable equilibria, small periodic oscillations, and quasi-chaotic oscillations.

When the rate of spacer deletion is small, the phase diagram no longer follows the predictions of the classical mean-field, predator–prey model. The phage extinction probability during exposure to CRISPR-bearing bacteria becomes non-classical and reentrant (Han and Deem 2017). Parameters affecting the phase diagram include rates of CRISPR acquisition and spacer deletion, rates of phage mutation and recombination, bacterial exposure rate, and multiple phage protospacers. The new, non-classical region appeared at a low rate of spacer deletion, as seen in Fig. 4. The population of phages progressed through three extinction phases and two abundance phases, as a function of bacterial exposure rate.

2.3 CRISPR Locus Length and Phage Diversity

The number of spacers in the CRISPR locus and the phage diversity are critical parameters affecting the bacteria and phage coevolution. In Levin et al. (2013), a well-defined, simple system was studied experimentally. Bacteria immune to a single type of phage via a single spacer were observed to be eventually invaded by phages. The single spacer caused incomplete resistance because of a high rate of CRISPR escape mutations. That is, the bacteria were invaded by phages that had made single mutations in their protospacer regions. Conversely, the CRISPR-Cas efficacy is predicted to increase rapidly with number of protospacers per phage genome (Iranzo et al. 2013).

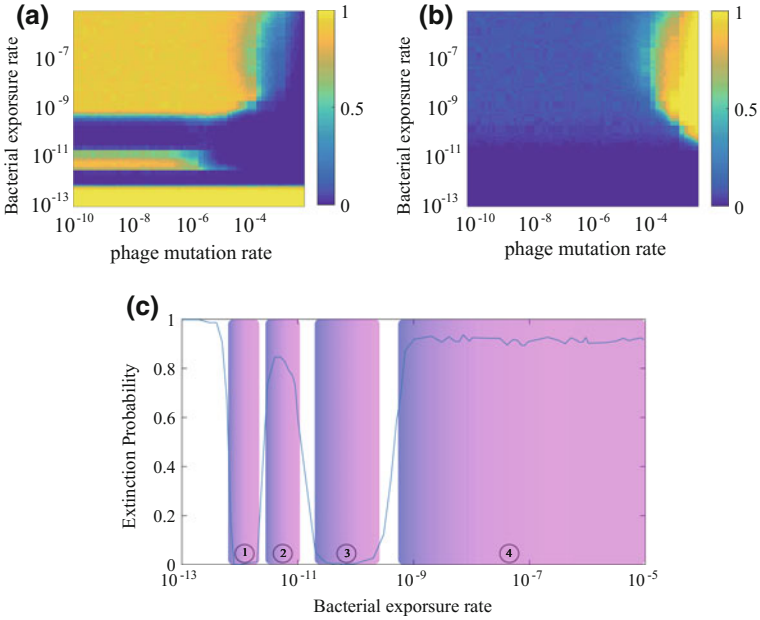


Fig. 4 Non-classical phase diagrams of **a** the phage extinction rate and **b** the bacterial extinction rate resulting from a coevolutionary model. These complex patterns of phage-bacteria coexistence represent the delicate balance in place among the bacterial exposure rate, phage evasion through mutation, number of available protospacers, and rate of spacer acquisition. **c** A small rate of CRISPR spacer deletion leads to the three observed phases of phage extinction and two phases of phage survival that depend on the rate of bacterial exposure. Reprinted with permission from Han and Deem (2017)

Aspects of the complex phage-bacteria coevolution were also studied theoretically. Protection and immunity can be non-monotonic in time because of the decreasing phage population diversity over time (Han et al. 2013). A stochastic, agent-based mathematical model of coevolution of host and phage shows CRISPR-Cas efficacy is dependent on population size, spacer incorporation efficiency, number of protospacers per phage, phage mutation rate, and fitness cost of maintaining a CRISPR-Cas system (Iranzo et al. 2013). The coevolution of the CRISPR-Cas immune system and lytic phages was modeled under evolutionary and ecological conditions, *i.e.*, coupling of host and phage reproduction and death rates, in which CRISPR-Cas immunity stabilizes phage-host coexistence, rather than extinction of phage. The overall phage diversity was observed to grow due to an increase of host and phage population size, not specifically due to CRISPR-Cas selection pressure on single protospacers. The CRISPR-Cas system was predicted to become ineffective at a certain phage diversity threshold and lost due to the associated fitness cost of maintaining *cas* genes.

Another model similarly showed the evolved average number of spacers in the CRISPR depended on the phage mutation rate and the spacer cost to fitness

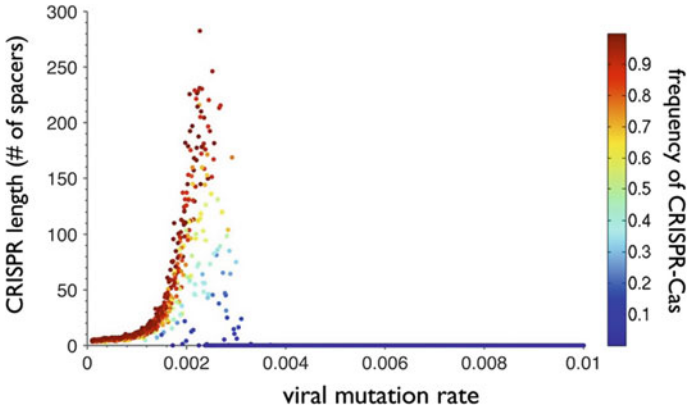


Fig. 5 A mathematical model of how the length and prevalence of the CRISPR array depend on the phage (viral) mutation rate. With low phage mutation rates, CRISPR-Cas systems are highly prevalent and select to retain low numbers of spacers to match the low diversity of phages. As the phage mutation rate increases, CRISPR-Cas systems become less frequent, though those that are present are collecting more spacers to keep up with the diversifying phage population. At a certain phage mutation threshold, the CRISPR locus becomes too long to be effectively maintained and is rapidly lost from the host population. Reprinted with permission from Weinberger et al. (2012)

(Weinberger et al. 2012). At low mutation rates, a limited number of spacers was sufficient to confer protection to the bacteria against the phage population of limited diversity, shown in Fig. 5. As the phage mutation rate increased, the CRISPR loci increased in length. At a critical threshold of phage mutation, the CRISPR array became unable to recognize the diverse phage population, and the average locus length fell rapidly to zero, even if the rate of spacer addition outpaced phage mutation rate. It was speculated that similar behavior would occur from an increasing immigration rate of new phages.

One hypothesis for the greater fraction of hyperthermophiles that have effective CRISPR-Cas systems compared to mesophiles is that the lower rates of mutation and fixation in thermal habitats lead to more effective, and therefore selected for, CRISPR systems in thermophiles. Additionally, another possible mechanism suggested theoretically is that CRISPR becomes ineffective in mesophiles because of larger population sizes (Iranzo et al. 2013).

3 Selection for Mutation and Recombination in the Phage

The bacterial immune system of CRISPR implies a selective advantage for those phages with mutations in the PAM or protospacer regions. That is, a mismatch between the crRNA sequence and PAM or protospacer of invading phage is likely to allow the phage to infect and replicate in the bacteria. Concomitantly, the mechanism of recombination can integrate multiple point mutations, increasing the

chance of a mismatch that would allow the phage to escape CRISPR recognition. In this setting, recombination can be a positive mechanism for generating genomic diversity.

3.1 Coevolutionary Implications

A number of coevolutionary dynamics models have captured the idea that not only can phages evade CRISPR-Cas via mutation or recombination of protospacers, but also bacteria can regain immunity through acquiring more spacers from the same phage. These models are reviewed in Koonin and Wolf (2015). In our own work, we first considered CRISPR arrays with between 2 and 30 spacers, considering the possibility of phage mutation. These refinements supported the main prediction that the diversity of spacers was found to decrease with position from the leader-proximal end (He and Deem 2010).

A combination of mathematical models, population dynamic experiments, and DNA sequence analyses have been used to understand CRISPR-containing-host and phage coevolutionary dynamics in the *S. thermophilus* CRISPR-Cas and virulent phage 2972 model systems (Levin et al. 2013). There was a particular interest in hosts that had gained resistance by the addition of novel spacers and phages that evaded resistance by mutation in their matching sequences. The coevolution between the phage and bacteria was termed an "arms race," perpetuated by the competing effects of spacer acquisition and protospacer mutation.

The effects of recombination depend on the degree of divergence between protospacer and spacer required for the phage to escape CRISPR surveillance. We showed there is little difference between the results from point mutation and those from recombination in the phage genome if the phage needs just one mismatch to escape (Han et al. 2013). However, when the phage needs two mismatches to escape, the difference is apparent in the immunity, the rate at which bacteria are able to kill phages. Recombination, by combining mutations, is a more rapid generator of protospacer diversity and is a more successful phage escape strategy when CRISPR has a higher mismatch tolerance with the protospacers, see Fig. 6. When the phage has multiple protospacers, a similar argument implies that recombination, now of the protospacers rather than of genetic material within a single protospacer, again leads to a more rapid escape of the phage than does point mutation alone. This result occurs because mutation in different protospacers can be recombined, making it substantially less likely for the CRISPR to recognize the recombined daughter phage. Thus, the phage recombination-mediated escape mechanism is also more successful when the phage has multiple protospacers. The immunity afforded by CRISPR is lower as mutation and recombination rates increase.

The interplay between the CRISPR pressure on the phage and the phage pressure on the bacteria leads to a phase diagram of coexistence. That is, only in some parameter regions do the phages and bacteria coexist. The pattern of coexistence is more complicated than the classical predator–prey model, due to the feedback of the

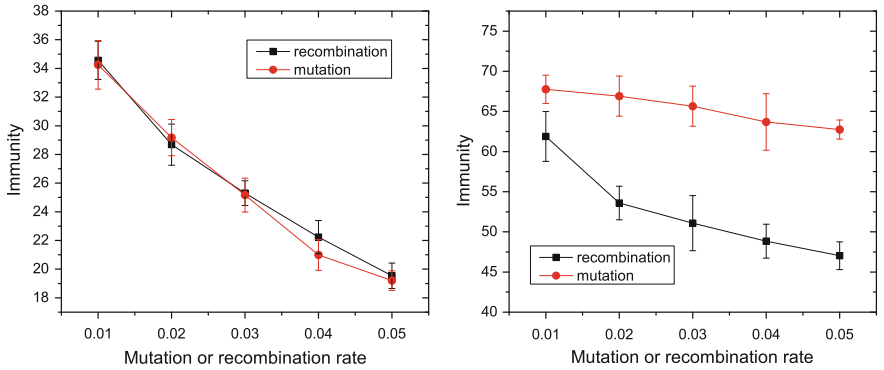


Fig. 6 A mathematical model shows that bacterial CRISPR immunity decreases with increasing phage mutation or recombination rate. When the bacteria's CRISPR-Cas system has a mismatch tolerance of just one nucleotide, there is little difference between the effect of phage mutation and recombination (*left*). However, when there is a higher mismatch tolerance of two nucleotides, recombination gives the phage a higher probability to survive (*right*). Reprinted with permission from Han et al. (2013)

CRISPR system on the phage. A low phage mutation rate can lead to a phage extinction probability that is a non-monotonic function of bacterial exposure rate (Han and Deem 2017). The resulting non-classical phase diagram in Fig. 4 shows the extinction and abundance tipping points that result from a complex relationship between the bacterial exposure rate and the phage mutation rates.

3.2 Constraints on Non-synonymous Mutations

One important difference between thermophiles, with habitats of 42 – 122 °C, and mesophiles, with habitats of 20 – 45 °C, is that protein stability is a more crucial factor in thermophiles. That is, mutations are more likely to be deleterious in thermophiles; because on average, proteins are more easily destabilized by mutation at higher temperatures (Berezovsky and Shakhnovich 2005). A stochastic model of phage-CRISPR coevolution was used to investigate why CRISPR-Cas systems are more prevalent in thermophiles than mesophiles (Weinberger et al. 2012). Mutations are more likely to be lethal in thermophilic environments because high temperatures reduce protein stability; therefore, there is selection for phages that mutate less. It was argued that the reason roughly 90% of archaea, which are mostly thermophiles, have CRISPR-Cas systems is because these CRISPR-Cas systems work well in this habitat. Further support for this hypothesis was that the CRISPR-Cas is more correlated with thermophilic environments than with archaeal taxonomy. This stability criterion was used to compute a phage mutation rate threshold, beyond which phages were selected against.

3.3 Benefits of CRISPR Versus Other Immune Mechanisms

Bacteria have other, innate mechanisms of resistance against phages. For example, the bacterial surface receptors that promote phage attachment and entry can undergo modification. A model was used to study the evolution of CRISPR-Cas positive and negative hosts as they encountered phages (Weinberger et al. 2012). The interaction events were either successful microbial protection against infection or successful phage infection. The CRISPR-Cas positive host could delete or lose the CRISPR system and the CRISPR-Cas negative host could acquire a CRISPR system by HGT. The fitness of the phages increased by productively infecting hosts and creating phage progeny, whereas the fitness of the hosts increased by acquiring CRISPR-Cas and useful spacers. There was a potential fitness cost to the hosts due to autoimmunity of the CRISPR system inhibiting normal bacterial gene function and restriction of potentially beneficial HGT events. The CRISPR-Cas system was found to be beneficial at intermediate levels of the innate resistance. There was little fitness advantage from CRISPR storing spacers of phages that the bacteria were unlikely to encounter again. When the bacteria survived two-thirds of its phage encounters without the help of CRISPR, maintaining the CRISPR system was too costly, *i.e.*, there was no benefit to having it.

3.4 Heterogeneous Environments

Many of the models of the CRISPR system assume a mean-field, homogeneous distribution of the phage and bacteria in space. Spatial effects and heterogenous environments, however, occur in the body and in nature, and these effects can have a significant effect on the outcome. Indeed, experiments carried out with *S. thermophilus* and phage 2972 have shown that a small percentage of acquired spacers matched the closely related phage 2766 that had migrated spatially (Paez-Espino et al. 2015). A mathematical model of bacteria-phage coexistence was used to take into account the effects of space on species coexistence and adaptive CRISPR defense (Haerter et al. 2011). In the model, bacteria and phage populations spread on a two-dimensional square. Parameters of the model included the effective infection rate of microcolony of bacteria, *i.e.*, the probability of infection, and the mean latency time, as a ratio of phage to bacteria mean replication. Two spatial arrangements of phage replication were explored: a well-mixed system, in which phage offspring were placed at random sites, and a slow diffusion model, in which phages only spread to neighboring sites. Bacteria dynamically acquired resistance through CRISPR-Cas to the diverse phage population while removing the oldest spacers. For successful phage survival, *i.e.*, not leading to depletion of bacterial hosts or exceeding available spatial carrying capacity, a balance was needed between the effective infection and phage replication rates. At least two phage strains were needed to allow stable coexistence with bacteria. Coexistence persisted

as long as the maximum number of CRISPR insertions was fewer than the total number of phage types.

Another strain-level model of the origin and diversification of CRISPR arrays in host and protospacers in phages was developed, taking density-dependent ecological dynamics into account (Childs et al. 2012). To understand the coevolution of strain diversities, as well as densities, three main components in the model were specified: coupling among host and phage reproduction and death rates, molecular scale CRISPR events based on sequence matches between spacers and protospacers, and evolutionary changes of phage protospacer mutation and CRISPR spacer acquisition. A maintenance of many coexisting strains in highly diverse communities was observed, with high strain similarity on short timescales but high dissimilarity over long timescales. Short-term changes in host diversity were driven by incomplete sweeps of newly-evolved high-fitness strains in low abundance, recurrence of ancestral strains that gained fitness advantage in low abundance, and invasions of multiple dominant coalitions that arose from having nearly identical immune phenotypes, but different genotypes, *i.e.*, similar protection afforded by the incorporation of different protospacers. A majority of new phage mutants did not have a significant increase in fitness, since mutation was random. In this model, only the first spacers were important to shaping selective coevolution because they provided the highest immunity, and the predicted spacer acquisition rate was more important to diversification than was CRISPR immunity failure.

4 Experimental Evidence for Phage Evasion of CRISPR

4.1 Synonymous Mutations

The evasion of CRISPR-Cas by phage evolution has been explored in a number of studies. The CRISPR-Cas system creates an evolutionary battle between phage and bacteria through addition or deletion of spacers in the bacteria and mutations or deletions in the phage genomes, as reviewed in Deveau et al. (2010). There is evidence for phage evasion: a small population of virulent phage mutants was observed to infect previously bacteriophage-insensitive bacteria. These infecting phages had single nucleotide changes or deletions within their targeted protospacer. Indeed, the CRISPR locus is subject to dynamic and rapid evolutionary changes driven by phage exposure, and therefore CRISPR spacers can be used to analyze past host-phage interactions.

A study of the lactic acid bacteria *S. thermophilus* and the role of its CRISPR1 locus in phage-host interactions was carried out by selecting two different bacteriophage-insensitive strains and exposing them to other phages to which they were sensitive (Deveau et al. 2008). The addition of one new spacer of about 30 nucleotides in CRISPR1 was the most frequent outcome of a phage challenge. Spacers were only acquired from protospacers with an AGAAW motif 2

nucleotides downstream from the protospacer. There was also evidence of spacer deletion. Successive phage challenges and subsequent addition of spacers increased the overall resistance of the host to the phage. Newly added spacers were required to be identical to the protospacer region in the phage genome to confer resistance to phage. Indeed, phages were able to evade CRISPR immunity through single nucleotide mutations and deletions. The most common mechanism of escape was mutation within the protospacers or AGAA flanking sequences.

Natural hot springs provide a rich source of microbial and phage diversity. A metagenomic analysis of the Yellowstone hot springs to study natural evolution of microbial and phage populations due to the CRISPR immune system was carried out (Heidelberg et al. 2009). Two thermophilic *Synechococcus* bacteria isolates were sequenced from microbial mats in the Octopus and Mushroom Springs in Yellowstone National Park, *Syn* OS-A from high-temperature areas and *Syn* OS-B' from low-temperature areas, to make comparisons with phage and prokaryotic metagenomic data collected from the same springs. The *Syn* OS-B' genome contained individuals from two types of CRISPR loci, while *Syn* OS-A contained individuals from three types, with these types distinguished by their repeat sequence. The Type III repeat sequence identified in *Syn* OS-A, but not *Syn* OS-B', was also present in another abundant microbe in the mat, *Roseiflexus* RS-1, suggesting recent DNA transfer between them. While CRISPR repeats were of course highly conserved within the microbial metagenome, spacer sequences were quite unique, and from the CRISPR sequence data, it was difficult to find matches to the phage metagenomic data. Nonetheless, several spacers matched lysozyme and lysin protein genes. These lysozyme enzymes attack the cell wall late in phage infection, causing cell lysis and release of phages. There were some silent or conservative mutations found in these phage lysozyme or lysin protein sequences that did not affect protein function, but most likely helped the phage to evade CRISPR identification.

4.2 Recombination

The phage genomic regions targeted by CRISPR are selected to have substitutions (Paez-Espino et al. 2015). Homologous recombination events between genetically related phages can then further diversify the phage population. Long-term coevolution experimental studies were carried out with *S. thermophilus* and phage 2972 for up to 232 days until the phage went extinct. An analysis of the *S. thermophilus* spacers revealed that spacers acquired during the experiments mapped unevenly over the phage genome. The immune pressure from CRISPR drove escape mutations located exclusively in the protospacer and PAM regions, as well as the accumulation of phage genome rearrangements. Phage mutation rates were much higher than that of the bacterial host, and the presence of phages also accelerated host genome evolution in the CRISPR array. The coexistence of multiple phages

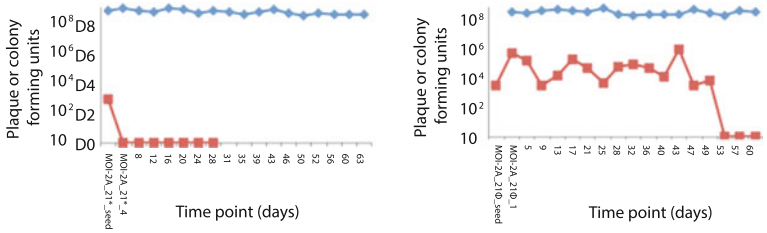


Fig. 7 CRISPR more effectively eradicates less diverse phage populations. Experimental evidence of the ability of two phage types to coexist with CRISPR bacteria (*right*) for a longer period of time than a single phage type could (*left*). The host cell population (CFU per milliliter) is shown in *blue*, and phage counts (PFU per milliliter) are shown in *red*. Reprinted with permission from Paez-Espino et al. (2015)

allowed recombination events that boosted the observed substitution rates beyond the bare mutation rate, termed the "rescue" effect, see Fig. 7.

Experiments show that the most recently acquired spacers match coexisting phages (Andersson and Banfield 2008). In other words, the phage population evolves, abrogating the utility of old spacers. Phages use extensive recombination to shuffle sequence motifs and evade the CRISPR spacers. Spacers of CRISPR loci recovered from *Leptospirillum* group III, I-plasma, E-plasma, A-plasma, and G-plasma microbial communities in biofilms collected from Richmond Mine (Redding, CA) were used to link phages to their coexisting host bacteria and archaea. A majority of spacers corresponded to phages, though some corresponded to other mobile genetic elements, such as plasmids and transposons. It was found that microbial cells targeted several different phage populations, and only a few CRISPR spacers were widely shared among bacterial strains. For example, some E-plasma cells target specific AMDV2, AMDV3, and AMDV4 phage variants, and some spacers match dominant phage sequence types. Yet many spacers match sequences characteristic of only one or a small number of genotypes within the population. The phage population is reshaped by extensive homologous recombination, as evident from combinatorial mixtures of small sequence motifs. This recombination resulted in genetic blocks shared by different phage individuals that were often no more than 25 nucleotides in length, creating new phage sequences that can disrupt the function of the CRISPR's 28 to 54-nucleotide spacers. A benefit of recombination is that if the preexisting sequence diversity in the phage population is mainly in the protospacer and PAM regions, recombination can increase this diversity in a combinatorial way without introducing novel diversity outside these regions. In this fashion, recombination creates new DNA sequences with a lower risk of altering protein function than does mutation and limits alterations to the phage genome outside of the CRISPR-recognized PAMs and protospacers.

4.3 Mutations in PAM, Seed, and Non-seed Regions Are Distinct

The original view that a single mutation in the protospacer region is sufficient to allow phage to escape from CRISPR has transformed into a more nuanced view of CRISPR recognition being quite sensitive to mutations in a seed region, which is usually defined as the protospacer's eight PAM-proximal nucleotides (Sorek et al. 2013), but less sensitive to mutations in the rest of the protospacer region. An experiment with a co-culture of *S. thermophilus* with phage 2972 for one week studied the impact of spacer acquisition and host population diversification on phage genome evolution (Sun et al. 2013). Tracking of CRISPR diversification and host-phage coevolution revealed a strong selective advantage for phages containing PAM or near PAM mutations. A genetically diverse bacterial population arose, with multiple subdominant strain lineages. All surviving *S. thermophilus* cells had at least one newly incorporated spacer against phage 2972. Of the two loci sequenced, CRISPR1 was the most active, and CRISPR3 only incorporated single spacers. In this experiment, all recovered phages contained a synonymous mutation eight nucleotides from the PAM, apparently leading to escape from *S. thermophilus* spacer1 in the CRISPR1 loci. The fixation of this mutation suggests that only phage that had it could rise in abundance. This phage sequence encodes a protein that recognizes the host, and there is strong selective pressure for the host to abrogate this recognition. Also in this experiment, 88% of recovered phage sequences contained synonymous mutations six nucleotides from PAM, likely leading to escape from spacer32. Finally, 92% of phage sequences contained non-synonymous mutations in the PAM corresponding to spacer6. There were no observed phage mutations in the regions targeted by any CRISPR3 spacers.

4.4 Long-Term Study of Heterogeneous Environments

A long-term metagenomic study of archaeal, bacteria, and phage populations in Lake Tyrrell (LT) from 2007 to 2010 was carried out, and the dynamics of their populations on timescales of months to years was studied (Emerson et al. 2013). Samples were collected and analyzed from LT over three summers and four winters. Overall, archaeal and bacterial populations were more stable than the phage populations, *i.e.*, over the timescale of years, phage populations were less stable than their prokaryotic prey. Analysis of CRISPR arrays indicated both rare and abundant phages were targeted, suggesting archaeal hosts attempt to balance protecting themselves against persistent, low-abundance phages and highly abundant phages that could destroy the host community. There was a high diversity of phages in the environment, and even in the absence of superinfection, the CRISPR array sampled extensively from this diversity.

4.5 Plasmid Evasion of CRISPR-Cas

While plasmids and phage bear a resemblance, interestingly the selection pressure on plasmids from CRISPR appears somewhat different from that on the phage. Namely, the rate of plasmid mutation is much slower than combined rate of loss of CRISPR immunity by spontaneous mutation or deletion. On the one hand, plasmids can confer positive features upon bacteria, such as antibiotic resistance. Analysis via experiments and computer modeling of the loss of CRISPR-Cas loci in the presence of an environment containing plasmids that increase the host's fitness has been carried out (Jiang et al. 2013). Conjugational transfer of the Staphylococcal plasmid pG0400 (*nickase* gene *nes*) into *Staphylococcus epidermidis*, which contained a spacer targeting this plasmid, was analyzed. Simulation results showed plasmid transfer into the host could occur if the plasmid mutated, the CRISPR lost the associated spacer, the CRISPR locus became deactivated or deleted, or the CRISPR response was subdued. Experiments showed the wild-type plasmid on CRIPSR-negative mutants only, meaning that instead of utilizing the phage mechanism of mutating their targeted regions to evade CRISPR-Cas, a plasmid "evasion" strategy could occur within the host, *i.e.*, loss of CRISPR locus allowed the host to receive the beneficial plasmid. In vitro experiments showed little to no intrinsic fitness cost of losing CRISPR.

5 Emergence of Game Theoretic Strategies in the Phage

5.1 Anti-CRISPR Proteins

The phages evade CRISPR by more than just mutation and recombination. Phages have evolved more complicated mechanisms that may be understood from a game theoretic point of view. Anti-CRISPR proteins have been identified in phage that is associated with inactivating Type I-E and I-F CRISPR-Cas systems (Maxwell 2016). These anti-CRISPR proteins, discovered to be encoded by *Pseudomonas aeruginosa* phages, circumvent CRISPR by inactivating the Cas proteins. Five distinct families of proteins that targeted Type I-F and four that targeted Type I-E were found. The existence of subsets of these genes among phages suggests HGT may have been responsible for a "mix and match" scenario of acquiring them. The mechanisms of action of three unique I-F interference inhibitors, AcrF1, AcrF2, and AcrF3, are illustrated in Fig. 8. Briefly, in Type I-F CRISPR-Cas systems, a Csy complex is guided by crRNA to bind to invading DNA, and Cas3 is recruited for target cleavage. AcrF1 binds along the full Csy3, which comprises three molecules in the Csy complex, but allosterically interferes with DNA binding. AcrF2 binds to the Csy1-Csy2 heterodimer in the Csy complex to block the 5' end of crRNA and directly prevents DNA binding. AcrF3 interacts with Cas3 to block its recruitment to the Csy complex. These mechanisms could potentially regulate lateral gene

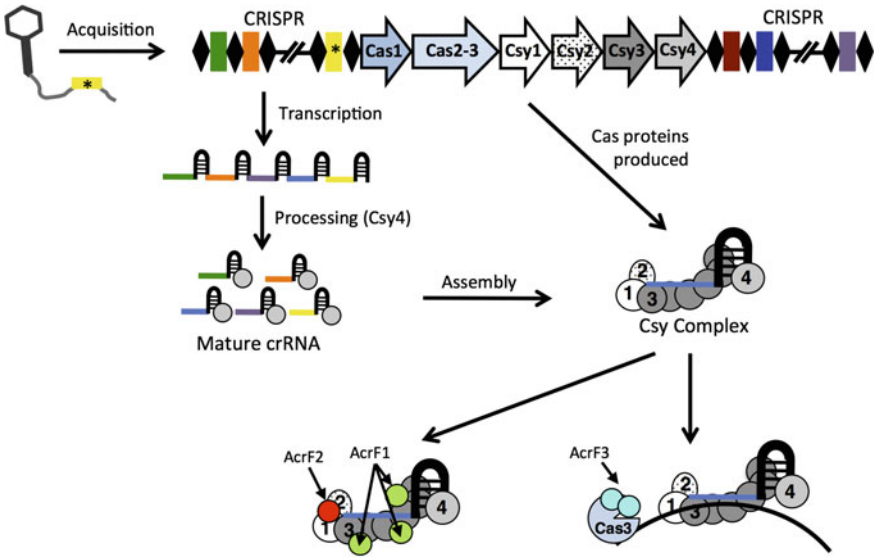


Fig. 8 Depiction of mechanisms used by three different anti-CRISPR proteins that target the Type I-F CRISPR system, which contains a crRNA: Csy surveillance complex and Cas3 cleavage protein. Anti-CRISPR proteins AcrF1 and AcrF2 attach directly to the Csy complex to block the crRNA from fully binding with the target DNA. Protein AcrF3 attaches to the Cas3 protein to prevent it from being recruited to cleave a bound target. Reprinted with permission from Maxwell (2016)

transfer to allow foreign DNA to bypass CRISPR-Cas inhibition. In a separate experiment, it was also found that another phage produced enzymes upon infection that induced a phenotypic phage resistance in sensitive bacteria, but killed bacteria with CRISPR, even in the presence of many phages (Levin et al. 2013).

5.2 *Phage Encoding Its Own CRISPR*

Amazingly, CRISPR elements have even been found in prophages and phage DNA segments. Whole-genome microarray analysis revealed that the *Clostridium difficile* genome contained mobile genetic elements, such as prophages, which contained CRISPR loci (Sebaihia et al. 2006). The complete genome sequence was determined for strain 630 of *C. difficile*, which is virulent and multidrug-resistant. An unusually large fraction of the genome, 11%, consists of mobile genetic elements (MGEs), including two prophages and a prophage-like element, responsible for acquisition of genes involved in antimicrobial resistance and virulence. Ten CRISPR DNA repeat regions were identified with no evidence of their expression or function. Interestingly, several were located on the prophages and

prophage-like element, suggesting the phages had incorporated a CRISPR with spacers against other phages.

Twenty-two CRISPR arrays were found in the phage sequences from a metagenomic study of the genetic composition of the phage population in the human gut microbiome (Minot et al. 2011). Simultaneously isolated at two time points, one of the phage's spacer sequences matched the sequence of another phage present in the same individual, suggesting a phage-phage competition mediated by CRISPR. Additionally, there was an array that showed greater than 95% identity in the repeat regions to the previously found CRISPR in *C. difficile* (Sebaihia et al. 2006), described above.

One bacteriophage has been observed to directly combat a host bacteria's immunity by using a CRISPR-Cas system (Seed et al. 2013). The bacterial *Vibrio cholerae* serogroup O1, which causes cholera, can be treated with the **I**nternational **C**entre for Diarrhoeal Disease Research, Bangladesh cholera phage 1 (ICphage1). The currently active *V. cholerae* strain "El Tor" does not contain a CRISPR-Cas system; however, it encodes an 18-kb **p**hage-inducible chromosomal island-**l**ike **e**lement (PLE). The phage-inducible chromosomal island is a highly mobile genetic element that contributes to HGT, host adaptation, and virulence, by using phages as helpers to promote the host's spread while simultaneously preventing these helper phages from reproducing. *V. cholerae* therefore uses its PLE to interfere with the ICphage1 reproductive cycle and increase its own virulence. As an evolutionary counterattack, the ICphage1 contains a CRISPR-Cas system, comprised of two loci and six *cas* genes, that actively targets the bacteria's PLE. A single spacer that targets the PLE is sufficient to allow the ICphage1 to destroy the PLE and then replicate successfully. It is unclear how the ICphage1 evolved to have this system; however, a comparison with existing characterized CRISPR systems reveals a high similarity to Type I-F.

6 Discussion

6.1 An Example Application in Biotechnology

The CRISPR system has powerful applications in the molecular biology and genomic editing fields. In the latter, CRISPR has been used to delete, add, activate or suppress targeted genes in many species, including human cells, resulting in the so-called "CRISPR craze" (Pennisi 2013). Self-targeting applications, to target and cleave the host chromosome, have been suggested. These include antimicrobial selection for specific microbial populations within a mixture, antibiotic-resistant gene targeting, and large-scale genomic deletion (Briner and Barrangou 2016; He and Deem 2010). Targeting of the uptake of external genetic material to inhibit the spread of resistance genes by HGT has also been suggested (He and Deem 2010). Here, we mention one successful experimental implementation of using

phage-delivered CRISPR-Cas to control antibiotic resistance and horizontal gene transfer (Yosef et al. 2015). In this study, temperate lambda phages were used to deliver a CRISPR-Cas system into *E. coli*. The locus was programmed to destroy antibiotic resistance-conferring plasmids and protect against lytic phages. Following its delivery, lytic phages were used to kill antibiotic-resistant bacteria. The combined result was to leave only antibiotic-sensitive bacteria, which could later be killed via antibiotics. This method selected for antibiotic-sensitized bacteria by creating and delivering to the bacteria a CRISPR-Cas system that contained spacers which would first self-target the antibiotic-resistant genes and then protect the bacteria from lytic phages engineered with matching protospacers. It was found that HGT of antibiotic-resistant elements was no longer possible to *E. coli* containing the system. Additionally, these antibiotic-sensitive bacteria were resistant to the engineered lytic phages. Importantly, the construction allowed for the selection of these bacteria and selection against antibiotic-resistant ones. This bacteria-sensitizing method delivered a CRISPR-Cas system to phages that then delivered this system to bacteria, which simulated relevant external environments such as a hospital fomites or the human skin. An important advantage, therefore, is that this approach would not require delivery of the CRISPR-Cas system to human host cells.

6.2 An Example Application in Microbiome Modification

Study of the microbiome is a burgeoning field, with implications ranging from health and disease to learning and mood. The bacterial flora in the gut is heavily influenced by the 15 $\times$ greater number of viruses there (Howe et al. 2015), 90% of which are bacteriophages (Scarpellini et al. 2015). A study of the genetic composition of the phage population in the human gut and their dynamic evolution in response to environmental perturbations was carried out (Minot et al. 2011). Metagenomic sequencing of the human virome from subjects on a dietary intervention of a high-fat and low fiber, a low-fat and high-fiber, or an ad-lib diet was performed on samples collected over 8 days to understand the structure and dynamics of the phage population. The controlled feeding regimen caused the virome to change and converge among individual subjects on the same diet. The sequencing identified many DNA segments responsible for phage functions required in lytic and lysogenic growth, as well as antibiotic resistance. Interestingly, the study found phages encoded CRISPR against other phages, as reported in the previous section. In particular, CRISPR arrays in *Bacteriodetes* from the gut contained spacers matching genetic material from the virome sequenced in the same individuals, suggesting a functional CRISPR existed.

7 Conclusion

CRISPR-Cas is now a major topic in applied molecular biology and genomic editing. It first rose to prominence, however, as a novel immune defense mechanism of bacteria against phages. Theory and modeling have shed light on the dynamics of CRISPR spacer acquisition, how the diversity of the phage population is reflected in the content of the spacers, and the potentially complicated patterns of coevolution that bacteria and phages exhibit. CRISPR puts selection pressure on phages to evolve, selecting for increased rates of substitution and recombination. The effect that heterogeneous environments have on phage-bacteria coexistence has also been explored. A number of experimental studies have borne out theoretical predictions regarding phage diversity, mutation and recombination, and heterogeneous environments. Interestingly, phages also encode strategies to combat the bacterial immune system, including anti-CRISPR proteins and their own CRISPR-Cas system. Theoretical work and modeling continue to play an integral role in developing a fundamental understanding of the CRISPR-Cas genetic adaptive immune system of prokaryotes and in designing applications in molecular biotechnology and genomic editing.

Acknowledgements This work was partially supported by the Center for Theoretical Biological Physics at Rice University, Houston, TX 77005, USA.

References

Abudayyeh OO, Gootenberg JS, Konermann S, Joung J, Slaymaker IM, Cox DBT, Shmakov S, Makarova KS, Semenova E, Minakhin L, Severinov K, Regev A, Lander ES, Koonin EV, Zhang F (2016) C2c2 is a single-component programmable RNA-guided RNA-targeting CRISPR effector. Science 353(6299):aaf5573

Al-Attar S, Westra ER, van der Oost J, Brouns SJ (2011) Clustered regularly interspaced short palindromic repeats (CRISPRs): The hallmark of an ingenious antiviral defense mechanism in prokaryotes. Biol Chem 392(4):277–289

Andersson AF, Banfield Jillian F (2008) Virus population dynamics and acquired virus resistance in natural microbial communities. Science 320(5879):1047–1050

Barrangou R, Fremaux C, Deveau H, Richards M, Boyaval P, Moineau S, Romero DA, Horvath P (2007) CRISPR provides acquired resistance against viruses in prokaryotes. Science 315 (5819):1709–1712

Berezovskaya FS, Wolf YI, Koonin EV, Karev GP (2014) Pseudo-chaotic oscillations in CRISPR-virus coevolution predicted by bifurcation analysis. Biol Direct 9(1):1–17

Berezovsky IN, Shakhnovich EI (2005) Physics and evolution of thermophilic adaptation. Proc Natl Acad Sci USA 102(36):12742–12747

Bhaya D, Davison M, Barrangou Rodolphe (2011) CRISPR-Cas systems in bacteria and archaea: Versatile small RNAs for adaptive defense and regulation. Annu Rev Genet 45:273–297

Bolotin A, Quinquis B, Sorokin A, Dusko Ehrlich S (2005) Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin. Microbiology 151 (8):2551–2561

Briner AE, Barrangou R (2016) Deciphering and shaping bacterial diversity through CRISPR. Curr Opin Microbiol 31:101–108

Childs LM, Held NL, Young MJ, Whitaker RJ, Weitz JS (2012) Multiscale model of CRISPR-induced coevolutionary dynamics: Diversification at the interface of Lamarck and Darwin. Evolution 66(7):2015–2029

Deveau H, Barrangou R, Garneau JE, Labonté J, Fremaux C, Boyaval P, Romero DA, Horvath P, Moineau Sylvain (2008) Phage response to CRISPR-encoded resistance in *Streptococcus thermophilus*. J Bacteriol 190(4):1390–1400

Deveau H, Garneau JE, Moineau S (2010) CRISPR/Cas system and its role in phage-bacteria interactions. Annu Rev Microbiol 64:475–493

Emerson JB, Andrade K, Thomas BC, Norman A, Allen EE, Heidelberg KB, Banfield JF (2013) Virus-host and CRISPR dynamics in archaea-dominated hypersaline Lake Tyrrell, Victoria, Australia. Archaea, 2013, p 370871

Haerter JO, Trusina A, Sneppen K (2011) Targeted bacterial immunity buffers phage diversity. J Virol 85(20):10554–10560

Han P, Deem MW (2017) Non-classical phase diagram for virus bacterial coevolution mediated by clustered regularly interspaced short palindromic repeats. J R Soc Interface 14(127):20160905

Han P, Niestemski LR, Barrick JE, Deem MW (2013) Physical model of the immune response of bacteria against bacteriophage through the adaptive CRISPR-Cas immune system. Phys Biol 10(2):025004

He J, Deem MW (2010) Heterogeneous diversity of spacers within CRISPR (clustered regularly interspaced short palindromic repeats). Phys Rev Lett 105(12):128102

Heidelberg JF, Nelson WC, Schoenfeld T, Bhaya D (2009) Germ warfare in a microbial mat community: CRISPRs provide insights into the co-evolution of host and viral genomes. PLoS ONE 4(1):e4169

Horvath P, Barrangou R (2010) CRISPR/Cas, the immune system of bacteria and archaea. Science 327(5962):167–170

Horvath P, Coûté-Monvoisin A-C, Romero DA, Boyaval P, Fremaux C, Barrangou R (2009) Comparative analysis of CRISPR loci in lactic acid bacteria genomes. Int J Food Microbiol 131 (1):62–70

Howe A, Ringus DL, Williams RJ, Choo Z-N, Greenwald SM, Owens SM, Coleman ML, Meyer F, Chang EB (2015) Divergent responses of viral and bacterial communities in the gut microbiome to dietary disturbances in mice. ISME J 10:1217–1227

Hsu PD, Lander ES, Zhang F (2014) Development and applications of CRISPR-Cas9 for genome engineering. Cell 157(6):1262–1278

Iranzo J, Lobkovsky AE, Wolf YI, Koonin EV (2013) Evolutionary dynamics of the prokaryotic adaptive immunity system CRISPR-Cas in an explicit ecological context. J Bacteriol 195 (17):3834–3844

Jiang W, Maniv I, Arain F, Wang Y, Levin BR, Marraffini LA (2013) Dealing with the evolutionary downside of CRISPR immunity: bacteria and beneficial plasmids. PLoS Genet 9 (9):e1003844

Koonin EV, Wolf YI (2015) Evolution of the CRISPR-Cas adaptive immunity systems in prokaryotes: Models and observations on virus-host coevolution. Mol BioSyst 11:20–27

Lander ES (2016) The heroes of CRISPR. Cell 164(1–2):18–28

Levin BR, Moineau S, Bushman M, Barrangou R (2013) The population and evolutionary dynamics of phage and bacteria with CRISPR—mediated immunity. PLoS Genet 9(3): e1003312

Lillestøl R, Redder P, Garrett RA, Brügger KIM (2006) A putative viral defence mechanism in archaeal cells. Archaea 2(1):59–72

Mahony J, van Sinderen D (2015) Novel strategies to prevent or exploit phages in fermentations, insights from phage–host interactions. Curr Opin Biotechnol 32:8–13

Makarova KS, Wolf YI, Alkhnbashi OS, Costa F, Shah SA, Saunders SJ, Barrangou R, Brouns SJ, Charpentier E, Haft DH, Horvath P, Moineau S, Mojica FJM, Terns RM, Terns MP, White MF, Yakunin AF, Garrett RA, van der Oost J, Backofen R, Koonin EV (2015) An

updated evolutionary classification of CRISPR-Cas systems. Nat Rev Microbiol 13(11):722–736

Marraffini LA, Sontheimer EJ (2010) CRISPR interference: RNA-directed adaptive immunity in bacteria and archaea. Nat Rev Genet 11(3):181–190

Maxwell KL (2016) Phages fight back: inactivation of the CRISPR-Cas bacterial immune system by anti-CRISPR proteins. PLoS Pathog 12(1):e1005282

Minot S, Sinha R, Chen J, Li H, Keilbaugh SA, Wu GD, Lewis JD, Bushman FD (2011) The human gut virome: inter-individual variation and dynamic response to diet. Genome Res 21 (10):1616–1625

Mojica FJM, Garca-Martnez J, Soria E, Diez-Villasenor C (2005) Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements. J Mol Evol 60 (2):174–182

Paez-Espino D, Sharon I, Morovic W, Stahl B, Thomas BC, Barrangou R, Banfield JF (2015) CRISPR immunity drives rapid phage genome evolution in *Streptococcus thermophilus*. mBio 6(2):e00262-15

Pennisi E (2013) The CRISPR craze. Science 341(6148):833–836

Peters JM, Silvis MR, Zhao D, Hawkins JS, Gross CA, Qi LS (2015) Bacterial CRISPR: accomplishments and prospects. Curr Opin Microbiol 27:121–126

Pourcel C, Salvignol G, Vergnaud G (2005) CRISPR elements in *Yersinia pestis* acquire new repeats by preferential uptake of bacteriophage DNA, and provide additional tools for evolutionary studies. Microbiology 151(3):653–663

Sapranauskas R, Gasiunas G, Fremaux C, Barrangou R, Horvath P, Siksnys V (2011) The *Streptococcus thermophilus* CRISPR/Cas system provides immunity in *Escherichia coli*. Nucleic Acids Res 39(21):9275–9282

Scarpellini E, Ianiro G, Attili F, Bassanelli C, De Santis A, Gasbarrini A (2015) The human gut microbiota and virome: potential therapeutic implications. Dig Liver Dis 47(12):1007–1012

Sebaihia M, Wren BW, Mullany P, Fairweather NF, Minton N, Stabler R, Thomson NR, Roberts AP, Cerdeño-Tárraga AM, Wang H et al (2006) The multidrug-resistant human pathogen *Clostridium difficile* has a highly mobile, mosaic genome. Nat Genet 38(7):779–786

Seed KD, Lazinski DW, Calderwood SB, Camilli A (2013) A bacteriophage encodes its own CRISPR/Cas adaptive response to evade host innate immunity. Nature 494(7438):489–491

Sorek R, Kunin V, Hugenholtz P (2008) CRISPR—a widespread system that provides acquired resistance against phages in bacteria and archaea. Nat Rev Microbiol 6(3):181–186

Sorek R, Lawrence CM, Wiedenheft B (2013) CRISPR-mediated adaptive immune systems in bacteria and archaea. Ann Rev Biochem 82:237–266

Sun CL, Barrangou R, Thomas BC, Horvath P, Fremaux C, Banfield JF (2013) Phage mutations in response to CRISPR diversification in a bacterial population. Environ Microbiol 15(2):463–470

Tyson GW, Banfield JF (2008) Rapidly evolving CRISPRs implicated in acquired resistance of microorganisms to viruses. Environ Microbiol 10(1):200–207

Weinberger AD, Wolf YI, Lobkovsky AE, Gilmore MS, Koonin EV (2012) Viral diversity threshold for adaptive immunity in prokaryotes. MBio 3(6):e00456–12

Yosef I, Manor M, Kiro R, Qimron U (2015) Temperate and lytic bacteriophages programmed to sensitize and kill antibiotic-resistant bacteria. Proc Natl Acad Sci USA 112(23):7267–7272

Hidden Silent Codes in Viral Genomes

Eli Goz, Hadas Zur and Tamir Tuller

Abstract Viruses are small infectious agents that replicate only inside the living cells of other organisms and comprise approximately 94% of the nucleic acid-containing particles in the oceans. They are believed to play a central role in evolution, are responsible for various human diseases, and have important contributions to biotechnology and nanotechnology. Viruses undergo evolutionary selection for efficient transmission from host to host by exploiting the host's gene expression machinery (e.g., ribosomes) for the expression of the genes encoded in their genomes. As a result, viral genes tend to be expressed via non-canonical mechanisms that are very rare in living organisms. Many of the gene expression stages and other aspects of the viral life cycle are encoded in the viral transcripts via 'silent codes', and are induced by mutations that are synonymous to the viral amino acid content. In a series of studies that included the analyses of dozens of organisms from the three domains of life, it was shown that there are overlapping 'silent codes' in the genetic code that are related to all stages of gene expression regulation. The aim of this chapter is to summarize the current knowledge related to the silent codes in viral genomes and the open questions in the field.

E. Goz and H. Zur—Equal contribution

E. Goz · H. Zur · T. Tuller (✉)
Department of Biomedical Engineering, Tel Aviv University,
P.O. Box 39040, 6997801 Tel Aviv, Israel
e-mail: tamirtul@post.tau.ac.il

E. Goz · T. Tuller
SynVaccine Ltd. Ramat Hachayal, Tel Aviv, Israel

T. Tuller
Sagol School of Neuroscience, Tel Aviv University,
P.O. Box 39040, 6997801 Tel Aviv, Israel

P. Pontarotti (ed.), *Evolutionary Biology: Self/Nonself Evolution, Species and Complex Traits Evolution, Methods and Concepts*,
DOI 10.1007/978-3-319-61569-1_5

1 Hidden Information Related to Gene Expression Regulation and Other Aspects is Encoded in the Transcripts and Affects Organismal/Viral Fitness

Proteins are the principal actors in all intracellular activities. Gene expression is the process by which the information encoded in a gene is used to synthesize the corresponding protein. The major cellular biophysical stages of gene expression are transcription, splicing (in eukaryotes), mRNA degradation, translation, and protein degradation; each of these stages has several substages (*e.g.* initiation, elongation, and termination of translation). For many years, researchers referred to the promoter (which mainly determines the transcription initiation rates) as the 'module' that includes almost all the information related to gene expression regulation, while the information related to protein structure is contained in the coding sequence via the genetic code.

However, in recent years, it was shown that such a *modularity is only a raw approximation of the reality* (Quax et al. 2015; Supek 2016; Sauna and Kimchi-Sarfaty 2013; Fredrick and Ibba 2010; Cannarozzi et al. 2010; Bahir et al. 2009; Gorgoni, et al. 2014; Gu et al. 2010; Zafrir and Tuller 2017; Tuller and Zur 2015; Ben-Yehezkel et al. 2015; Zafrir and Tuller 2015a, Yofe et al. 2014; Diament et al. (in press); Dana and Tuller 2014a; Zur and Tuller 2012, 2013, 2016; Tuller et al. 2010a, b, 2011a; Zafrir et al. 2016; Goz et al. 2017). Various signals (codes) related to all the stages of gene expression regulation, including its dynamics and amplitude, appear also in the coding sequence (ORF) itself and in the untranslated regions (UTRs), and are involved in biophysical interactions with the other segments of the transcript, and various macromolecules involved in gene expression regulation (Quax et al. 2015; Supek 2016; Sauna and Kimchi-Sarfaty 2013; Fredrick and Ibba 2010; Cannarozzi et al. 2010; Bahir et al. 2009; Gorgoni et al. 2014; Gu et al. 2010; Zafrir and Tuller 2015b; 2017; Tuller and Zur 2015; Ben-Yehezkel et al. 2015; Yofe et al. 2014; Diament et al. (in press); Dana and Tuller 2014b; Zur and Tuller 2012; 2013; 2016; Tuller et al. 2010a, b, 2011a; Zafrir et al. 2016; Goz et al. 2017) (see Figs. 1 and 2). Transcripts tend to also include information related to/affecting additional phenomena such as co-translational protein folding (Thommen et al. 2016; Chaney and Clark 2015) and regulation by the bacterial immune system (Terns and Terns 2011).

Specifically, it is interesting to emphasize that many of these 'silent' codes are encoded in the coding regions via the redundancy of the genetic code. A certain protein can be encoded by an exponential number of codon combinations; replacing a codon with a synonymous one can significantly affect the expression of the transcript (Quax et al. 2015; Supek 2016; Sauna and Kimchi-Sarfaty 2013; Fredrick and Ibba 2010; Cannarozzi et al. 2010; Bahir et al. 2009; Gorgoni et al. 2014; Gu et al. 2010; Tuller and Zur 2015, 2017; Ben-Yehezkel et al. 2015; Zafrir and Tuller 2015a; Yofe et al. 2014; Diament et al. (in press); Dana and Tuller 2014b; 2016; Tuller et al. 2010a, b, 2011a; Zur and Tuller 2012, 2013; Zafrir et al. 2016; Goz et al. 2017; Bazzini et al. 2016; Morgunov et al. 2014; Sin et al. 2016).

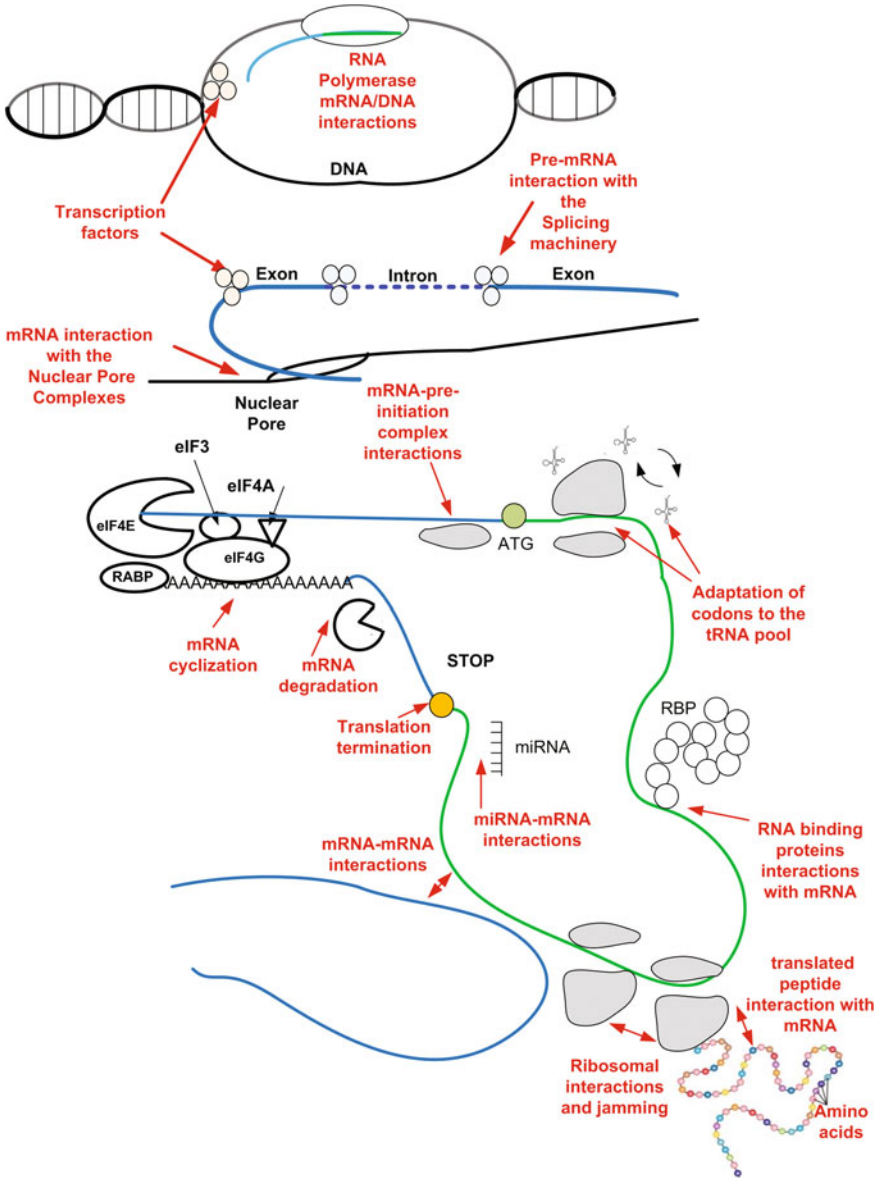


Fig. 1 Some of the interactions of the mRNA molecule with the gene expression machinery. The affinities of these interactions are encoded in the UTRs and ORFs of the genes

The information related to these codes is considered 'hidden' as it is partially encoded in synonymous/'silent' aspects of the transcript, and is much harder to model than the genetic code. Regulation of gene expression is clearly at the heart of

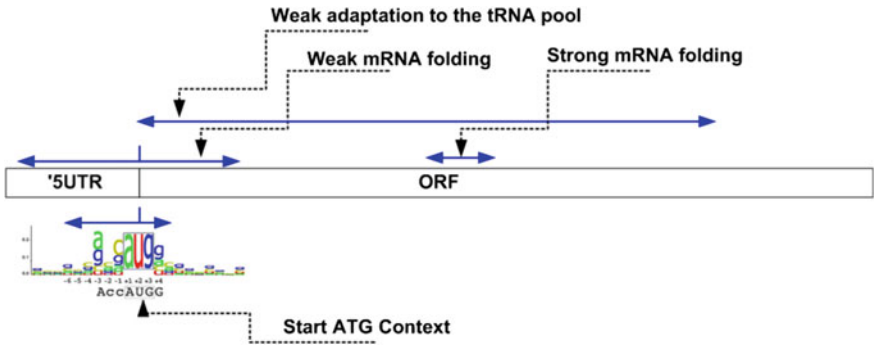


Fig. 2 Some signals related to gene translation regulation that are encoded in the coding sequence and 5'UTR (see references in the main text)

every biological system. Thus, understanding how aspects of this process are encoded in the transcript should have important ramifications to every biomedical discipline (e.g., human health, synthetic biology, molecular evolution, genetics, systems biology, etc.).

It is important to emphasize that while there are studies that suggest codon usage bias is related to mutation drift (Bulmer 1991), various lines of evidence have recently demonstrated that codon usage bias directly affects the translation elongation speed. Specifically, based on direct experimental measurements of ribosome densities (which are related to elongation rates) over the entire transcriptome at a single codon resolution, it was shown that different codons have different elongation rates, which correlate with corresponding tRNA levels (Dana and Tuller 2014b; Gardin et al. 2014). It was also experimentally shown that changing the codon content of a protein directly affects protein levels (Ben-Yehezkel et al. 2015; Gustafsson et al. 2004), and thus the organism's fitness.

2 The Importance of Understanding the 'Silent' Information Encoded in Viral Genomes

Viruses are small infectious agents that replicate only inside the living cells of other organisms. They are comprised of genetic material (RNA or DNA molecule(s)) and often additional enzymes that are enclosed within a protective coat of lipids and proteins. The viral genome contains all necessary information to initiate and complete a replication cycle within a cell. Viruses can infect all living organisms (fungi, plants, bacteria, mammals, etc.); we eat and breathe billions of virus particles regularly and carry viral genomes as part of our own genetic material.

Viruses are by far the most abundant biological entities in the oceans, comprising approximately 94% of the nucleic acid-containing particles (Zimmer 2011). They are believed to play a central role in evolution as they are important natural

means of transferring genes between different species (Zimmer 2011). In addition, viruses are responsible for various human diseases: Some of them are common (*e.g.,* common cold, influenza, chickenpox, cold sores), others are severe and fatal (*e.g.,* ebola, AIDS, avian influenza, and SARS). Viruses also have important implications to biotechnology (*e.g.,* they are often used as vectors to introduce genes into cells), and even to materials science and nanotechnology (*e.g.,* they can be used as organic nanoparticles) (Fischlechner and Donath 2007).

The viral genomes undergo evolutionary selection for efficient transmission from host to host, and for exploiting the gene expression machinery of the host (*e.g.,* ribosomes, various transcription/translation factors, etc.) for efficient synthesis of the encoded proteins and the efficient expression of various types of genes (*e.g.,* see Gale et al. 2000). As a result, viral genes tend to be expressed via non-canonical mechanisms that are either specific only to viruses, or very rare in living organisms (*e.g.,* see Gale et al. 2000; Firth and Brierley 2012; Rohde et al. 1994; Brierley 1995; Lopez-Lastra et al. 2010; Fig. 3). For example, viruses tend to include

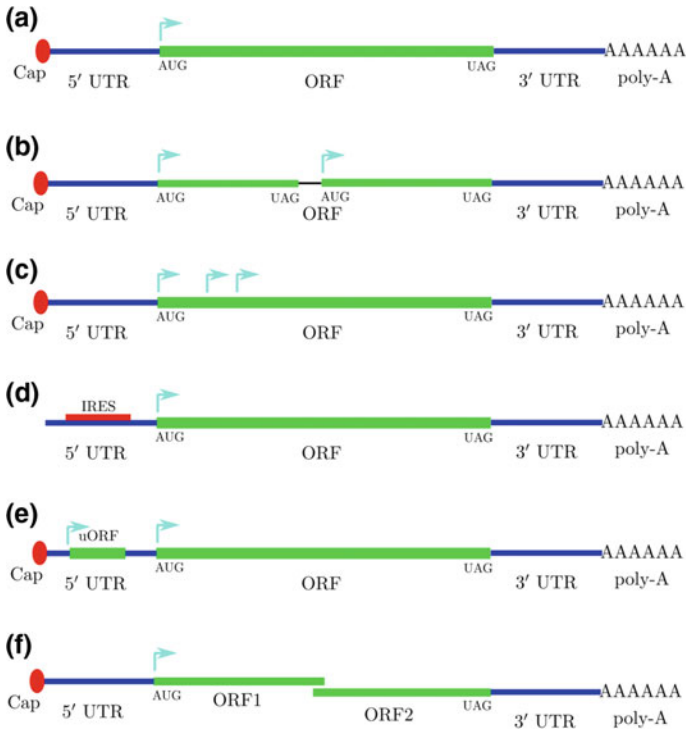


Fig. 3 Examples of non-canonical viral mRNA translation. **a** Canonical mRNA translation, which translates a single ORF from the start codon (AUG) to the stop codon (UAG, or UAA or UGA). **b**, **c**, **d**, **e** and **f** Non-canonical translation mechanisms: **b** Ribosomal re-initiation. **c** Diversity of start codon and near-cognate start codons. **d** Cap-independent translation via IRES (internal ribosome entry site). **e** Upstream-ORF (uORF). **f** Multiple/internal ORFs (either in-frame or out-of-frame)

overlapping ORFs and they often include a long ORF translated into a single polyprotein that is cleaved posttranslationaly into a set of mature proteins. Viruses also tend to initiate translation from internal ribosome entry sites (IRES), and not via canonical scanning from the 5' end of the transcript. Furthermore, frequently viral genes contain functional mRNA structures related to all stages of their expression regulation. Furthermore, frequently they include strong mRNA structure related to all stages of their gene expression regulation. Regularly, the viral genetic material is RNA and not DNA and can undergo a series of transformations before translation into proteins. Finally, most of the viral genomes are very compact and include all their gene expression information in a very short genome (typically a few thousand nucleotides), etc.

One gene expression aspect common to all viruses is the fact that all types of viruses must use the ribosomes (and other expression machinery) of their host.

It is important to emphasize that viruses evolve to include non-canonical gene expression mechanisms since these non-canonical regulatory mechanisms contribute to their fitness. Specifically, often during viral development, the canonical gene expression mechanisms in the cell are 'shut down' (e.g., due to down-regulation of relevant initiation factors); since viruses bypass these canonical mechanisms (via non-canonical mechanisms, e.g. IRES), they can successfully exploit the intracellular gene expression resources (e.g., ribosomes and tRNAs). Some of these non-canonical mechanisms (e.g., overlapping ORFs) enable a more efficient (in terms of energy) production of viruses, and decreasing the probability of deleterious mutations (Holmes 2009). Among others, this means that it is less trivial to understand the viral silent gene expression codes as they are relatively rare and unique (Gale et al. 2000; Firth and Brierley 2012; Rohde et al. 1994; Brierley 1995; Lopez-Lastra et al. 2010; Adrian et al. 2005; Holland 2012).

3 Previous Relevant Studies Concerning 'Silent' Information Related to Gene Expression in Viruses

Various studies in recent years have provided statistical evidence that silent aspects in the viral genomes are related to their fitness.

Specifically, among others, it was suggested that *very* basic features, such as mRNA folding, codon decoding times, codon or nucleotide pairs distributions (or other low order statistics of genomic sequences), may be induced by synonymous mutations and play an important role in controlling the viral life cycle (Bahir et al. 2009; Cuevas et al. 2012; Lobo et al. 2009; Jenkins et al. 2001; Greenbaum et al. 2008; van Hemert et al. 2007; Pride et al. 2006; Cardinale and Duffy 2011; Shackelton et al. 2006; Carbone 2008; Gu et al. 2004; Sau et al. 2005a, 2007; Zhao et al. 2008; Cheng et al. 2012; Lucks et al. 2008; Mueller et al. 2006). In this subsection, detail the different silent aspects of viral gene expression that have been reported thus far.

3.1 Selection for Codon Preference in Viral Genomes

The most basic property of the viral coding sequences is the frequencies of the different codons. The tendency to choose specific codons has been shown to affect/regulate intracellular mechanisms (Supek 2016; Sauna and Kimchi-Sarfaty 2013; Tuller and Zur 2015; Novoa et al. 2012): for example, it may affect translation elongation (Dana and Tuller 2014b; Gardin et al. 2014; Ben-Yehezkel et al. 2015), translation initiation (Tuller and Zur 2015; Zur and Tuller 2013; Kozak 1986), splicing (Zafrir and Tuller 2015b; Chamary and Hurst 2005), mRNA folding (Gu et al. 2010; Zur and Tuller 2012; Tuller et al. 2010), protein folding (Pechmann and Frydman 2013; Kramer et al. 2009), and more; thus, we expect that viral codon bias will be under selection pressure.

Indeed, many studies have suggested that viral codons may be under selection to improve the viral fitness, for example, via adaptation to the host tRNA pool (or other translation resources) (Bahir et al. 2009; Burns et al. 2006; Tao et al. 2009; Jia et al. 2009; Zhou et al. 2010; Liu et al. 2010, 2011; Das et al. 2006; Cai et al. 2009; Sau et al. 2005b; Wong et al. 2010; Zhong et al. 2007; Zhang et al. 2013; Novella et al. 2004; Michely et al. 2013; Roychoudhury and Mukherjee 2010; Ma et al. 2011; Aragones et al. 2010; Tsai et al. 2007; Su et al. 2009; Bull et al. 2012; Zhao et al. 2005). The adaptation of the viral codon usage bias to the tRNA pool is expected to improve translation efficiency via better allocation of the limited translation resources (e.g., ribosomes and tRNA molecules) (Dana and Tuller 2014b; Rocha 2004; Sharp et al. 2005).

In order to study the effects and extents of codon usage bias many measures have been developed (Sharp and Li 1987; dos Reis et al. 2004; Sabi et al. 2016; Wright 1990).

For example, Bahir et al. (Bahir et al. 2009) analyzed a large data set of viruses that infect hosts ranging from bacteria to humans. They show that bacteria-infecting viruses are strongly adapted to their specific hosts in terms of codon usage bias but that they differ from other unrelated bacterial hosts. Viruses that infect humans, but not those that infect other mammals or aves, show a strong resemblance to most mammalian and avian hosts, in terms of codon preferences. This observation can be partially explained by the following points: (1) There is similarity in the codon usages among most mammals (Bahir et al. 2009). (2) The codon usage bias among bacteria is very high (Bahir et al. 2009). (3) Bacteria (and thus probably also their viruses) usually undergo stronger selection for codon usage bias and for various aspects of translation optimality (among others due to their larger population size) relatively to most eukaryotes (dos Reis et al. 2004; dos Reis and Wernisch 2009). (4) Additional explanations may be related to the recent expansion of humans and the coevolution of their viruses, or to the hypothesis that large portions of the human genome are actually of viral origin (Bahir et al. 2009; Kazazian 2004).

Pavesi et al. suggested that the fact viruses undergo selection to include specific codons can help detecting new and ancestral viral coding regions (Pavesi et al. 2013). Aragonès et al. suggested that the Hepatitis A virus undergoes various types of adaptations to fine-tune the translation kinetics, among others, via selection on

codon usage bias (Aragones et al. 2010). A study by Bull et al. (2012) has shown that when reeingineering the major capsid gene of the bacteriophage T7 with varying levels of suboptimal synonymous codons , the fitness of the constructs declines linearly with the number of suboptimal changes. These experiments/analyses suggest a direct relation between codon usage bias and fitness/fitness-recovery. Similarly, a related study by Lauring et al. (2012) compared the wild-type poliovirus to synthetic viruses carrying reengineered capsid sequences with hundreds of synonymous mutations. They found that such mutations are related to the rewiring of the population's mutant network which reduced its robustness to mutations and attenuated the virus in an animal model of infection.

It is important to mention that some of these codon usage bias patterns may be associated with regulatory signals not necessarily directly related to tRNA levels (Gog et al. 2007); alternative or partial explanations to viral codon usage bias are mutational bias, asymmetrical mutational bias in two DNA strands, temperature, viral replication mechanisms, protein folding, dinucleotide distribution, mRNA folding, and more (Das et al. 2006; Zhang et al. 2011, 2013; Sau and Deb 2009; Adams and Antoniw 2004; Cardinale et al. 2013; Berkhout et al. 2002; Pinto et al. 2007; Cladel et al. 2008; Choi et al. 2005; Zhou et al. 2013; Burns et al. 2009; Liu et al. 2012). The effect on chromatin structure and nucleosome positioning is another potential constraint on the viral codon frequency distribution, as viruses are exposed to histones produced by the host (Eslami-Mossallam et al. 2016; Cohanim and Haran 2009; Babbitt and Schulze 2012).

Interestingly, a recent line of studies suggested that codon pairs' distribution is an important feature under selection in various viruses, which may be used for their attenuation for developing new vaccines (Coleman et al. 2008; Mueller et al. 2008; Martrus et al. 2013). However, there is a debate regarding this feature, while some researchers believe that it is related *directly* to the distribution of codon pairs (Coleman et al. 2008; Mueller et al. 2008; Martrus et al. 2013), others have suggested that it is related to the distribution of dinucleotides (Tulloch et al. 2014) which affect RNA folding (see, e.g., Babak et al. 2007), or may be related to the enhanced innate immune responses to viruses with elevated CpG/UpA dinucleotide frequencies rather than the viruses themselves being intrinsically defective (Tulloch et al. 2014; Belalov and Lukashev 2013). These possible explanations still connect the viral fitness to silent features of its genome, demonstrating their importance and influence on viral fitness and evolution. Finally, it is important to emphasize the fact that many silent viral codes are localized to specific regions within the genome (Dumans et al. 2004).

3.2 *Evidence for Condition Specific Adaptation to Codon Bias*

Intriguingly, a recent study has provided evidence of selection for distinct compositions of synonymous codons in viral genes that are expressed at different stages of the viral life cycle (e.g., early and late viral genes): It was shown that in the

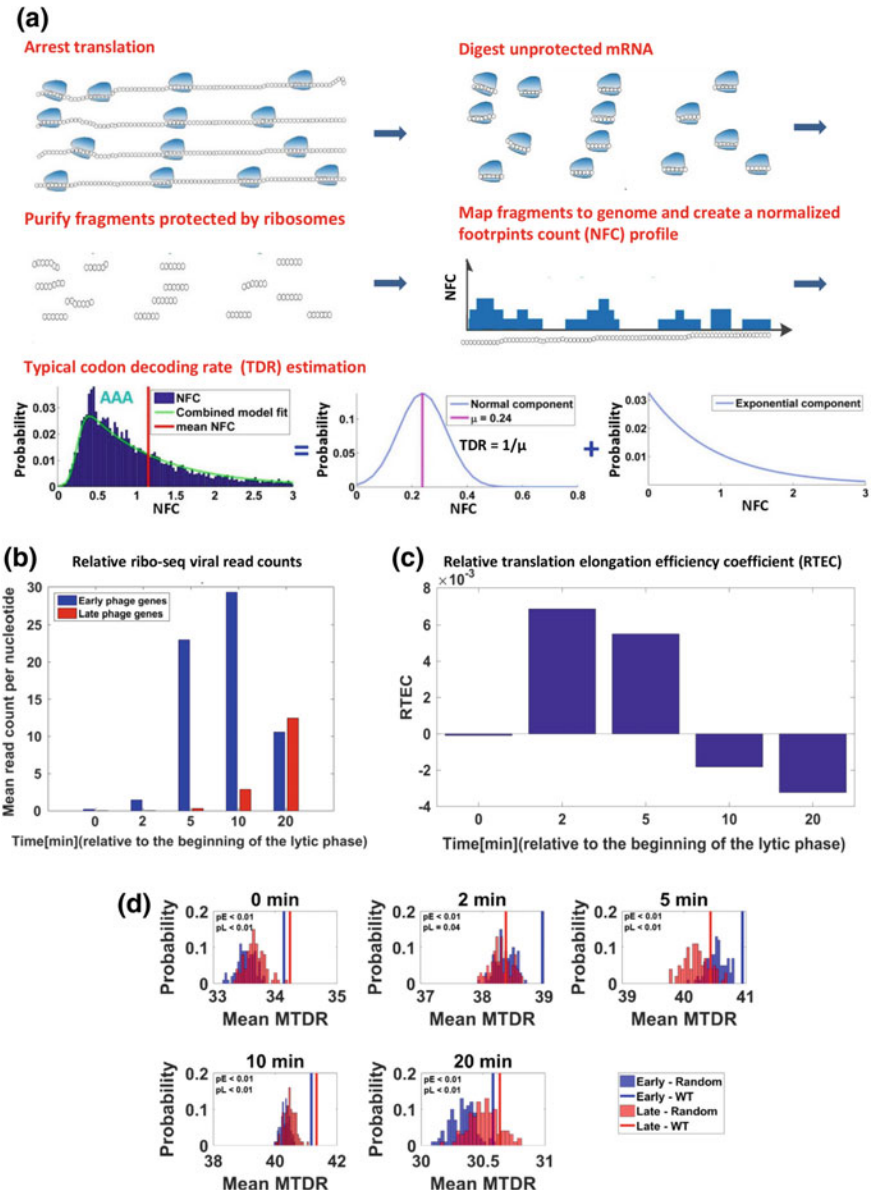
bacteriophage lambda, evolution of viral coding regions is driven, among others, by codon 'selection' which is specific to the expression time of the gene during the viral development (e.g., early expressed genes versus late expressed genes). Specifically, during the initial/progressive stages of infection, the decoding rates in early/late genes were found to be superior to those in late/early genes, respectively (Goz et al. 2017) (Fig. 4). This study is important since it is the first to show that the selection for codon usage in the virus is directly related to translation elongation rates. In addition, it was shown for the first time that codon elongation rates change during viral evolution; thus, this is expected to affect the codons 'selected' for each viral gene based on its expression time during the viral development cycle. Currently, due to the absence of experimental measurements, this result has been demonstrated only in one virus (bacteriophage lambda), since to perform such an analysis one needs to infer the codon decoding rates in different time points of the viral development. This can be achieved only via relevant experiments (Ingolia et al. 2009) and data filtering (Dana and Tuller 2014b), and the viral genome alone is not enough.

Specifically, one type of relevant experiment is Ribo-Seq which provides large-scale information (the entire transcriptome) related to the probability for seeing a ribosome over each codon in the trasncriptome in vivo (Ingolia et al. 2009). These experiments, when performed in different viral development stages, can be used for estimating the decoding rates of different codons in different viral conditions (Goz et al. 2017; Liu et al. 2013). Since Ribo-Seq data includes various sources of bias and noise, the data should be analyzed with tools tailored specifically for parameter estimation and bias filtering in the Ribo-Seq experiments (Dana and Tuller 2014b; Diament and Tuller 2016).

As can be seen in (Fig. 4a), to estimate codon decoding rates we do not compute a simple average of the normalized Ribo-Seq footprint count (NFC). The main reasons that a simple average does not work are related to: (1) The fact that Ribo-Seq includes various types of non-trivial biases (e.g., very extreme values in certain positions due to the biochemistry of the protocol) (Dana and Tuller 2012, 2014b; Diament and Tuller 2016; Gerashchenko and Gladyshev 2017). (2) Codons upstream of slower codons will have more reads due to traffic jams (Dana and Tuller 2014b). (3) Codons downstream of slower codons will have more reads due to incomplete halting of the ribosomes movement during the Ribo-Seq experiment (Hussmann et al. 2015).

Consequently, the NFC always has a very thick right tail. It was shown via simulations of the Ribo-Seq procedure (Dana and Tuller 2014a, b) that without the aforementioned problems, the NFC distribution is close to normal (resembles a Gaussian without the thick right tail). Thus, to estimate the nominal decoding rate, we must filter the right tail. It was shown via Ribo-Seq simulations that the suggested filtering estimates the correct decoding times, but due to the reasons explained above merely taking the mean of the entire NFC distribution does not correlate with the actual decoding times (Dana and Tuller 2014b).

We believe that in the future, similar results will be reported for additional viruses.


(a)
Arrest translation
Digest unprotected mRNA
Purify fragments protected by ribosomes
Map fragments to genome and create a normalized footprints count (NFC) profile
NFC
Typical codon decoding rate (TDR) estimation
AAA
NFC
Combined model fit
mean NFC
Probability
NFC
Normal component
μ = 0.24
TDR = 1/μ
Exponential component
(b)
Relative ribo-seq viral read counts
Early phage genes
Late phage genes
Mean read count per nucleotide
Time[min](relative to the beginning of the lytic phase)
(c)
Relative translation elongation efficiency coefficient (RTEC)
RTEC
Time[min](relative to the beginning of the lytic phase)
(d)
0 min
2 min
5 min
10 min
20 min
Probability
Mean MTDR
pE < 0.01
pL < 0.01
pL = 0.04
Early - Random
Early - WT
Late - Random
Late - WT

◂**Fig. 4** **a** Schematic description of the ribosome profiling method, generation of the NFC distributions, and estimation of typical decoding rates of codons. Translation of mRNA codons (*black circles*) by ribosomes (*blue shapes*) is arrested, then exposed mRNA is digested. Protected mRNA footprints are then sequenced, mapped onto the genome, and normalized per gene by their mean read count value, resulting in NFC profiles. Then, NFC values of each specific codon type (NFC values of codons of type 'AAA' are demonstrated) are collected from all analyzed genes and presented via a histogram, where the *x-axis* represents the NFC values and the *y-axis* represents the fraction of the time (probability) each NFC value appears in the analyzed genes, thus creating the NFC distribution of a codon. (e.g, the codon 'AAA' appears with an NFC value that equals 1 in the analyzed genes in 1.6% of the times.) The combined normal/exponential model fitting of codon NFC distribution is plotted as a *curve*. The position of the mean NFC value is presented with a *vertical line*. The NFC distribution can be decomposed into a normal and an exponential component using a log-likelihood fitting. The mean of the normal component is used for computing the Mean of the Typical Decoding Rate (MTDR) of coding regions. **b** Relative expression levels of each of the lambda phage gene groups (early/late) in Ribo-Seq read count per nucleotide. **c** Adaptation of translation elongation efficiency in early and late genes to different bacteriophage development stages genes. Relative translation elongation efficiency coefficient, $RTEC = (mean\,MTDR_E - mean\,MTDR_L)/(mean\,MTDR_E + mean\,MTDR_L)$, as a function of time from the beginning of the lytic stage (0–20 min), where $MTDR_E$ and $MTDR_L$ are the MTDR of early and late genes, respectively. We can see that the RTEC of early genes is higher at the beginning and becomes lower with time (as expected); the first point (t = 0), when there are no measurements of expression is ignored. **c** Selection for translation elongation efficiency in bacteriophage coding regions. At each time point, average MTDR values of wild-type early/late genes (*vertical bars*) were compared to MTDR values of 100 corresponding randomized variants (histograms). Average wild-type MTDR values of each group are significantly higher ($p < 0.05$) than expected in random. The late genes were sampled to control for the length factor

3.3 mRNA Structures of the Coding Sequences and UTRs

Various previous studies have suggested that the UTRs of many viruses include important functional structures (Watts et al. 2009; Firth et al. 2011; Brown et al. 1992; Hyde et al. 2014; Abbink and Berkhout 2003). For example, it was demonstrated that extensive structural elements that modulate RNA replication via different conformations appear in the 5′ and 3′ UTRs of Dengue and other flaviviruses. The promoter for Dengue virus RNA synthesis is a large stem-loop structure located at the 5′ end of the genome. This structure specifically interacts with the viral polymerase NS5 and promotes RNA synthesis at the 3′ end of a circularized genome. The circular conformation of the viral genome is mediated by long-range RNA–RNA interactions that span thousands of nucleotides (Fig. 5).

As another example, the genomes of human hepatitis C virus (HCV), and the animal pestiviruses responsible for bovine viral diarrhea (BVDV) and hog cholera (HChV), have a conserved (and probably functional) stem-loop structure in the 3' 200 bases of the 5'UTR (Brown et al. 1992). A different study (Hyde et al. 2014) suggested that the pathogenic alphaviruses use secondary structural motifs within the 5'UTR as part of an evasion mechanism by which viruses avoid immune restriction.

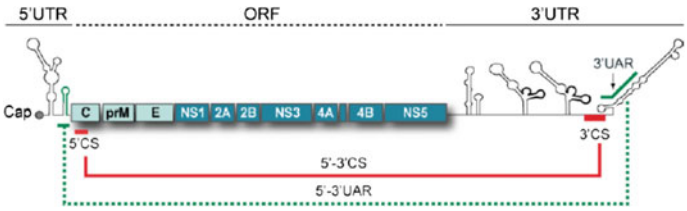


Fig. 5 Illustration of the functional RNA structures at the UTRs of the Dengue virus. Among others, these structures are related to genome cyclization, which is mediated by long-range RNA–RNA interactions and enables the polymerase to reach the 3′ end of long RNA molecules (adapted from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3187688/)

Interestingly, Firth et al. (2011) analyzed the ∼150nt 3′-adjacent to the stop codon (UGA) in Sindbis, Venezuelan equine encephalitis related alphaviruses, and in the plant virus genera (Furovirus, Pomovirus, Tobravirus, Pecluvirus and Benyvirus); they found a phylogenetically conserved stem-loop structure. Mutational analysis of the predicted structure demonstrated that the stem-loop increases read-through by up to ten-fold. Thus, this structure has an important function: increasing read-through probability.

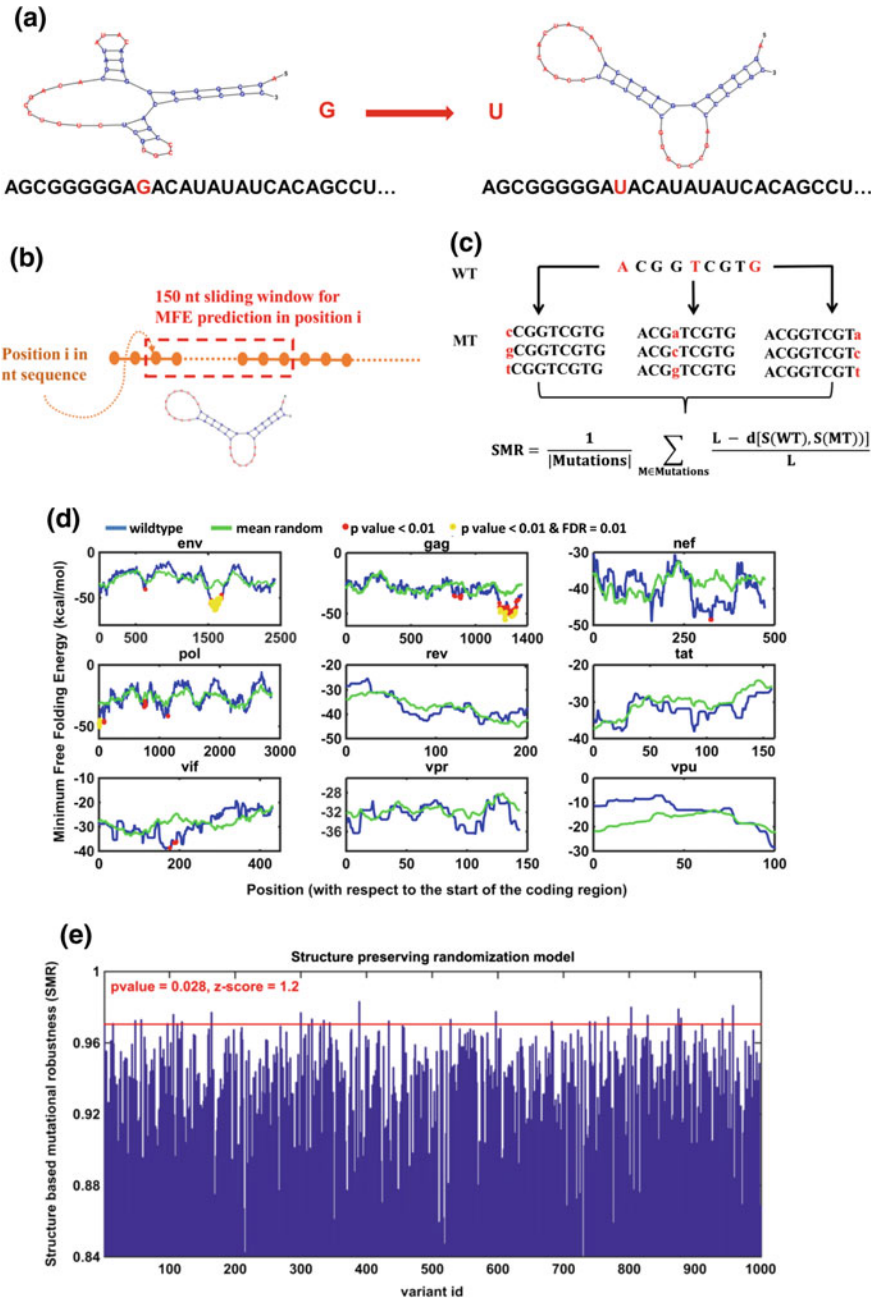
An interesting question is related to the possibility that such important functional structures appear inside the coding regions of viruses. To check this possibility, the strength of the structures within the coding regions of viral genomes can be compared to the ones we 'expect to obtain' under a 'null evolutionary model' that generates viral genomes with similar properties to the original genome (such as, encoded proteins, GC content, codon frequencies, identical distances/alignment-scores between the viral strains of the same virus). Two recent studies have performed such analyses (Goz and Tuller 2015, 2016).

In these papers (Goz and Tuller 2015, 2016), 1666 genomes of the four Dengue serotypes and the HIV genome were analyzed, using statistical/computational analyses to detect dozens of positions suspected to undergo selection for weak/strong local mRNA folding (probably many of them are related to viral fitness), while controlling for the false discovery rate.

An extensive position-specific selection for global and local mRNA structures in these viruses was demonstrated (Goz and Tuller 2015, 2016) (see also Goz et al. 2017). In addition, since robustness to mutations is an important factor that influences viral evolution (expressly in the case of RNA viruses) (Lauring et al. 2013), it was specifically interesting to provide evidence related to the robustness of some of these structures to mutations/errors (Goz and Tuller 2016) (Fig. 6).

Inference of the HIV RNA structure (Watts et al. 2009) suggested that there is correlation between high levels of RNA structure and sequences that encode inter-domain loops in HIV proteins.

It was shown that RNA structure can effect translation elongation rates (Tuller et al. 2011a; Dana and Tuller 2012); it was also shown that the elongation rates can effect co-translational folding (Zur and Tuller 2016; Yang et al. 2014; Faure et al.


(a)
G
U
AGCGGGGGAGACAUAUAUCACAGCCU...
AGCGGGGGAUACAUAUAUCACAGCCU...
(b)
150 nt sliding window for
MFE prediction in position i
Position i in
nt sequence
(c)
WT
A C G G T C G T G
MT
cCGGTCGTG ACGaTCGTG ACGGTCGTa
gCGGTCGTG ACGcTCGTG ACGGTCGTc
tCGGTCGTG ACGgTCGTG ACGGTCGTt
SMR = 1/|Mutations| ∑ M∈Mutations (L − d[S(WT), S(MT))]/L
(d)
wildtype
mean random
p value < 0.01
p value < 0.01 & FDR = 0.01
env
gag
nef
pol
rev
tat
vif
vpr
vpu
Minimum Free Folding Energy (kcal/mol)
Position (with respect to the start of the coding region)
(e)
Structure preserving randomization model
pvalue = 0.028, z-score = 1.2
Structure based mutational robustness (SMR)
variant id

◂**Fig. 6** **a** Modification of the wild-type secondary structure (*left*) after introducing a single-point G → U mutation (*right*); the mutated nucleotides are marked in *red*; the distance between the wild-type and mutated secondary structures (number of changes in the base–pair connections required to transfer one structure into another) in this example its d = 13. **b** Prediction of MFE (minimum free folding energy) in local windows (*red broken-line square*) along the coding sequence (*brown*): each position i in the sequence was assigned with the MFE value predicted in the 150nt window starting at this position. **c** Computation of the structure-based mutational robustness (SMR): L—sequence length; d—base–pair distance between the secondary structure of the wild-type sequence (S(WT)) and the secondary structure of the mutant (S(MT)), averaged over all single-point mutants at all positions along the sequence (total of 3L mutants). **d Evidence that specific regions of HIV structural genes undergo an evolutionary selection for strong folding**. Each panel corresponds to wild-type (*blue*) and mean randomized (*green*) MFE profiles for one gene: The y-axis corresponds to the MFE (kcal/mol) in the 150nt genomic window starting at positions specified along the x-axis (nucleotide coordinate given with respect to the start of the coding region); *red points*—positions with MFE related p-value < 0.01 (in these positions the wild-type folding is stronger than in 99% of than randomized variants); *yellow points*—MFE-selected positions (MFE p-value < 0.01 and BH-FDR = 0.01), these positions span genomic regions that are conjectured to undergo an evolutionary selection for strong folding. We can see clusters of MFE-selected positions in each one of the structural genes (env, gag, pol); in other genes, no evidence of selection for strong folding was found. **e Structure-based mutation robustness of RRE**. X-axis—variant id: 1—wild type; 2—1001-randomized (structure preserving and dinucleotide and amino acid preserving variants). Y-axis—Structure-based mutational robustness (SMR). The *red line* corresponds to the wild-type SMR value. The p-value (portion of randomized variants with a higher robustness than in wild-type) and z-score (number of standard deviations the wild-type SMR is higher than the mean randomized SMR) are specified in *red*. We can conclude that RRE is significantly more robust than in random, and this robustness cannot be explained by the specific secondary structure of the corresponding region, its folding strength and/or other sequence attributes such as composition of dinucelotides and amino acids

2016). Thus, it is possible that, among others, the RNA structure modulates ribosome elongation to promote native protein folding. It was shown that unstructured RNA regions tend to include splice site acceptors and hypervariable regions. The HIV genome also includes a functional ribosomal gag-pol frameshift stem-loop.

These results suggest that the coding regions, and not only the UTRs, of various viruses are populated with local RNA structures that are important for the viral life cycle and fitness.

3.4 Longer Hidden Codes in the Viral Coding Sequence

As we mentioned in the previous section, we expect the viral coding region to include many codes/patterns that are important for the viral fitness and are longer and more complex than the single codon distribution. Thus, to show this we recently performed large-scale analyses of most all the viruses with available genomes and their hosts with a novel method for detecting hidden silent codes (that cannot be explained by codon bias) (Zur and Tuller 2015) in the viral genetic material. The new statistical measure compares that mean repetitive patterns in the

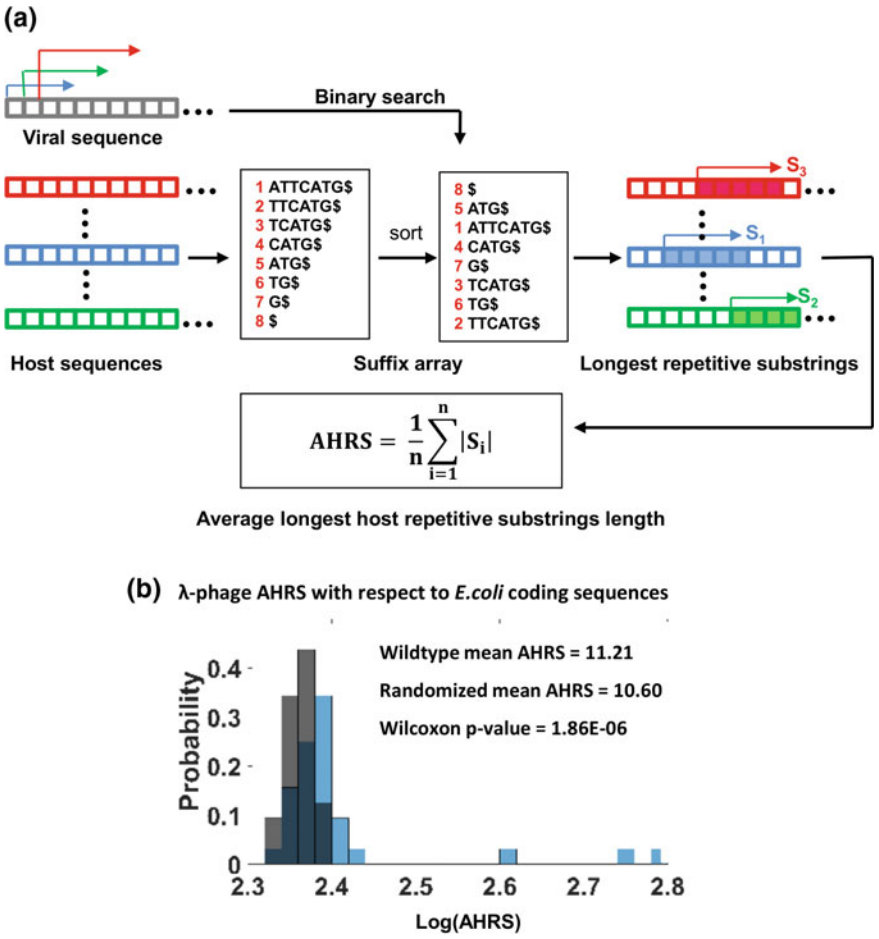


Fig. 7 **a** The statistical approach for evaluating the tendency of a viral coding region to include long subsequences that tend to appear in the host. At each position in the coding region, the length of the longest subsequence starting in this position that also appears in the host is computed. The average longest host repetitive score (AHRS) is the average of all these lengths. **b** To evaluate the statistical significance of the AHRS in the viral coding sequences, the score was compared to the ones obtained for randomized versions of the viral genomes maintaining the proteins, codon frequencies, dinucleotide frequencies, and GC content. The figure includes the analysis for the bacteriophage lambda (Goz et al. 2017)

viral and host genome and identify signals that are not expected to appear in these genomes based only on the distribution of single codons (Goz et al. 2017; Zur and Tuller 2015; Goz and Tuller 2017) (see Fig. 7).

Based on this analysis, we were able to detect significant patterns of such codes (repetitive sequences) in a high percentage of the analyzed viruses (33–90% for different groups of viruses classified according to their host) and in 90% of the bacteriophages (Goz et al. 2017; Goz and Tuller 2017).

4 Discussion

Engineering viruses

It is important to mention that there are some preliminary studies regarding gene expression engineering/modeling and other related aspects (see,e.g., Gorgoni et al. 2014; Tuller et al. 2011a; Sin et al. 2016; Konur et al. 2016; Sanassy et al. 2015; Wu et al. 2016; Schoech and Zabet 2014; Cheng et al. 2016; Pan et al. 2016; Haldane et al. 2014; Raveh et al. 2016; Reuveni et al. 2011), but none dealing with complete viruses. Thus, one open question is related to the development of practical strategies for engineering viruses based on the hidden/silent information. Developing approaches for controlling these codes should enable us to manipulate (e.g., increase or decrease) the expression levels of viral genes, and thus to modulate various viral phenotypes such as replication rates.

Therefore, based on such an approach, it will be possible to efficiently engineer viruses (Wimmer et al. 2009) for various objectives related to human health such as the design of live attenuated and killed vaccines (Lauring et al. 2010). Today, almost all the approaches for designing vaccines are based on non-synonymous alterations of the viral genomes, ignoring the largest fraction of the information (i.e., the silent information) encoded in the viral genome. Indeed, some preliminary studies have suggested that modulating simple features, such as codon and codon-pair usage, and local mRNA folding, can be used for the development of live attenuated vaccines (Coleman et al. 2008; Goz and Tuller 2015; Nogales et al. 2014).

Such an approach can also be generalized to engineer bacteriophages for various objectives such as 'fighting' pathogenic bacteria resistance to antibiotics, and engineering the human microbiome. It may also be used to design better oncolytic viruses with improved replication/fitness in cancerous cells but not in healthy ones.

Is it possible that some of the silent codes are related to the immune system?

In this book chapter, we emphasized the relation between sequence patterns in the viral coding sequences and transcripts, and viral fitness, via their effect on gene expression. However, it is possible that some of these patterns are related not only to gene expression, but also to the evolution of the virus for escaping the host immune system. It is important to emphasize that in most of the analyses, we and others reported (some are mentioned above), the *amino acid content* of the viral genes was controlled for. Thus, the reported signals cannot trivially be attributed only to the classical mechanisms, such as viral recognition by the host (e.g., antibodies), as these mechanisms are traditionally believed to be based on interactions between proteins. However, it is very plausible that they are related to alternative known and/or unknown mechanisms.

One very relevant such mechanism in bacteria is clustered regularly interspaced short palindromic repeats (CRISPR; see Fig. 8) (Marraffini 2015; Horvath and Barrangou 2010). This mechanism is based on creating fragments in the viral genome that are transcribed to short RNA molecules (crRNAs); these short RNA molecules match a certain region in the viral genome and 'guide' a protein complex

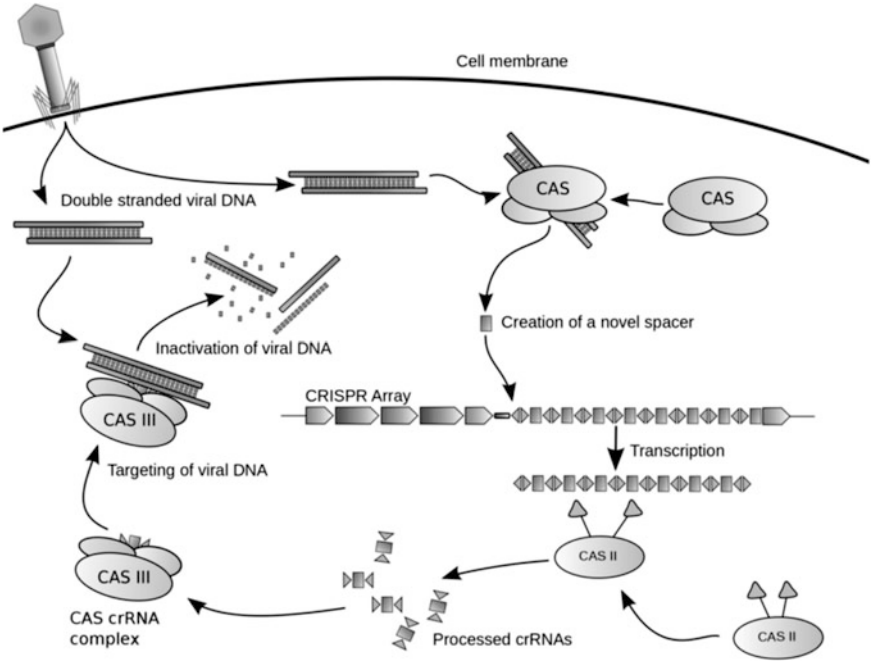


Fig. 8 The short palindromic repeat (CRISPR)-Cas system provides adaptive immunity against foreign elements in prokaryotes: Upon viral injection, a small sequence of the viral genome, known as a spacer, is integrated into the CRISPR locus to immunize the host cell. Spacers are transcribed into small RNA guides that direct the cleavage of the viral DNA by Cas nucleases (Horvath and Barrangou 2010)

(CAS-crRNA complex) that cuts the viral DNA in this region and inactivates the virus. Since this mechanism is based on the recognition of short DNA subsequences that should appear in the virus/phage but not in the host, this may trigger evolution of the nucleotide composition of the virus/phage to be similar to the host. This may result in similar patterns of codons, and longer sequences that appear in the phage and the host, explaining some of the results reported here (Goz and Tuller 2017).

Relation to horizontal gene transfer (HGT)

Finally, it is important to emphasize that similarly to viral adaptation to the host, silent features of the coding regions are expected to affect related phenomena such as horizontal gene transfer (HGT). In this case, a transferred gene is expected to be successfully expressed in a new host if its silent features are compatible (Tuller et al. 2011b; Tuller 2011, 2012; Roller et al. 2013; Medrano-Soto et al. 2004). Thus, many of the results reported here may be generalized to the case of HGT.

It is important to emphasize that a central HGT mechanism is transduction, the process in which bacterial DNA is moved from one bacterium to another by a virus/bacteriophage (Soucy et al. 2015). Thus, the reported relations between (1) the

host silent codes and (2) the transferred gene silent codes have much overlap: The fact that viral fitness is related to the similarity of its silent aspects/codes to the host should directly improve its ability to transfer genes; it is also directly related to the fact that the silent aspects/codes in the transferred genes are more adapted to the new host since the virus undergoes evolution to be better adapted to the host.

Some preliminary studies of heterologous gene expression have suggested that introducing a foreign gene with a distinct codon distribution to the host results in a decrease in the host's fitness and the gene's protein levels (Gustafsson et al. 2004; Ben-Yehezkel et al. 2015; Tuller et al. 2011; Welch et al. 2009). Computational models have suggested that this is partially due to the fact that such genes recruit more ribosomes (slower codons result in ribosomes spending more time on the mRNA), the number of available ribosomes decreases, the global initiation rates of the host genes decreases, and thus the host fitness decreases (Raveh et al. 2016; Tuller et al. 2011b; Tuller 2011) (though many additional explanations exist (Tuller and Zur 2015; Tuller 2012; Welch et al. 2009; Angov 2011)). However, additional experimental studies should be performed to better understand the effect of the codon bias of a transferred gene on the transferred gene expression and the host fitness.

Acknowledgements This study was supported in part by a fellowship from the Edmond J. Safra Center for Bioinformatics at Tel-Aviv University. TT is partially supported by the Minerva ARCHES award.

References

Abbink TE, Berkhout B (2003) A novel long distance base-pairing interaction in human immunodeficiency virus type 1 RNA occludes the gag start codon. J Biol Chem 278:11601–11611

Adams MJ, Antoniw JF (2004) Codon usage bias amongst plant viruses. Arch Virol 149:113–135

Adrian JG et al (2005) Molecular basis of virus evolution. University Press, Cambridge

Angov E (2011) Codon usage: nature's roadmap to expression and folding of proteins. Biotechnol J 6:650–659

Aragones L et al (2010) Fine-tuning translation kinetics selection as the driving force of codon usage bias in the hepatitis A virus capsid. PLoS Pathog 6:e1000797

Babak T et al (2007) Considerations in the identification of functional RNA structural elements in genomic alignments. BMC Bioinform 8:33

Babbitt GA, Schulze KV (2012) Codons support the maintenance of intrinsic DNA polymer flexibility over evolutionary timescales. Genome Biol Evol 4:954–965

Bahir I et al (2009) Viral adaptation to host: a proteome-based analysis of codon usage and amino acid preferences. Mol Syst Biol 5:1–14

Bazzini AA et al (2016) Codon identity regulates mRNA stability and translation efficiency during the maternal-to-zygotic transition. EMBO J 35:2087–2103

Ben-Yehezkel T et al (2015) Rationally designed, heterologous S. cerevisiae transcripts expose novel expression determinants. RNA Biol 12:972–984

Belalov IS, Lukashev AN (2013) Causes and implications of codon usage bias in RNA viruses. PLoS One 8:e56642

Ben-Yehezkel T et al (2015) Rationally designed, heterologous S. cerevisiae transcripts expose novel expression determinants. RNA Biol 12:972–984

Berkhout B et al (2002) Codon and amino acid usage in retroviral genomes is consistent with virus-specific nucleotide pressure. AIDS Res Hum Retroviruses 18:133–141
Brierley I (1995) Ribosomal frameshifting viral RNAs. J Gen Virol 76(Pt 8):1885–1892
Brown EA et al (1992) Secondary structure of the 5' nontranslated regions of hepatitis C virus and pestivirus genomic RNAs. Nucleic Acid Res 20:5041–5045
Bull JJ et al (2012) Slow fitness recovery in a codon-modified viral genome. Mol Biol Evol 29:2997–3004
Bulmer M (1991) The selection-mutation-drift theory of synonymous codon usage. Genetics 129:897–907
Burns CC et al (2006) Modulation of poliovirus replicative fitness in HeLa cells by deoptimization of synonymous codon usage in the capsid region. J Virol 80:3259–3272
Burns CC et al (2009) Genetic inactivation of poliovirus infectivity by increasing the frequencies of CpG and UpA dinucleotides within and across synonymous capsid region codons. J Virol 83:9957–9969
Cai MS et al (2009) Characterization of synonymous codon usage bias in the duck plague virus UL35 gene. Intervirology 52:266–278
Cannarozzi G et al (2010) A role for codon order in translation dynamics. Cell 141:355–367
Carbone A (2008) Codon bias is a major factor explaining phage evolution in translationally biased hosts. J Mol Evol 66:210–223
Cardinale DJ, Duffy S (2011) Single-stranded genomic architecture constrains optimal codon usage. Bacteriophage 1:219–224
Cardinale DJ et al (2013) Base composition and translational selection are insufficient to explain codon usage bias in plant viruses. Viruses 5:162–181
Chamary JV, Hurst LD (2005) Biased codon usage near intron-exon junctions: selection on splicing enhancers, splice-site recognition or something else? Trends Genet 21:256–259
Chaney JL, Clark PL (2015) Roles for synonymous codon usage in protein biogenesis. Annu Rev Biophys 44:143–166
Cheng Z et al (2016) Differential dynamics of the mammalian mRNA and protein expression response to misfolding stress. Mol Syst Biol 12:855
Cheng XF et al (2012) High codon adaptation in citrus tristeza virus to its citrus host. Virol J 9:113
Choi IR et al (2005) An internal RNA element in the P3 cistron of wheat streak mosaic virus revealed by synonymous mutations that affect both movement and replication. J Gen Virol 86:2605–2614
Cladel NM et al (2008) CRPV genomes with synonymous codon optimizations in the CRPV E7 gene show phenotypic differences in growth and altered immunity upon E7 vaccination. PLoS One 3:e2947
Cohanim AB, Haran TE (2009) The coexistence of the nucleosome positioning code with the genetic code on eukaryotic genomes. Nucleic Acid Res 37:6466–6476
Coleman JR et al (2008) Virus attenuation by genome-scale changes in codon pair bias. Science 320:1784–1787
Cuevas JM et al (2012) The fitness effects of synonymous mutations in DNA and RNA viruses. Mol Biol Evol 29:17–20
Dana A, Tuller T (2012) Determinants of translation elongation speed and ribosomal profiling biases in mouse embryonic stem cells. PLoS Comput Biol 8:e1002755
Dana A, Tuller T (2014a) Properties and determinants of codon decoding time distributions. BMC Genomics 15(Suppl 6):S13
Dana A, Tuller T (2014b) The effect of tRNA levels on decoding times of mRNA codons. Nucleic Acid Res 42:9171–9181
Das S et al (2006) Synonymous codon usage in adenoviruses: influence of mutation, selection and protein hydropathy. Virus Res 117:227–236
Diament A, Tuller T (2016) Estimation of ribosome profiling performance and reproducibility at various levels of resolution. Biol Direct 11:24
Diament A et al (2014) Three dimensional genomic organization of eukaryotic genes is correlated with their expression and function. Nat Commun (in press)

dos Reis M, Wernisch L (2009) Estimating translational selection in eukaryotic genomes. Mol Biol Evol 26:451–461
dos Reis M et al (2004) Solving the riddle of codon usage preferences: a test for translational selection. Nucleic Acid Res 32:5036–5044
Dumans AT et al (2004) Synonymous genetic polymorphisms within Brazilian human immunodeficiency virus Type 1 subtypes may influence mutational routes to drug resistance. J Infect Dis 189:1232–1238
Eslami-Mossallam B et al (2016) Multiplexing genetic and nucleosome positioning codes: a computational approach. PLoS One 11:e0156905
Faure G et al (2016) Role of mRNA structure in the control of protein folding. Nucleic Acid Res 44:10898–10911
Firth AE, Brierley I (2012) Non-canonical translation in RNA viruses. J Gen Virol 93:1385–1409
Firth AE et al (2011) Stimulation of stop codon readthrough: frequent presence of an extended 3' RNA structural element. Nucleic Acid Res 39:6679–6691
Fischlechner M, Donath E (2007) Viruses as building blocks for materials and devices. Angew Chem Int Ed Engl 46:3184–3193
Fredrick K, Ibba M (2010) How the sequence of a gene can tune its translation. Cell 141:227–229
Gale M Jr et al (2000) Translational control of viral gene expression in eukaryotes. Microbiol Mol Biol Rev 64:239–280
Gardin J et al. (2014) Measurement of average decoding rates of the 61 sense codons in vivo. Elife 3:10.7554/eLife.03735
Gerashchenko MV, Gladyshev VN (2017) Ribonuclease selection for ribosome profiling. Nucleic Acids Res 45(2):e6
Gorgoni B et al (2014) Controlling translation elongation efficiency: tRNA regulation of ribosome flux on the mRNA. Biochem Soc Trans 42:160–165. doi:10.1042/BST20130132
Goz E et al (2017) Evidence of translation efficiency adaptation of the coding regions of the bacteriophage lambda. DNA Res
Greenbaum BD et al (2008) Patterns of evolution and host gene mimicry in influenza and other RNA viruses. PLoS Pathog 4:e1000079
Gu W et al (2010) A universal trend of reduced mRNA stability near the translation-initiation site in prokaryotes and eukaryotes. PLoS Comput Biol 2010(6):1–8
Gog JR et al (2007) Codon conservation in the influenza A virus genome defines RNA packaging signals. Nucleic Acid Res 35:1897–1907
Goz E, Tuller T (2015) Widespread signatures of local mRNA folding structure selection in four dengue virus serotypes. BMC Genom 16(Suppl 10):S4
Goz E, Tuller T (2016) Evidence of a direct evolutionary selection for strong folding and mutational robustness within HIV coding regions. J Comput Biol 23:641–650
Goz E, Tuller T (2017) Widespread selection for complex patterns of synonymous information in viral coding regions. In Review
Gu W et al (2004) Analysis of synonymous codon usage in SARS Coronavirus and other viruses in the Nidovirales. Virus Res 101:155–161
Gustafsson C et al (2004) Codon bias and heterologous protein expression. Trends Biotechnol 22:346–353
Haldane A et al (2014) Biophysical fitness landscapes for transcription factor binding sites. PLoS Comput Biol 10:e1003683
Holland JJ (2012) Genetic diversity of RNA viruses. Springer Science & Business Media
Holmes EC (2009) The evolution and emergence of RNA viruses. Oxford University Press Inc, New York
Horvath P, Barrangou R (2010) CRISPR/Cas, the immune system of bacteria and archaea. Science 327:167–170
Hussmann JA et al (2015) Understanding biases in ribosome profiling experiments reveals signatures of translation dynamics in yeast. PLoS Genet 11:e1005732
Hyde JL et al (2014) A viral RNA structural element alters host recognition of nonself RNA. Science 343:783–787

Ingolia NT et al (2009) Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. Science 324:218–223
Jenkins GM et al (2001) Evolution of base composition and codon usage bias in the genus Flavivirus. J Mol Evol 52:383–390
Jia R et al (2009) Analysis of synonymous codon usage in the UL24 gene of duck enteritis virus. Virus Genes 38:96–103
Kazazian HH Jr (2004) Mobile elements: drivers of genome evolution. Science 303:1626–1632
Konur S et al (2016) An integrated in silico simulation and biomatter compilation approach to cellular computation. In: Adamatzky A. (ed) Advances in unconventional computing, vol 23. pp. 655–676
Kozak M (1986) Point mutations define a sequence flanking the AUG initiator codon that modulates translation by eukaryotic ribosomes. Cell 44:283–292
Kramer G et al (2009) The ribosome as a platform for co-translational processing, folding and targeting of newly synthesized proteins. Nat Struct Mol Biol 16:589–597
Lauring AS et al (2010) Rationalizing the development of live attenuated virus vaccines. Nat Biotechnol 28:573–579
Lauring AS et al (2012) Codon usage determines the mutational robustness, evolutionary capacity, and virulence of an RNA virus. Cell Host Microbe 12:623–632
Lauring AS et al (2013) The role of mutational robustness in RNA virus evolution. Nat Rev Microbiol 11:327–336
Liu YS et al (2010) Analysis of synonymous codon usage in porcine reproductive and respiratory syndrome virus. Infect Genet Evol 10:797–803
Liu YS et al (2011) The characteristics of the synonymous codon usage in enterovirus 71 virus and the effects of host on the virus in codon usage pattern. Infect Genet Evol 11:1168–1173
Liu XS et al (2012) Patterns and influencing factor of synonymous codon usage in porcine circovirus. Virol J 9:68
Liu X et al (2013) High-resolution view of bacteriophage lambda gene expression by ribosome profiling. Proc Natl Acad Sci USA 110:11928–11933
Lopez-Lastra M et al (2010) Translation initiation of viral mRNAs. Rev Med Virol 20:177–195
Lobo FP et al (2009) Virus-host coevolution: common patterns of nucleotide motif usage in Flaviviridae and their hosts. PLoS One 4:e6282
Lucks JB et al (2008) Genome landscapes and bacteriophage codon usage. PLoS Comput Biol 4: e1000001
Marraffini LA (2015) CRISPR-Cas immunity in prokaryotes. Nature 526:55–61
Martrus G et al (2013) Changes in codon-pair bias of human immunodeficiency virus type 1 have profound effects on virus replication in cell culture. Retrovirology 10:78
Ma MR et al (2011) The characteristics of the synonymous codon usage in hepatitis B virus and the effects of host on the virus in codon usage pattern. Virol J 8:544
Medrano-Soto A et al (2004) Successful lateral transfer requires codon usage compatibility between foreign genes and recipient genomes. Mol Biol Evol 21:1884–1894
Michely S et al (2013) Evolution of codon usage in the smallest photosynthetic eukaryotes and their giant viruses. Genome Biol Evol 5:848–859
Morgunov AS, Babu MM (2014) Optimizing membrane-protein biogenesis through nonoptimal-codon usage. Nat Struct Mol Biol 21:1023–1025. doi:10.1038/nsmb.2926
Mueller S et al (2006) Reduction of the rate of poliovirus protein synthesis through large-scale codon deoptimization causes attenuation of viral virulence by lowering specific infectivity. J Virol 80:9687–9696
Mueller S et al (2008) Live attenuated influenza virus vaccines by computer-aided rational design virus attenuation by genome-scale changes in codon pair bias. Nat Biotechnol 28:723–726
Nogales A et al (2014) Influenza A virus attenuation by codon deoptimization of the NS gene for vaccine development. J Virol 88:10525–10540
Novella IS et al (2004) Positive selection of synonymous mutations in vesicular stomatitis virus. J Mol Biol 342:1415–1421

Novoa EM, Ribas de Pouplana L (2012) Speeding with control: codon usage, tRNAs, and ribosomes. Trends Genet 28:574–81. doi:10.1016/j.tig.2012.07.006. Epub 23 Aug 2012
Pan W et al (2016) Online model selection for synthetic gene networks. CDC 776–782
Pavesi A et al (2013) Viral proteins originated de novo by overprinting can be identified by codon usage: application to the "gene nursery" of Deltaretroviruses. PLoS Comput Biol 9:e1003162
Pechmann S, Frydman J (2013) Evolutionary conservation of codon optimality reveals hidden signatures of cotranslational folding. Nat Struct Mol Biol 20:237–243. doi:10.1038/nsmb.2466. Epub 23 Dec 2012
Pinto RM et al (2007) Codon usage and replicative strategies of hepatitis A virus. Virus Res 127:158–163
Pride DT et al (2006) Evidence of host-virus co-evolution in tetranucleotide usage patterns of bacteriophages and eukaryotic viruses. BMC Genomics 7:8
Quax TE et al (2015) Codon bias as a means to fine-tune gene expression. Mol Cell 59:149–161
Raveh A et al (2016) A model for competition for ribosomes in the cell. J R Soc Interface 13:20151062
Reuveni S et al (2011) Genome-scale analysis of translation elongation with a ribosome flow model. PLoS Comput Biol 1–18
Rocha EP (2004) Codon usage bias from tRNA's point of view: redundancy, specialization, and efficient decoding for translation optimization. Genome Res 14: 2279–2286. Epub 12 Oct 2004
Rohde W et al (1994) Plant viruses as model systems for the study of non-canonical translation mechanisms in higher plants. J Gen Virol 75(Pt 9):2141–2149
Roller M et al (2013) Environmental shaping of codon usage and functional adaptation across microbial communities. Nucleic Acid Res 41:8842–8852
Roychoudhury S, Mukherjee D (2010) A detailed comparative analysis on the overall codon usage pattern in herpesviruses. Virus Res 148:31–43
Sabi R et al (2016) stAIcalc: tRNA adaptation index calculator based on species-specific weights. Bioinformatics
Sanassy D et al (2015) Meta-stochastic simulation of biochemical models for systems and synthetic biology. ACS Synth Biol 4:39–47
Sau K, Deb A (2009) Temperature influences synonymous codon and amino acid usage biases in the phages infecting extremely thermophilic prokaryotes. Silico Biol 9:1–9
Sau K et al (2005a) Factors influencing the synonymous codon and amino acid usage bias in AT-rich Pseudomonas aeruginosa phage PhiKZ. Acta Biochim Biophys Sin (Shanghai) 37:625–633
Sau K et al (2005b) Synonymous codon usage bias in 16 Staphylococcus aureus phages: implication in phage therapy. Virus Res 113:123–131
Sau K et al (2007) Studies on synonymous codon and amino acid usage biases in the broad-host range bacteriophage KVP40. J Microbiol 45:58–63
Sauna ZE, Kimchi-Sarfaty C (2013) Understanding the contribution of synonymous mutations to human disease. Nat Rev Genet 12:683–691
Schoech AP, Zabet NR (2014) Facilitated diffusion buffers noise in gene expression. Phys Rev E Stat Nonlin Soft Matter Phys 90:032701
Shackelton LA et al (2006) Evolutionary basis of codon usage and nucleotide composition bias in vertebrate DNA viruses. J Mol Evol 62:551–563
Sharp PM, Li WH (1987) The codon adaptation index—a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acid Res 15:1281–1295
Sharp PM et al (2005) Variation in the strength of selected codon usage bias among bacteria. Nucleic Acid Res 33:1141–1153. Print 2005
Sin C et al (2016) Quantitative assessment of ribosome drop-off in E. coli. Nucleic Acid Res 44:2528–2537
Soucy SM et al (2015) Horizontal gene transfer: building the web of life. Nat Rev Genet 16:472–482
Su MW et al (2009) Categorizing host-dependent RNA viruses by principal component analysis of their codon usage preferences. J Comput Biol 16:1539–1547

Supek F (2016) The code of silence: widespread associations between synonymous codon biases and gene function. J Mol Evol 82:65–73
Tao P et al (2009) Analysis of synonymous codon usage in classical swine fever virus. Virus Genes 38:104–112
Terns MP, Terns RM (2011) CRISPR-based adaptive immune systems. Curr Opin Microbiol 14:321–327
Thommen M et al (2016) Co-translational protein folding: progress and methods. Curr Opin Struct Biol 42:83–89
Tsai CT et al (2007) Analysis of codon usage bias and base compositional constraints in iridovirus genomes. Virus Res 126:196–206
Tuller T et al (2010a) Translation efficiency is determined by both codon bias and folding energy. Proc Natl Acad Sci USA 107:3645–3650
Tuller T et al (2010b) An evolutionarily conserved mechanism for controlling the efficiency of protein translation. Cell 141:344–354
Tuller T (2011) Codon bias, tRNA pools, and horizontal gene transfer. Mob Genet Elem
Tuller T (2012) The effect of codon usage on the success of horizontal gene transfer. In: In lateral gene transfer in evolution
Tuller T et al (2011a) Composite effects of gene determinants on the translation speed and density of ribosomes. Genome Biol 12:R110
Tuller T et al (2011b) Association between translation efficiency and horizontal gene transfer within microbial communities. Nucleic Acid Res 22
Tuller T, Zur H (2015) Multiple roles of the coding sequence 5' end in gene expression regulation. Nucleic Acid Res 43:13–28
Tulloch F et al (2014) RNA virus attenuation by codon pair deoptimisation is an artefact of increases in CpG/UpA dinucleotide frequencies. Elife 3:e04531
van Hemert FJ et al (2007) Host-related nucleotide composition and codon usage as driving forces in the recent evolution of the Astroviridae. Virology 361:447–454
Watts JM et al (2009) Architecture and secondary structure of an entire HIV-1 RNA genome. Nature 460:711–716
Welch M et al (2009) Design parameters to control synthetic gene expression in Escherichia coli. PLoS ONE 4:1–10
Wimmer E et al (2009) Synthetic viruses: a new opportunity to understand and prevent viral disease. Nat Biotechnol 27:1163–1172
Wong EH et al (2010) Codon usage bias and the evolution of influenza A viruses. Codon usage biases of influenza virus. BMC Evol Biol 10:253
Wright F (1990) The 'effective number of codons' used in a gene. Gene 87:23–29
Wu H et al (2016) Multiensemble Markov models of molecular thermodynamics and kinetics. Proc Natl Acad Sci USA 113:E3221–E3230
Yang JR et al (2014) Codon-by-codon modulation of translational speed and accuracy via mRNA folding. PLoS Biol 12:e1001910. doi:10.1371/journal.pbio.1001910. eCollection Jul 2014
Yofe I et al (2014) An intronic code for gene expression regulation in S.cerevisiae. PLoS Genet 10: e1004407
Zafrir Z, Tuller T (2015a) Selection for nucleotide composition adjacent to intronic splice sites improves splicing efficiency via its effect on pre-mRNA local folding in fungi. RNA 21: 1704–1718
Zafrir Z, Tuller T (2015b) Nucleotide sequence composition adjacent to intronic splice sites improves splicing efficiency via its effect on pre-mRNA local folding in fungi. RNA 21: 1704–1718
Zafrir Z, Tuller T (2017) Unsupervised detection of regulatory gene expression information in different genomic regions enables gene expression ranking. BMC Bioinform 18:77
Zafrir Z et al (2016) Selection for reduced translation costs at the intronic 5' end in fungi. DNA Res 23:377–394
Zhang Y et al (2011) Analysis of synonymous codon usage in hepatitis A virus. Virol J 8:174

Zhang Z et al (2013) Analysis of synonymous codon usage patterns in torque tenosus virus 1 (TTSuV1). Arch Virol 158:145–154
Zhao KN et al (2005) Gene codon composition determines differentiation-dependent expression of a viral capsid gene in keratinocytes in vitro and in vivo. Mol Cell Biol 25:8643–8655
Zhao S et al (2008) Analysis of synonymous codon usage in 11 human bocavirus isolates. Biosystems 92:207–214
Zhong J et al (2007) Mutation pressure shapes codon usage in the GC-Rich genome of foot-and-mouth disease virus. Virus Genes 35:767–776
Zhou JH et al (2010) Analysis of synonymous codon usage in foot-and-mouth disease virus. Vet Res Commun 34:393–404
Zhou JH et al (2013) The effects of the synonymous codon usage and tRNA abundance on protein folding of the 3C protease of foot-and-mouth disease virus. Infect Genet Evol 16:270–274
Zimmer C (2011) A Planet of Viruses. University Of Chicago Press, Chicago
Zur H, Tuller T (2012) Strong association between mRNA folding strength and protein abundance in S. cerevisiae. EMBO Rep
Zur H, Tuller T (2013) New universal rules of Eukaryotic translation initiation fidelity. PLoS Comput Biol 9:e1003136
Zur H, Tuller T (2015) Exploiting hidden information interleaved in the redundancy of the genetic code without prior knowledge. Bioinformatics 31:1161–1168
Zur H, Tuller T (2016) Predictive biophysical modeling and understanding of the dynamics of mRNA translation and its evolution. Nucleic Acid Res 44:9031–9049

Self and Nonself from a Genomic Perspective: Transposable Elements

Marie Fablet, Judit Salces-Ortiz, Bianca Fraga Menezes, Marlène Roy and Cristina Vieira

Abstract Transposable elements (TEs) are sequences that can move and multiply along the chromosomes. Considered for a long time as genomic parasites, they are now acknowledged as key players of genome function and evolution. Accordingly, the presence of TEs in a genome may affect the chromatin structure of the regions in which they are inserted. TEs allow us to revisit self and nonself distinction at the genomic level, through the complex relationships they display with the genome and the epigenome and their interaction with the environment.

1 Introduction

Genomes are not mere strings of genes. They are also scattered with—apparently—dispensable sequences, which we know as transposable elements (TEs) due to their ability to move from one locus to another along the chromosomes. They were discovered in the 1950s thanks to the pioneering work of Barbara McClintock (1950), who was awarded the Nobel Prize in 1983. The historic study of TE research is a textbook case of how ideas evolve within the scientific community. First considered as "junk DNA," devoid of scientific interest, they are now rec-

M. Fablet (✉) · C. Vieira (✉)
43 boulevard du 11 novembre 1918, 69622 Villeurbanne Cedex, France
e-mail: marie.fablet@univ-lyon1.fr

C. Vieira
e-mail: cristina.vieira@univ-lyon1.fr

M. Fablet · J. Salces-Ortiz · B.F. Menezes · M. Roy · C. Vieira
Laboratoire de Biométrie et Biologie Evolutive, UMR CNRS 5558, Université Lyon 1, Université de Lyon, Villeurbanne, France
e-mail: judit.salces-ortiz@univ-lyon1.fr

B.F. Menezes
e-mail: bianca.menezes@univ-lyon1.fr

M. Roy
e-mail: marlene.roy@etu.univ-lyon1.fr

P. Pontarotti (ed.), *Evolutionary Biology: Self/Nonself Evolution, Species and Complex Traits Evolution, Methods and Concepts*,
DOI 10.1007/978-3-319-61569-1_6

ognized as key functional and evolutionary components of genomes (Biémont 2010).

TEs—also known as mobile elements—are generally shorter than 10 kb and are able to mobilize, i.e., to *transpose*, and multiply within genomes thanks to the enzymatic machinery they encode. TE copies resulting from transposition events of a given TE share sequence similarity with each other and constitute TE families. Based on their transposition mechanisms, TE families are broadly organized into two classes (a more comprehensive classification is proposed in Wicker et al. (2007)) (Fig. 1). Class II TEs are also called *transposons* and are mobilized through a *cut and paste* mechanism. They encode one enzyme called *transposase* and are bordered by Terminal Inverted Repeats (TIRs). Class I TEs are termed *retrotransposons* and get mobilized through a *copy and paste* scenario via an RNA intermediate. Two subclasses of retrotransposons are distinguished. LTR retrotransposons are bordered by direct Long Terminal Repeats (LTRs), which include regulatory sequences, and display three Open Reading Frames (ORFs): *gag*—encodes proteins of the capsid, *pol*—encodes the retrotransposition enzymes, i.e., reverse transcriptase, RNase-H, integrase, protease, and sometimes *env*—encodes proteins involved in the formation of an envelope. Endogenous retroviruses (ERVs) are sequences of viral origin integrated into the genome and transmitted from parent to progeny, like genes. They are included into the LTR retrotransposon subclass. The other retrotransposon subclass is made of non-LTR retrotransposons. Among them, LINEs (Long Interspersed Nuclear Elements) display two ORFs, sharing similarity with *gag* and *pol*, and are terminated by a polyA tail, and SINEs (Short Interspersed Nuclear Elements) are non-autonomous elements, meaning they do not encode their own machinery and are mobilized by LINEs. SINEs are non-coding and result from accidental transposition of RNA polymerase III transcripts (such as transfer RNAs) (Wicker et al. 2007).

What makes the transition from a genomic parasite to a domesticated sequence? This is a sensitive question around TEs. We intend to answer it throughout this

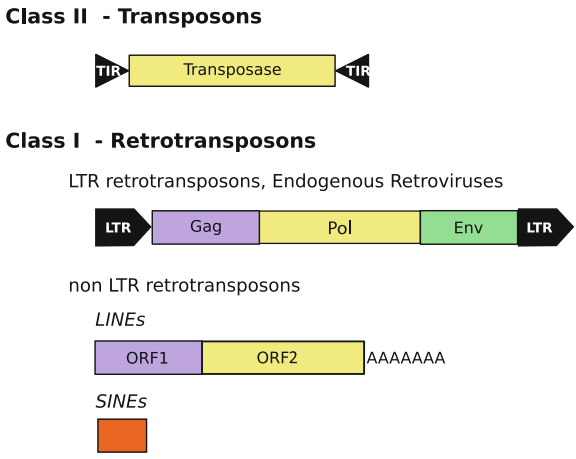


Fig. 1 Structures of the archetypes of TE classes and subclasses. TIR: Terminal Inverted Repeat. LTR: Long Terminal Repeat. ORF: Open Reading Frame. SINEs are short non-coding sequences that derive from RNA polymerase III transcripts

chapter by examining ways the host genome controls TEs, the effects of TE insertions, and the way they can be fully integrated into the host genome functioning and evolution. The long-lasting interaction between genomes and TEs makes the latter considered sometimes as genomic nonself and sometimes as genomic self.

2 TEs as Nonself: Genomes Fight Against TEs

TEs are a threat to genome integrity because of both their mobility and their repetitive nature. Indeed, TEs may be harmful when a new insertion disrupts the coding sequence of a gene or regulatory sequences. They are also harmful when ectopic recombination occurs between two distant copies of the same TE family (Petrov et al. 2003), which leads to deleterious chromosomal rearrangements. Population genetic models were proposed and considered the deleterious effects of TE insertions and the deleterious consequences of genomic rearrangements (Charlesworth and Langley 1989; Langley et al. 1988). The elimination of TEs is thus expected in the long-term process of evolution. However, what we know from genomic analyses is that is it generally not the case. The resolution of this conundrum was allowed by the discovery of defense mechanisms against TE expression that can act rapidly, at a shorter evolutionary timescale. We will dedicate this section to these mechanisms.

2.1 Genomes Control TEs Through the PiRNA Pathway

We have known for almost 15 years that TEs are controlled by RNA interference mechanisms via the Piwi-interacting RNA (piRNA) class of small RNAs, which were originally called rasiRNAs (Aravin et al. 2007; Klenov et al. 2007; Saito et al. 2006; Vagin et al. 2006). Most of our knowledge comes from fly studies which will be presented below, but similar mechanisms also occur in other taxa.

piRNAs are 23–30-nt-long single-stranded RNA molecules (Fig. 2). They are loaded onto proteins of the PIWI subfamily (Argonaute family): Aubergine (Aub), Argonaute 3 (Ago3), and Piwi (Meister 2013; Siomi et al. 2011), which display RNase-H activity (Jinek and Doudna 2009). In *Drosophila melanogaster* ovaries, antisense piRNAs are loaded onto Aub and target sense TE transcripts. The latter are then sliced into sense piRNAs, which are loaded onto Ago3 and target antisense TE transcripts. This leads to the so-called *ping-pong loop* (Brennecke et al. 2007), which occurs in the *nuage*, a dense cytoplasmic region located near the nucleus. Such piRNAs are termed secondary piRNAs. Secondary piRNAs loaded onto Ago3 show a 10 nt sequence overlap at their 5′ end with secondary piRNAs loaded onto Aub, which is called the *ping-pong signature* (Brennecke et al. 2007).

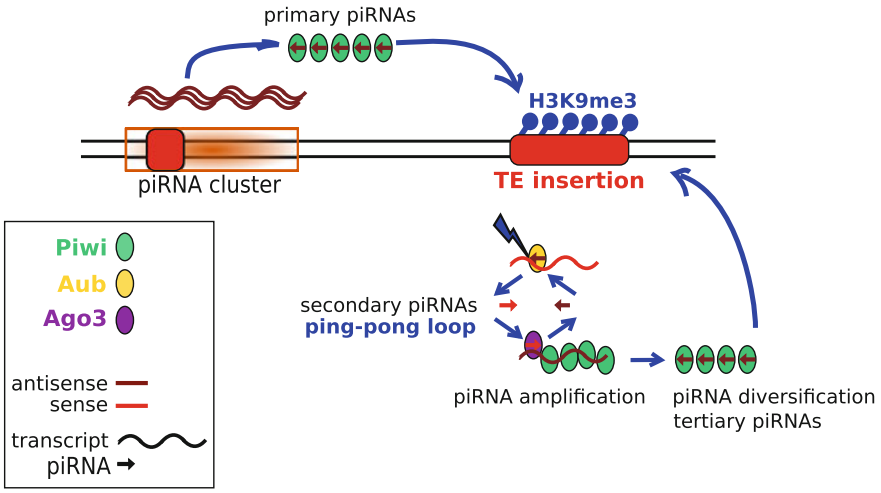


Fig. 2 piRNA pathway. Antisense piRNAs are loaded onto Aub and target sense TE transcripts. The latter are then sliced into sense piRNAs, which are loaded onto Ago3 and target antisense TE transcripts. This leads to the *ping-pong loop*, producing secondary piRNAs. Piwi-bound piRNAs go to the nucleus and promote the transcription of piRNA clusters. This generates primary piRNAs, which target TE insertion sites triggering their heterochromatinization (H3K9me3 histone marks). In addition, tertiary piRNAs result from Piwi-induced phasing and allow the diversification of piRNA sequences

Antisense piRNAs may also be loaded onto Piwi and go back to the nucleus. They then promote the transcription of particular loci that are called piRNA clusters. These loci are heterochromatic and, in *Drosophila*, are predominantly filled with TE sequences. The best known *Drosophila* cluster is *flamenco* (X chromosome, ~180 kb), which controls the *gypsy* endogenous retrovirus (Pélisson et al. 1994; Prud'homme et al. 1995) among others. The transcription of piRNA clusters leads to the production of long precursors that are cleaved into primary piRNAs. These are loaded onto Piwi and target TE insertion sites triggering their heterochromatinization via the recruitment of histone-modifying enzymes (Akkouche et al. 2013; Le Thomas et al. 2013; Rozhkov et al. 2013a; Sienski et al. 2012). Such a mechanism ensures TE control both, at the transcriptional and posttranscriptional levels.

In addition, it was recently discovered that Piwi-bound piRNAs displayed sequence *phasing*, meaning that the 3′ ends of such piRNAs are immediately followed by the 5′ end of the next Piwi-bound piRNAs (Han et al. 2015; Mohn et al. 2015). Such phased piRNAs are considered tertiary piRNAs (Siomi and Siomi 2015) and go back to the nucleus and achieve transcriptional TE silencing. These phased Piwi-bound piRNAs are mostly generated from Ago3-bound secondary piRNAs (Wang et al. 2015) (Fig. 2). While the ping-pong loop provides secondary piRNAs in high amounts, those phased tertiary piRNAs allow increases in the diversity of target sequences (Han et al. 2015).

In *Drosophila* ovaries, secondary piRNAs are produced only in germline cells while primary piRNAs are produced both in germline and somatic cells of the ovary (Malone et al. 2009). Which piRNA cluster is transcribed also depends on the cell type. Feeding of the ping-pong loop can be provided by the maternal transmission of Aub-bound secondary piRNAs. In addition, while it was considered for a long time that the *Drosophila* piRNA pathway took place exclusively in the ovaries, it is now known that an efficient secondary piRNA pathway also takes place in the fat body (Jones et al. 2016), widening the potential roles of piRNA regulation.

2.2 Genomes and TEs: A Red Queen Relationship

It is common today to consider the piRNA pathway as an immune pathway acting at the genomic level (Aravin et al. 2007; Fablet 2014; Malone and Hannon 2009; Senti and Brennecke 2010; Siomi et al. 2011). Such a parallel is even more relevant when performing an evolutionary analysis of the genes involved in this system. Classical host-pathogen interactions are often described using the *Red Queen* metaphor, after Lewis Carroll's novel (van Valen 1973). It illustrates the fact that both partners of this antagonistic relationship are constantly evolving. For instance, concerning antiviral immunity in *Drosophila*, which is achieved through the siRNA pathway, Obbard et al. (2006, 2009) showed that genes involved in the siRNA pathway displayed strong signatures of recurrent positive selection, which is distinctive of rapid, constant evolution. It is noteworthy that genes involved in the piRNA pathway (GIPPs) also showed strong, significant signatures of recurrent positive selection in the same study (Obbard et al. 2009). Other works, using different samples and methods, confirmed this general trend of rapid sequence evolution for GIPPs (Fablet et al. 2014; Kolaczkowski et al. 2011; Obbard et al. 2009b). In addition, GIPPs also display rapid expression rate evolution (Fablet et al. 2014). All these elements suggest that the Red Queen metaphor also applies to the genome (at the level of GIPPs) versus TE interaction, meaning that the relationship that host genomes develop toward TEs could be considered host-pathogen-like, at least to some extent. However, the molecular mechanisms of the interaction and the rapid evolution are still a topic of further investigation.

Some authors even go further and apply the hygiene hypothesis to the case of TEs and GIPPs. The hygiene hypothesis states that, in *Homo sapiens* populations, the increase of pathogen abundance during the neolithic period constituted a selective pressure for an efficient immune system, which is now assumed to be responsible for an increase in the incidence of autoimmune diseases in the aseptic-ized post-industrial era (Rook 2012; Strachan 1989). Blumenstiel et al. (2016) propose that genomic autoimmunity could result from off-targets of a too efficient piRNA pathway and would reinforce the evolutionary tension between GIPPs and

the genome. Nevertheless, there is still a need for empirical data to sustain this hypothesis.

2.3 *Acquired Genomic Immunity Is Provided by PiRNA Clusters*

The piRNA pathway is considered as an acquired immune pathway since it provides protection only against those TE families that have one copy inserted into a piRNA cluster (Malone et al. 2009; Zanni et al. 2013). Indeed, a clear causal relationship between the presence of the TE family within a cluster and the repression of its mobility was demonstrated in the case of the *gypsy* and *ZAM* retrotransposons, which are controlled by the *flamenco* cluster (Zanni et al. 2013). Zanni et al. (2013) also suggested that although their chromosomal locations are conserved across closely related species (Malone et al. 2009), piRNA clusters such as *flamenco* are highly dynamic regions. Indeed, insertions, deletions, and rearrangements frequently occur, even at the intraspecific scale. Again, this rapid evolution fits with the above-mentioned Red Queen hypothesis. As a consequence, the dynamics of transposition can also be rapidly evolving, since a specific TE family can reacquire activity if ever the piRNA cluster looses the corresponding insertion.

TEs are a threat to the genome at the time they invade it, for instance, after a horizontal transfer event or after reactivation of an old family. Theoretical work on *Drosophila* indicates that they remain a threat up until a copy inserts within a piRNA cluster, which triggers control through the piRNA pathway (Lu and Clark 2010). Indeed, once the piRNA pathway controls a TE family, the different copies of these families are neither able to transpose—because they are transcriptionally and posttranscriptionally silenced by piRNAs—nor involved in ectopic recombination, since recombination is less likely to occur in heterochromatinized regions. Khurana et al. (2011) experimentally demonstrated that *P* element introduction into a naive genome first leads to a high transpositional, potentially deleterious activity. However, as the host ages, control of the *P* element can be established. The authors propose that this control is possible due to the insertion of a *P* copy within a piRNA cluster (Khurana et al. 2011) at some point along the life of the host individual. Similarly, Rozhkov et al. (2013b) introduced a *Penelope* TE into a naive *D. melanogaster* genome and found that after some generations, *Penelope* control was established due to insertions into piRNA clusters.

We propose that a TE family is recognized as genetically nonself as long as it is not inserted into a piRNA cluster. Once silenced, the TE family can enter a new relationship with the host genome, eventually leading to genomic assimilation into self.

3 TEs as Self: Genomes Recruit TEs

3.1 TEs as Raw Material for Genome Evolution

TE mobilization is known to be responsible for deleterious mutations or to be involved in the first steps of cancer and other diseases in humans (Belancio et al. 2009; Hancks and Kazazian 2016; Hancks and Kazazian 2012; Kim et al. 2007). They can also be at the origin of phenotypic variation, without deleterious effects, as for instance, body size diversity in dogs (Sutter et al. 2007). The historic example is that of Barbara McClintock, which led her to the discovery of TEs: Variation in pigmentation patterns in maize kernels is due to the activity of the *Ac* and *Ds* transposons (McClintock 1950). Other examples of color variation due to transposable elements have been recorded in plants. For instance, color variation in maize pericarp can be due to the insertion of an *Ac* transposon into the *p1* gene (Zhang and Peterson 2005), and petal color variation in *Anthirrhinum* is associated with different insertions of the *Tam* transposon on both sides of the *nivea* gene, responsible for flower pigmentation (Uchiyama et al. 2013). In animals, TE insertions are associated with coat color variability in dogs and cats (Clark et al. 2006; David et al. 2014), or the classical example of the agouti color in mice (Morgan et al. 1999).

TE mobilization contributes to raw material for genome evolution. While TE insertions into genes lead to deleterious mutations, most of the time, the insertion of a TE in the vicinity of genes will increase the possibilities and modalities of gene expression. For example, the insertion of a *Doc* TE within the *CHKov* gene leads to alternative transcripts that provide *Drosophila* with pesticide resistance (Aminetzach et al. 2005). On a more global scale, Vieira et al. (1999) observed an increase of the global TE content in derived strains of *D. melanogaster* as strains were further away from the ancestral, African area. They proposed that such an increase in TE amounts could be adaptive in the frame of the colonization of new environments. As reviewed by Casacuberta and González (2013), TEs may play major roles in environmental adaptation. Similarly, but at a shorter timescale, Perrat et al. (2013) found that transposition occurred in memory-relevant neurons in the *Drosophila* brain and proposed that this genomic heterogeneity was a conserved feature of the brain potentially producing behavioral variability between individuals. Prior work on rodents and humans (Coufal et al. 2009; Muotri et al. 2005) also suggested that transposon-mediated genomic heterogeneity may be a conserved characteristic of certain neurons.

TEs are made up of coding and regulatory sequences. Long-term, safe relationships between the host genome and TEs may lead to the recruitment of TE sequences, so that such insertions are considered as *domesticated* (Miller et al. 1999; Volff 2006). For instance, at least 25% of human gene promoters derive from TEs (Jordan et al. 2003). The most famous cases of domestication are *rag1* and *rag2* genes—involved in the vertebrate V(D)J recombination giving birth to immune cell receptors (Agrawal et al. 1998), *syncytin* genes—remnants of an ERV

env gene, allowing placenta formation (Blaise et al. 2003; Mi et al. 2000), and *TART* and *HeT-A*—non-LTR retrotransposons insuring telomere maintenance in *Drosophila* (Biessmann et al. 1990; Levis et al. 1993). New examples continue to be published. For instance, the extensively taught industrial melanism mutation of peppered moth is now known to be due to a TE insertion (Van't Hof et al. 2016).

3.2 Genomes Recruit TE Regulatory Sequences

At the time they invade a genome, most TEs possess all necessary sequences for their expression and mobilization. For instance, they all display promoter sequences —even the non-autonomous SINEs—and transcription termination signals. Most TEs also contain transcription factor binding sites (Chuong et al. 2016; Zemojtel and Vingron 2012). As mentioned previously, TEs are the targets of silencing mechanisms from chromatin modifications to DNA methylation (Slotkin and Martienssen 2007) that minimize transposition potential. Such epigenetic modifications also modulate the expression of neighboring genes (Hollister and Gaut 2009; Rebollo et al. 2011) by the spreading of chromatin marks. This abundant source of regulatory sequences may be hijacked by the genome for its own purpose. For instance, TEs may be a source of promoter sequences for duplicated genes (e.g., Fablet et al. 2009). In addition, the fact of having the same repetitive sequences all along the chromosomes may transform TEs into conductors of regulatory networks (Chuong et al. 2017; Rebollo et al. 2012). For instance, Chuong et al. (2016) recently demonstrated that *MER41.AIM2*, an old human endogenous retrovirus, behaved as an enhancer and provided susceptibility to interferon in the inflammatory response to infection.

The number of reported cases of TEs being involved in gene regulation is exponentially increasing. What we mention in this chapter is merely the tip of the iceberg. Among the most remarkable examples, TEs may be involved in sex determination. In melon, a TE insertion in the promoter of a transcription factor spreads DNA methylation and results in the transition from male to female flowers (Martin et al. 2009). The determination of new sex chromosomes was also shown to be driven by TEs and the epigenetic marks that are associated with the neo-Y chromosome in *Drosophila miranda* (Zhou et al. 2013).

3.3 Genomes Recruit TE Coding Sequences

While it is rather common to find TEs recruited for their regulatory regions, it is less frequent, although more spectacular, to encounter TEs recruited for their coding sequences. The V(D)J recombination system in immune cells of vertebrates has been suggested for a long time to result from the domestication of a putative *RAG* transposon (Agrawal et al. 1998). It is only very recently that an active *ProtoRAG*

TE was actually found in the lancelet genome and is considered as a relative of the long-sought *RAG* TE (Huang et al. 2016).

The mouse *Fv1* gene is involved in virus resistance and derives from a TE *gag* gene (Yan et al. 2009). *Gag* genes of LTR retrotransposons and endogenous retroviruses encode capsid proteins, which, in the case of *Fv1*, interacts with exogenous virus capsid proteins, blocking progression of the viral cycle (Hilditch et al. 2011).

Hoen and Bureau (2015) performed a comprehensive search of novel genes derived from TEs in the *Arabidopsis thaliana* model plant genome and found dozens of such genes. This particular new relationship between the genome and TEs implies that each of these domesticated TEs had lost its mobilization ability and is now expressed—while non-domesticated TEs are generally silenced.

Kokošar and Kordiš (2013) thoroughly studied such phenomena in mammalian genomes and found that it had been particularly frequent in the ancestor of placental mammals. They propose a scenario for the transition from TEs to domesticated genes. They suggest that nucleotide changes first allow neofunctionalization, along with exonization of TE domains. The recruitment of *cis*-regulatory regions, such as promoters, is the next compulsory step in the acquisition of functionality. It can be achieved through different mechanisms such as the recruitment of bidirectional promoters, or promoter capture via the evolution of 5′-UTR exon/intron structures (Fablet et al. 2009; Kaessmann et al. 2009; Kokošar and Kordiš 2013).

4 TEs as Nonself Sensors

From genomic parasites to genomic full components, from genomic nonself to genomic self, TEs present complex relationships with the host genome. TEs have been shaping genomes for millions of years, and they may play a key role in the evolution of gene regulatory networks. This may explain the wide range of rapid phenotypic evolution and origin of morphological novelties found within and among populations (Cowley and Oakey 2013). In that sense, TEs are eventually used by the genomes as nonself sensors.

4.1 TEs as Sensors of Divergent Genomes

In the 1960s and 1970s, studies on *D. melanogaster* populations revealed the phenomenon of *hybrid dysgenesis*, which refers to aberrant phenotypic traits observed in the F1 of crosses between particular strains or natural populations. This was attributed to differences in TE content between the parental lines (Kidwell 1977; Picard 1976). Owing to the very recent development of the epigenomic field, we now know that the observed disruption of genomic stability in hybrids is the consequence of a high rate of TE mobilization, which itself is the result of the destabilization of epigenetic regulatory networks (Chambeyron et al. 2008; Jensen et al. 2008; Todeschini et al. 2010).

Since these pioneering studies of hybrid dysgenesis in *Drosophila*, other cases of TE reactivation during divergent crosses were recorded in other species of *Drosophila*, mice, and wallabies (Labrador et al. 1999; Lopez-Maestre et al. 2017; Metcalfe et al. 2007; O'Neill et al. 1998; Vela et al. 2014). The challenge is now to understand whether TEs are a primary contribution to population isolation, as for the *P* element in *D. melanogaster*, which was acquired by horizontal transfer, or whether they are only participating in the hybrid incompatibility once speciation has occurred (Craddock 2016; Rebollo et al. 2010). Both processes are probably involved. For example, when crossing closely related *Drosophila* species such as *D. mojavensis* and *D. arizonae*, a well-suited model to study speciation, authors showed that only very few TEs are misexpressed in hybrids (Carnelossi et al. 2014; Lopez-Maestre et al. 2017). Increasing the evolutionary distance between species will lead to a genome-wide misregulation of TEs, which has been attributed to divergence in the GIPP proteins, that fail to control TEs (García Guerreiro (personal communication), Kelleher et al. 2012). Thus, it appears that TEs are sensitive to genomic divergence and therefore fundamental guardians of genomic self.

4.2 TEs as Sensors of the Abiotic Environment

As previously mentioned, TEs bear transcription binding sites, which can contribute to making them sensitive to the environment. For instance, they harbor heat shock responsive elements (Pietzenuk et al. 2016), ecdysone binding sites (Micard et al. 1988), etc. An interesting example is the *Bari* insertion upstream of *D. melanogaster Jheh2* and *Jheh3* genes that adds antioxidant response elements and provides oxidative stress resistance to individuals bearing it (Guio et al. 2014).

TEs are also the targets of epigenetic modifications, through histone modifications or DNA methylation, depending on the organisms (Lister et al. 2008; Rebollo et al. 2011; Sienski et al. 2012). We may notice that it is a way for the genome to distinguish genomic self from genomic nonself. Although we do not understand the exact molecular mechanisms, we know that epigenetic modifications are sensitive to the environment. This TE/epigenetics/environment tripartite relationship makes TEs important players of evolvability (Fablet and Vieira 2011), promoting phenotypic diversity and environmental adaptation (Song and Cao 2017) (Fig. 3). The agouti coat color of mice is an emblematic illustration. Variation in coat color in mice can be due to variability in the methylation level of a TE insertion upstream of the *A* gene responsible for coat color (Morgan et al. 1999). Waterland and Jirtle (2003) showed that a methyl-rich diet provided to the mother could impact the methylation level of this particular TE and had direct effect on the coat color of the progeny. Evolutionary implications are wide. For instance, we may imagine that, in the case of melon sexual determination mentioned above (Martin et al. 2009), populations placed in environments differing for their methyl richness would display different sex ratios, therefore impacting the effective population size and the efficiency of natural selection.

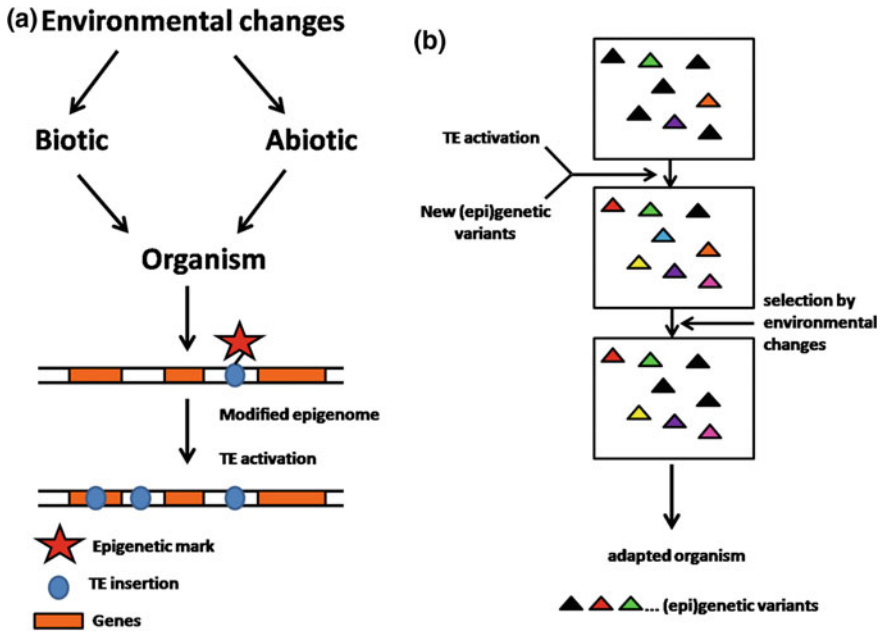


Fig. 3 TE, epigenome, and environment relationship. **a** Environmental changes could be perceived by the organisms through the epigenome and impair the control of transposition, leading to TE-mediated insertional mutagenesis and the induction of phenotypic variability. **b** (Epi)genetic variability increases due to induced TE activity in response to environmental changes on which selection could act providing favorable mutations to promote adaptation

Such sensitivity to environmental conditions is referred to as phenotypic plasticity, i.e., the ability of a genotype to express different phenotypes according to the environment. As explained above, TEs appear as major actors of plasticity, and therefore play a significant role in the adaptive potential of species (Fig. 3) (Casacuberta and González 2013; Fablet and Vieira 2011).

4.3 TEs as Sensors of the Biotic Environment

TEs may even behave as sensors of the biotic environment. For instance, the tobacco *Tnt1* TE is known to be reactivated in case of fungal infection (Melayah et al. 2001). Response to fungal infection in *A. thaliana* is controlled by the targeted demethylation of TE sequences located within promoters of particular genes resulting in downregulation of many stress response genes (Le et al. 2014). This sensitivity to pathogens may even be directly involved in immune pathways. This is the case when TE insertions display antiviral properties, such as the above-mentioned *Fv1* example. This is also true when TEs provide interferon-sensitive sequences, as in the case of *MER41.AIM2* (Chuong et al. 2016). The authors of this study suggest that such an

involvement of TEs from the endogenous retrovirus class in immunity is not unexpected. Indeed, they propose that it results from ancient viral adaptations allowing exploitation of immune pathways.

Such an idea is illustrated in the results of Karijolich et al. (2015). In mice, during murine gammaherpesvirus 68 infection, transcription of the *B2* SINE is enhanced and activates the antiviral NF-κB pathway. However, at the same time, this SINE expression is hijacked by the virus and stimulates viral replication (Karijolich et al. 2015).

5 Conclusion

The relationship between a TE family and the host genome is complex (Fig. 4). Firstly, because it evolves with time from the moment the TE family invades the genome to the moment the interaction is stable. Secondly, as TEs are repeated, the different copies of a given family may not have the same fate. One copy may be

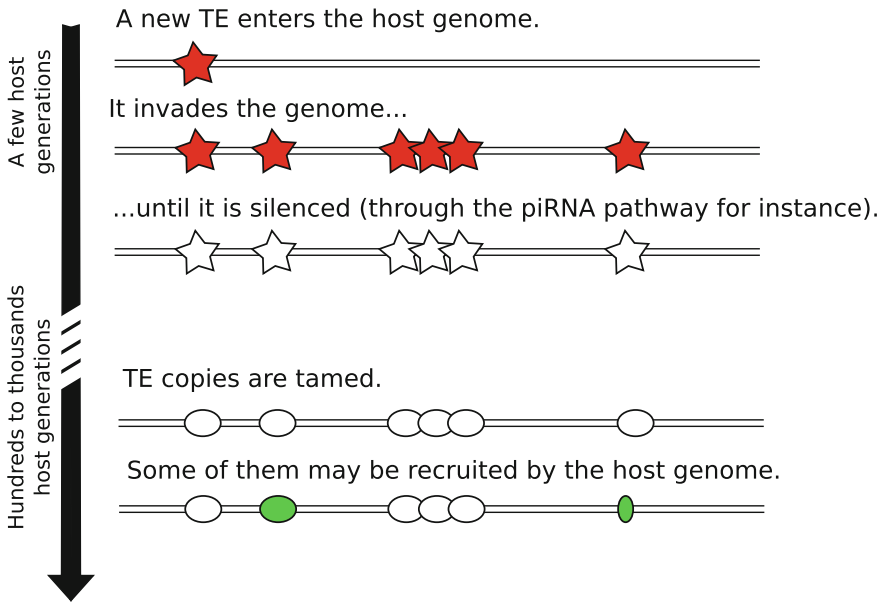


Fig. 4 A model for host genome/TE interactions. A new TE can enter the genome, for instance, due to horizontal transmission or to crossing between divergent genomes, as in the case of hybrid dysgenesis. At that time, it is recognized as genomic nonself. This TE will proliferate into the host genome up until one copy integrates into a piRNA cluster, which sets up transcriptional and posttranscriptional silencing. Controlled TE copies are the targets of heterochromatic marks, so are still identified as distinct from the canonical genome. However, they are no longer harmful. Some TE copies, or at least some parts of TE copies (e.g., transcription factor binding sites), may be recruited by the host genome. These copies are indicated in *green* in the figure. They now fulfill genomic functions. They are said to be domesticated and make part of genomic self

domesticated, while the others are silenced by the host genome. Such ambivalent behavior leads to complex and very sensitive interaction mechanisms—mostly through epigenetic pathways, and allows a wealth of implications of TEs in host genome evolution. Only the comprehensive knowledge of exact TE insertion sites in their epigenomic context will allow us to disentangle self from nonself in each TE-genome relationship.

References

Agrawal A, Eastman QM, Schatz DG (1998) Transposition mediated by RAG1 and RAG2 and its implications for the evolution of the immune system. Nature 394:744–751. doi:10.1038/29457

Akkouche A, Grentzinger T, Fablet M, Armenise C, Burlet N, Braman V, Chambeyron S, Vieira C (2013) Maternally deposited germline piRNAs silence the tirant retrotransposon in somatic cells. EMBO Rep 14:458–464. doi:10.1038/embor.2013.38

Aminetzach YT, Macpherson JM, Petrov DA (2005) Pesticide resistance via transposition-mediated adaptive gene truncation in Drosophila. Science 309:764–767. doi:10.1126/science.1112699

Aravin AA, Hannon GJ, Brennecke J (2007) The Piwi-piRNA pathway provides an adaptive defense in the transposon arms race. Science 318:761–764. doi:10.1126/science.1146484

Belancio VP, Deininger PL, Roy-Engel AM (2009) LINE dancing in the human genome: transposable elements and disease. Genome Med. 1:97. doi:10.1186/gm97

Biémont C (2010) A brief history of the status of transposable elements: from junk DNA to major players in evolution. Genetics 186:1085–1093. doi:10.1534/genetics.110.124180

Biessmann H, Mason JM, Ferry K, d'Hulst M, Valgeirsdottir K, Traverse KL, Pardue ML (1990) Addition of telomere-associated HeT DNA sequences "heals" broken chromosome ends in Drosophila. Cell 61:663–673

Blaise S, de Parseval N, Bénit L, Heidmann T (2003) Genomewide screening for fusogenic human endogenous retrovirus envelopes identifies syncytin 2, a gene conserved on primate evolution. Proc. Natl. Acad. Sci. U. S. A. 100:13013–13018. doi:10.1073/pnas.2132646100

Blumenstiel JP, Erwin AA, Hemmer LW (2016) What Drives Positive Selection in the Drosophila piRNA Machinery? The Genomic Autoimmunity Hypothesis. Yale J. Biol. Med. 89:499–512

Brennecke J, Aravin AA, Stark A, Dus M, Kellis M, Sachidanandam R, Hannon GJ (2007) Discrete small RNA-generating loci as master regulators of transposon activity in Drosophila. Cell 128:1089–1103. doi:10.1016/j.cell.2007.01.043

Carnelossi EAG, Lerat E, Henri H, Martinez S, Carareto CMA, Vieira C (2014) Specific activation of an I-like element in Drosophila interspecific hybrids. Genome Biol. Evol. 6:1806–1817. doi:10.1093/gbe/evu141

Casacuberta E, González J (2013) The impact of transposable elements in environmental adaptation. Mol Ecol 22:1503–1517. doi:10.1111/mec.12170

Chambeyron S, Popkova A, Payen-Groschêne G, Brun C, Laouini D, Pelisson A, Bucheton A (2008) piRNA-mediated nuclear accumulation of retrotransposon transcripts in the Drosophila female germline. Proc. Natl. Acad. Sci. U. S. A. 105:14964–14969. doi:10.1073/pnas.0805943105

Charlesworth B, Langley CH (1989) The population genetics of Drosophila transposable elements. Annu Rev Genet 23:251–287. doi:10.1146/annurev.ge.23.120189.001343

Chuong EB, Elde NC, Feschotte C (2016) Regulatory evolution of innate immunity through co-option of endogenous retroviruses. Science 351:1083–1087. doi:10.1126/science.aad5497

Chuong EB, Elde NC, Feschotte C (2017) Regulatory activities of transposable elements: from conflicts to benefits. Nat Rev Genet 18:71–86. doi:10.1038/nrg.2016.139

Clark LA, Wahl JM, Rees CA, Murphy KE (2006) Retrotransposon insertion in SILV is responsible for merle patterning of the domestic dog. Proc. Natl. Acad. Sci. U. S. A. 103:1376–1381. doi:10.1073/pnas.0506940103

Coufal NG, Garcia-Perez JL, Peng GE, Yeo GW, Mu Y, Lovci MT, Morell M, O'Shea KS, Moran JV, Gage FH (2009) L1 retrotransposition in human neural progenitor cells. Nature 460:1127–1131. doi:10.1038/nature08248

Cowley M, Oakey RJ (2013) Transposable elements re-wire and fine-tune the transcriptome. PLoS Genet 9:e1003234. doi:10.1371/journal.pgen.1003234

Craddock EM (2016) Profuse evolutionary diversification and speciation on volcanic islands: transposon instability and amplification bursts explain the genetic paradox. Biol. Direct 11:44. doi:10.1186/s13062-016-0146-1

David VA, Menotti-Raymond M, Wallace AC, Roelke M, Kehler J, Leighty R, Eizirik E, Hannah SS, Nelson G, Schäffer AA, Connelly CJ, O'Brien SJ, Ryugo DK (2014) Endogenous retrovirus insertion in the KIT oncogene determines white and white spotting in domestic cats. G3 Bethesda Md 4:1881–1891. doi:10.1534/g3.114.013425

Fablet M (2014) Host control of insect endogenous retroviruses: small RNA silencing and immune response. Viruses 6:4447–4464. doi:10.3390/v6114447

Fablet M, Akkouche A, Braman V, Vieira C (2014) Variable expression levels detected in the Drosophila effectors of piRNA biogenesis. Gene 537:149–153. doi:10.1016/j.gene.2013.11.095

Fablet M, Bueno M, Potrzebowski L, Kaessmann H (2009) Evolutionary origin and functions of retrogene introns. Mol Biol Evol 26:2147–2156. doi:10.1093/molbev/msp125

Fablet M, Vieira C (2011) Evolvability, epigenetics and transposable elements. BioMol Concepts 2:333–341

Guio L, Barrón MG, González J (2014) The transposable element Bari-Jheh mediates oxidative stress response in Drosophila. Mol Ecol 23:2020–2030. doi:10.1111/mec.12711

Han BW, Wang W, Li C, Weng Z, Zamore PD (2015) Noncoding RNA. piRNA-guided transposon cleavage initiates Zucchini-dependent, phased piRNA production. Science 348:817–821

Hancks DC, Kazazian HH (2016) Roles for retrotransposon insertions in human disease. Mob DNA 7. doi:10.1186/s13100-016-0065-9

Hancks DC, Kazazian HH Jr (2012) Active human retrotransposons: variation and disease. Curr. Opin. Genet. Dev. Molecular and genetic bases of disease 22:191–203. doi:10.1016/j.gde.2012.02.006

Hilditch L, Matadeen R, Goldstone DC, Rosenthal PB, Taylor IA, Stoye JP (2011) Ordered assembly of murine leukemia virus capsid protein on lipid nanotubes directs specific binding by the restriction factor, Fv1. Proc. Natl. Acad. Sci. U. S. A. 108:5771–5776. doi:10.1073/pnas.1100118108

Hoen DR, Bureau TE (2015) Discovery of novel genes derived from transposable elements using integrative genomic analysis. Mol Biol Evol 32:1487–1506. doi:10.1093/molbev/msv042

Hollister JD, Gaut BS (2009) Epigenetic silencing of transposable elements: a trade-off between reduced transposition and deleterious effects on neighboring gene expression. Genome Res 19:1419–1428. doi:10.1101/gr.091678.109

Huang S, Tao X, Yuan S, Zhang Y, Li P, Beilinson HA, Zhang Y, Yu W, Pontarotti P, Escriva H, Le Petillon Y, Liu X, Chen S, Schatz DG, Xu A (2016) Discovery of an Active RAG Transposon Illuminates the Origins of V(D)J Recombination. Cell 166:102–114. doi:10.1016/j.cell.2016.05.032

Jensen PA, Stuart JR, Goodpaster MP, Goodman JW, Simmons MJ (2008) Cytotype regulation of P transposable elements in Drosophila melanogaster: repressor polypeptides or piRNAs? Genetics 179:1785–1793. doi:10.1534/genetics.108.087072

Jinek M, Doudna JA (2009) A three-dimensional view of the molecular machinery of RNA interference. Nature 457:405–412. doi:10.1038/nature07755

Jones BC, Wood JG, Chang C, Tam AD, Franklin MJ, Siegel ER, Helfand SL (2016) A somatic piRNA pathway in the Drosophila fat body ensures metabolic homeostasis and normal lifespan. Nat. Commun. 7:13856. doi:10.1038/ncomms13856

Jordan IK, Rogozin IB, Glazko GV, Koonin EV (2003) Origin of a substantial fraction of human regulatory sequences from transposable elements. Trends Genet. TIG 19:68–72

Kaessmann H, Vinckenbosch N, Long M (2009) RNA-based gene duplication: mechanistic and evolutionary insights. Nat Rev Genet 10:19–31. doi:10.1038/nrg2487

Karijolich J, Abernathy E, Glaunsinger BA (2015) Infection-Induced Retrotransposon-Derived Noncoding RNAs Enhance Herpesviral Gene Expression via the NF-κB Pathway. PLoS Pathog 11:e1005260. doi:10.1371/journal.ppat.1005260

Kelleher ES, Edelman NB, Barbash DA (2012) Drosophila interspecific hybrids phenocopy piRNA-pathway mutants. PLoS Biol 10:e1001428. doi:10.1371/journal.pbio.1001428

Khurana JS, Wang J, Xu J, Koppetsch BS, Thomson TC, Nowosielska A, Li C, Zamore PD, Weng Z, Theurkauf WE (2011) Adaptation to P element transposon invasion in Drosophila melanogaster. Cell 147:1551–1563. doi:10.1016/j.cell.2011.11.042

Kidwell MG (1977) Reciprocal differences in female recombination associated with hybrid dysgenesis in Drosophila melanogaster. Genet Res 30:77–88

Kim D-S, Huh J-W, Kim H-S (2007) Transposable elements in human cancers by genome-wide EST alignment. Genes Genet. Syst. 82:145–156

Klenov MS, Lavrov SA, Stolyarenko AD, Ryazansky SS, Aravin AA, Tuschl T, Gvozdev VA (2007) Repeat-associated siRNAs cause chromatin silencing of retrotransposons in the Drosophila melanogaster germline. Nucleic Acids Res 35:5430–5438. doi:10.1093/nar/gkm576

Kokošar J, Kordiš D (2013) Genesis and regulatory wiring of retroelement-derived domesticated genes: a phylogenomic perspective. Mol Biol Evol 30:1015–1031. doi:10.1093/molbev/mst014

Kolaczkowski B, Hupalo DN, Kern AD (2011) Recurrent adaptation in RNA interference genes across the Drosophila phylogeny. Mol Biol Evol 28:1033–1042. doi:10.1093/molbev/msq284

Labrador M, Farré M, Utzet F, Fontdevila A (1999) Interspecific hybridization increases transposition rates of Osvaldo. Mol Biol Evol 16:931–937

Langley CH, Montgomery E, Hudson R, Kaplan N, Charlesworth B (1988) On the role of unequal exchange in the containment of transposable element copy number. Genet Res 52:223–235

Le Thomas A, Rogers AK, Webster A, Marinov GK, Liao SE, Perkins EM, Hur JK, Aravin AA, Tóth KF (2013) Piwi induces piRNA-guided transcriptional silencing and establishment of a repressive chromatin state. Genes Dev 27:390–399. doi:10.1101/gad.209841.112

Le T-N, Schumann U, Smith NA, Tiwari S, Au PCK, Zhu Q-H, Taylor JM, Kazan K, Llewellyn DJ, Zhang R, Dennis ES, Wang M-B (2014) DNA demethylases target promoter transposable elements to positively regulate stress responsive genes in Arabidopsis. Genome Biol 15:458. doi:10.1186/s13059-014-0458-3

Levis RW, Ganesan R, Houtchens K, Tolar LA, Sheen FM (1993) Transposons in place of telomeric repeats at a Drosophila telomere. Cell 75:1083–1093

Lister R, O'Malley RC, Tonti-Filippini J, Gregory BD, Berry CC, Millar AH, Ecker JR (2008) Highly integrated single-base resolution maps of the epigenome in Arabidopsis. Cell 133:523–536. doi:10.1016/j.cell.2008.03.029

Lopez-Maestre H, Carnelossi EAG, Lacroix V, Burlet N, Mugat B, Chambeyron S, Carareto CMA, Vieira C (2017) Identification of misexpressed genetic elements in hybrids between Drosophila-related species. Sci. Rep. 7:40618. doi:10.1038/srep40618

Lu J, Clark AG (2010) Population dynamics of PIWI-interacting RNAs (piRNAs) and their targets in Drosophila. Genome Res 20:212–227. doi:10.1101/gr.095406.109

Malone CD, Brennecke J, Dus M, Stark A, McCombie WR, Sachidanandam R, Hannon GJ (2009) Specialized piRNA pathways act in germline and somatic tissues of the Drosophila ovary. Cell 137:522–535. doi:10.1016/j.cell.2009.03.040

Malone CD, Hannon GJ (2009) Small RNAs as guardians of the genome. Cell 136:656–668. doi:10.1016/j.cell.2009.01.045

Martin A, Troadec C, Boualem A, Rajab M, Fernandez R, Morin H, Pitrat M, Dogimont C, Bendahmane A (2009) A transposon-induced epigenetic change leads to sex determination in melon. Nature 461:1135–1138. doi:10.1038/nature08498

McClintock B (1950) The origin and behavior of mutable loci in maize. Proc. Natl. Acad. Sci. U. S. A. 36:344–355

Meister G (2013) Argonaute proteins: functional insights and emerging roles. Nat Rev Genet 14:447–459. doi:10.1038/nrg3462

Melayah D, Bonnivard E, Chalhoub B, Audeon C, Grandbastien MA (2001) The mobility of the tobacco Tnt1 retrotransposon correlates with its transcriptional activation by fungal factors. Plant J. Cell Mol. Biol. 28:159–168

Metcalfe CJ, Bulazel KV, Ferreri GC, Schroeder-Reiter E, Wanner G, Rens W, Obergfell C, Eldridge MDB, O'Neill RJ (2007) Genomic instability within centromeres of interspecific marsupial hybrids. Genetics 177:2507–2517. doi:10.1534/genetics.107.082313

Mi S, Lee X, Li X, Veldman GM, Finnerty H, Racie L, LaVallie E, Tang XY, Edouard P, Howes S, Keith JC, McCoy JM (2000) Syncytin is a captive retroviral envelope protein involved in human placental morphogenesis. Nature 403:785–789. doi:10.1038/35001608

Micard D, Couderc JL, Sobrier ML, Giraud G, Dastugue B (1988) Molecular study of the retrovirus-like transposable element 412, a 20-OH ecdysone responsive repetitive sequence in Drosophila cultured cells. Nucleic Acids Res 16:455–470

Miller WJ, McDonald JF, Nouaud D, Anxolabéhère D (1999) Molecular domestication–more than a sporadic episode in evolution. Genetica 107:197–207

Mohn F, Handler D, Brennecke J (2015) Noncoding RNA. piRNA-guided slicing specifies transcripts for Zucchini-dependent, phased piRNA biogenesis. Science 348:812–817

Morgan HD, Sutherland HG, Martin DI, Whitelaw E (1999) Epigenetic inheritance at the agouti locus in the mouse. Nat Genet 23:314–318. doi:10.1038/15490

Muotri AR, Chu VT, Marchetto MCN, Deng W, Moran JV, Gage FH (2005) Somatic mosaicism in neuronal precursor cells mediated by L1 retrotransposition. Nature 435:903–910. doi:10.1038/nature03663

Obbard DJ, Gordon KHJ, Buck AH, Jiggins FM (2009a) The evolution of RNAi as a defence against viruses and transposable elements. Philos Trans R Soc Lond B Biol Sci 364:99–115. doi:10.1098/rstb.2008.0168

Obbard DJ, Jiggins FM, Halligan DL, Little TJ (2006) Natural selection drives extremely rapid evolution in antiviral RNAi genes. Curr. Biol. CB 16:580–585. doi:10.1016/j.cub.2006.01.065

Obbard DJ, Welch JJ, Kim K-W, Jiggins FM (2009b) Quantifying adaptive evolution in the Drosophila immune system. PLoS Genet 5:e1000698. doi:10.1371/journal.pgen.1000698

O'Neill RJ, O'Neill MJ, Graves JA (1998) Undermethylation associated with retroelement activation and chromosome remodelling in an interspecific mammalian hybrid. Nature 393:68–72. doi:10.1038/29985

Pélisson A, Song SU, Prud'homme N, Smith PA, Bucheton A, Corces VG (1994) Gypsy transposition correlates with the production of a retroviral envelope-like protein under the tissue-specific control of the Drosophila flamenco gene. EMBO J 13:4401–4411

Perrat PN, DasGupta S, Wang J, Theurkauf W, Weng Z, Rosbash M, Waddell S (2013) Transposition-driven genomic heterogeneity in the Drosophila brain. Science 340:91–95. doi:10.1126/science.1231965

Petrov DA, Aminetzach YT, Davis JC, Bensasson D, Hirsh AE (2003) Size matters: non-LTR retrotransposable elements and ectopic recombination in Drosophila. Mol Biol Evol 20:880–892. doi:10.1093/molbev/msg102

Picard G (1976) Non-mendelian female sterility in Drosophila melanogaster: hereditary transmission of I factor. Genetics 83:107–123

Pietzenuk B, Markus C, Gaubert H, Bagwan N, Merotto A, Bucher E, Pecinka A (2016) Recurrent evolution of heat-responsiveness in Brassicaceae COPIA elements. Genome Biol 17:209. doi:10.1186/s13059-016-1072-3

Prud'homme N, Gans M, Masson M, Terzian C, Bucheton A (1995) Flamenco, a gene controlling the gypsy retrovirus of Drosophila melanogaster. Genetics 139:697–711

Rebollo R, Horard B, Hubert B, Vieira C (2010) Jumping genes and epigenetics: Towards new species. Gene 454:1–7. doi:10.1016/j.gene.2010.01.003

Rebollo R, Karimi MM, Bilenky M, Gagnier L, Miceli-Royer K, Zhang Y, Goyal P, Keane TM, Jones S, Hirst M, Lorincz MC, Mager DL (2011) Retrotransposon-induced heterochromatin spreading in the mouse revealed by insertional polymorphisms. PLoS Genet 7:e1002301. doi:10.1371/journal.pgen.1002301

Rebollo R, Romanish MT, Mager DL (2012) Transposable elements: an abundant and natural source of regulatory sequences for host genes. Annu Rev Genet 46:21–42. doi:10.1146/annurev-genet-110711-155621

Rook GAW (2012) Hygiene hypothesis and autoimmune diseases. Clin Rev Allergy Immunol 42:5–15. doi:10.1007/s12016-011-8285-8

Rozhkov NV, Hammell M, Hannon GJ (2013a) Multiple roles for Piwi in silencing Drosophila transposons. Genes Dev 27:400–412. doi:10.1101/gad.209767.112

Rozhkov NV, Schostak NG, Zelentsova ES, Yushenova IA, Zatsepina OG, Evgen'ev MB (2013b) Evolution and dynamics of small RNA response to a retroelement invasion in Drosophila. Mol Biol Evol 30:397–408. doi:10.1093/molbev/mss241

Saito K, Nishida KM, Mori T, Kawamura Y, Miyoshi K, Nagami T, Siomi H, Siomi MC (2006) Specific association of Piwi with rasiRNAs derived from retrotransposon and heterochromatic regions in the Drosophila genome. Genes Dev 20:2214–2222. doi:10.1101/gad.1454806

Senti K-A, Brennecke J (2010) The piRNA pathway: a fly's perspective on the guardian of the genome. Trends Genet. TIG 26:499–509. doi:10.1016/j.tig.2010.08.007

Sienski G, Dönertas D, Brennecke J (2012) Transcriptional silencing of transposons by Piwi and maelstrom and its impact on chromatin state and gene expression. Cell 151:964–980. doi:10.1016/j.cell.2012.10.040

Siomi H, Siomi MC (2015) RNA. Phased piRNAs tackle transposons. Science 348:756–757. doi:10.1126/science.aab3004

Siomi MC, Sato K, Pezic D, Aravin AA (2011) PIWI-interacting small RNAs: the vanguard of genome defence. Nat Rev Mol Cell Biol 12:246–258. doi:10.1038/nrm3089

Slotkin RK, Martienssen R (2007) Transposable elements and the epigenetic regulation of the genome. Nat Rev Genet 8:272–285. doi:10.1038/nrg2072

Song X, Cao X (2017) Transposon-mediated epigenetic regulation contributes to phenotypic diversity and environmental adaptation in rice. Curr Opin Plant Biol 36:111–118. doi:10.1016/j.pbi.2017.02.004

Strachan DP (1989) Hay fever, hygiene, and household size. BMJ 299:1259–1260

Sutter NB, Bustamante CD, Chase K, Gray MM, Zhao K, Zhu L, Padhukasahasram B, Karlins E, Davis S, Jones PG, Quignon P, Johnson GS, Parker HG, Fretwell N, Mosher DS, Lawler DF, Satyaraj E, Nordborg M, Lark KG, Wayne RK, Ostrander EA (2007) A Single IGF1 Allele Is a Major Determinant of Small Size in Dogs. Science 316:112. doi:10.1126/science.1137045

Todeschini A-L, Teysset L, Delmarre V, Ronsseray S (2010) The epigenetic trans-silencing effect in Drosophila involves maternally-transmitted small RNAs whose production depends on the piRNA pathway and HP1. PLoS ONE 5:e11032. doi:10.1371/journal.pone.0011032

Uchiyama T, Hiura S, Ebinuma I, Senda M, Mikami T, Martin C, Kishima Y (2013) A pair of transposons coordinately suppresses gene expression, independent of pathways mediated by siRNA in Antirrhinum. New Phytol 197:431–440. doi:10.1111/nph.12041

Vagin VV, Sigova A, Li C, Seitz H, Gvozdev V, Zamore PD (2006) A distinct small RNA pathway silences selfish genetic elements in the germline. Science 313:320–324. doi:10.1126/science.1129333

van Valen L (1973) A new evolutionary law. Evol Theory 10:71–74

Van't Hof AE, Campagne P, Rigden DJ, Yung CJ, Lingley J, Quail MA, Hall N, Darby AC, Saccheri IJ (2016) The industrial melanism mutation in British peppered moths is a transposable element. Nature 534:102–105. doi:10.1038/nature17951

Vela D, Fontdevila A, Vieira C, García Guerreiro MP (2014) A genome-wide survey of genetic instability by transposition in Drosophila hybrids. PLoS ONE 9:e88992. doi:10.1371/journal.pone.0088992

Vieira C, Lepetit D, Dumont S, Biémont C (1999) Wake up of transposable elements following Drosophila simulans worldwide colonization. Mol Biol Evol 16:1251–1255

Volff J-N (2006) Turning junk into gold: domestication of transposable elements and the creation of new genes in eukaryotes. BioEssays News Rev. Mol. Cell. Dev. Biol. 28:913–922. doi:10.1002/bies.20452

Wang W, Han BW, Tipping C, Ge DT, Zhang Z, Weng Z, Zamore PD (2015) Slicing and Binding by Ago3 or Aub Trigger Piwi-Bound piRNA Production by Distinct Mechanisms. Mol Cell 59:819–830. doi:10.1016/j.molcel.2015.08.007

Waterland RA, Jirtle RL (2003) Transposable elements: targets for early nutritional effects on epigenetic gene regulation. Mol Cell Biol 23:5293–5300

Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, Flavell A, Leroy P, Morgante M, Panaud O, Paux E, SanMiguel P, Schulman AH (2007) A unified classification system for eukaryotic transposable elements. Nat Rev Genet 8:973–982. doi:10.1038/nrg2165

Yan Y, Buckler-White A, Wollenberg K, Kozak CA (2009) Origin, antiviral function and evidence for positive selection of the gammaretrovirus restriction gene Fv1 in the genus Mus. Proc. Natl. Acad. Sci. U. S. A. 106:3259–3263. doi:10.1073/pnas.0900181106

Zanni V, Eymery A, Coiffet M, Zytnicki M, Luyten I, Quesneville H, Vaury C, Jensen S (2013) Distribution, evolution, and diversity of retrotransposons at the flamenco locus reflect the regulatory properties of piRNA clusters. Proc. Natl. Acad. Sci. U. S. A. 110:19842–19847. doi:10.1073/pnas.1313677110

Zemojtel T, Vingron M (2012) P53 binding sites in transposons. Front. Genet. 3:40. doi:10.3389/fgene.2012.00040

Zhang F, Peterson T (2005) Comparisons of Maize pericarp color1 Alleles Reveal Paralogous Gene Recombination and an Organ-Specific Enhancer Region. Plant Cell 17:903–914. doi:10.1105/tpc.104.029660

Zhou Q, Ellison CE, Kaiser VB, Alekseyenko AA, Gorchakov AA, Bachtrog D (2013) The epigenome of evolving Drosophila neo-sex chromosomes: dosage compensation and heterochromatin formation. PLoS Biol 11:e1001711. doi:10.1371/journal.pbio.1001711

Mammalian-Specific Traits Generated by LTR Retrotransposon-Derived *SIRH* Genes

Tomoko Kaneko-Ishino, Masahito Irie and Fumitoshi Ishino

Abstract What is the mechanism by which mammalian-specific genes derived from long terminal repeat (LTR) retrotransposons played a role in generating mammalian-specific traits, such as a unique viviparous reproductive system and a highly developed central nervous system? A series of knockout mouse studies has clearly demonstrated that at least some sushi-ichi-related retrotransposon homologues (*SIRH* genes) play essential roles in placental development and brain function. Some *SIRH* genes are conserved in all eutherians, whereas other *SIRH* genes became pseudogenes in a species- or lineage-specific manner, implying that LTR retrotransposons served as a critical driving force in mammalian evolution and diversification by generating mammalian-specific and species- or lineage-specific genes, respectively. Interestingly, most *SIRH* genes are located on the X chromosome. We discuss whether there is a specific reason for or advantage of having an X-linked chromosomal location, and we also discuss the role of X chromosome inactivation during this process.

1 Introduction

Only 1.5% of the human genome consists of protein-coding genes, whereas approximately half is occupied by retrotransposons, such as long interspersed nuclear elements (LINEs, 20%), short interspersed nuclear elements (SINEs, 13%), and LTR retrotransposons/retroviruses (8%). In 2000, Mi et al. reported that the human *syncytin-1* gene, which is derived from an *Env* gene of an endogenous retrovirus (ERVW-1), exhibits cell fusion activity in vitro (Mi et al. 2000). Then, they proposed that the product of *syncytin-1* functions to form the syncytiotro-

T. Kaneko-Ishino · M. Irie
School of Health Sciences, Tokai University, Isehara-shi, Japan

M. Irie · F. Ishino (✉)
Department of Epigenetics, Medical Research Institute, Tokyo Medical and Dental University (TMDU), 1-5-45 Yushima, Bunkyo-ku, Tokyo 113-8510, Japan
e-mail: fishino.epgn@mri.tmd.ac.jp

P. Pontarotti (ed.), *Evolutionary Biology: Self/Nonself Evolution, Species and Complex Traits Evolution, Methods and Concepts*,
DOI 10.1007/978-3-319-61569-1_7

phoblast in the placenta because it is highly expressed in this part of the placenta. The next year, the first two long terminal repeat (LTR) retrotransposon-derived genes, *Paternally expressed 10* (*Peg10*) (Ono et al. 2001) and *Paternally expressed 11/Retrotransposon-like 1* (*Peg11/Rtl1*) (Charlier et al. 2001), were identified from comprehensive studies on genomic imprinting, a mammalian-specific (more precisely, a therian-specific) epigenetic mechanism (Surani et al. 1984; McGrath and Solter 1984; Mann and Lovell-Badge 1984; Cattanach and Kirk 1985). In 2006 and 2008, we demonstrated that *Peg10* is the major imprinted gene responsible for early embryonic lethality of parthenogenetic embryos (Ono et al. 2006) comprising two maternally derived genomes (Surani et al. 1984; McGrath and Solter 1984; Mann and Lovell-Badge 1984) and that *Peg11/Rtl1* is the major imprinted gene responsible for late embryonic to neonatal lethality associated with several morphological defects (Sekita et al. 2008) observed in paternal/maternal disomy of mouse chromosome 12 (Ch12) (Cattanach and Beechey 1990; Georgiades et al. 2000), respectively. *Peg10* and *Peg11/Rtl1* play critical roles in the formation and maintenance of placentas, respectively (Ono et al. 2006; Sekita et al. 2008). As the placenta is an essential organ for viviparous reproductive systems in therians (eutherians and marsupials), these studies reveal that LTR retrotransposons/retroviruses are not mere garbage of the genome (Gould and Vrba 1982; Brosius and Gould 1992) but that at least some of these became endogenous genes that could drive mammalian evolution and diversification (Kaneko-Ishino and Ishino 2010, 2012, 2015). Retrotransposon-derived genes have been increasingly attracting attention given their roles in mammalian developmental systems and in mammalian evolution.

These genes emerged in the course of mammalian evolution and are therefore evolutionarily young. *Syncytin-1* is derived from a primate-specific ERV (Mi et al. 2000), whereas *syncytins* in other eutherian species have a different origin and are derived from their own lineage-specific ERVs (Dupressoir et al. 2005, 2009; Lavialle et al. 2013; Nakaya et al. 2013; Cornelis et al. 2015). Both PEG10 and PEG11/RTL1 proteins exhibit homologies to the Gag and Pol proteins of a sushi-ichi retrotransposon, suggesting that both are derived from a certain sushi-ichi-related retrotransposon (Ono et al. 2001; Charlier et al. 2001; Volff et al. 2001). The sushi-ichi retrotransposon is a gypsy LTR retrotransposon isolated from fugu fish (puffer fish), but it does not exist in birds, reptiles, and mammals (Poulter and Butler 1998). *PEG10* is therian-specific, i.e., conserved in both marsupials and eutherians (Suzuki et al. 2007), whereas *PEG11/RTL1* is a eutherian-specific gene that is absent in marsupials (Edwards et al. 2008). These findings promoted screening for new candidates for such LTR retrotransposon-derived genes. Nine additional sushi-ichi-related retrotransposon homologues (*SIRH3-SIRH11* genes, also called *MART* genes) were discovered in human and mouse genomes by comprehensive screening of sushi-ichi Gag-like genes (Ono et al. 2006; Brandt et al. 2005; Youngson et al. 2005; Campillos et al. 2006). All of the *SIRH3-11* genes were domesticated in the eutherian ancestor, as was *PEG11/RTL1/SIRH2,* and all except for *SIRH9* have only a Gag-like region. *SIRH9* contains regions corresponding to Gag and Pol, such as *PEG10/SIRH1* and *PEG11/RTL1/SIRH2*.

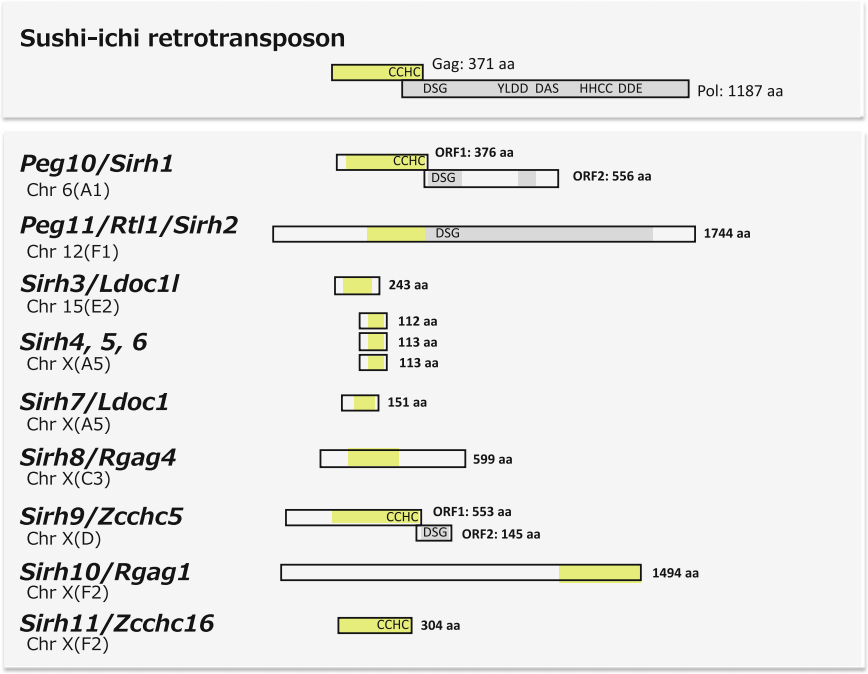


Fig. 1 *SIRH* genes in mice. Chromosomal location of each gene is indicated under the name of each gene. *SIRH* genes are also called *mammalian retrotransposon-derived* (*Mart*) or *sushi-ichi retrotransposon-derived* (*Sushi*) genes. Gag- and Pol-like parts are shown in *yellow* and *gray*, respectively. CCHC: RNA-binding motif, DSG: protease active site, YLDD: reverse transcriptase motif, DAS: RNase H highly conserved motif, HHCC: integrase DNA binding motif, and DDE: integrase catalytic motif

Marsupials have *PEG10* and an additional marsupial-specific *SIHR12* that is absent from eutherians. It was identified in the tammar wallaby, an Australian marsupial species; however, it became a pseudogene in the gray short-tailed opossum, a South American marsupial species. Interestingly, there are three degenerated copies of suchi-ichi-like retrotransposons compared with two *SIRH* genes in the tammar wallaby genomes, whereas no such remnants exist in the human and mouse genomes compared with 11 *SIRH* genes (Ono et al. 2011) although some eutherian *SIRH* genes became pseudogenes in a species- and/or lineage-specific manner as described below.

In this review, we focus on these LTR retrotransposon-derived *SIRH* genes (Fig. 1) and review their biological roles in mammalian developmental systems as determined by a series of analyses of knockout (KO) mice. We also discuss their impacts on mammalian evolution and diversification. Interestingly, some *SIRH* genes are conserved in all eutherian lineages, whereas others became pseudogenized in a species- or lineage-specific manner (Irie et al. 2015, 2016). These results suggest that the former contributed to the establishment of basic eutherian

characters, whereas the latter contributed to the diversification of eutherians in a species- or lineage-specific manner (Kaneko-Ishino and Ishino 2015; Irie et al. 2015, 2016). By revealing their biological functions, we can trace and reconstruct the processes of evolution and diversification in a step-by-step manner. Interestingly, eight out of 11 of the *SIRH* genes are located on the X chromosome: *PEG10*, *PEG11/RTL1,* and *SIRH3* are autosomal, whereas *SIRH4-11* genes are X-linked (Ono et al. 2006; Brandt et al. 2005; Youngson et al. 2005; Campillos et al. 2006). In the last section, we discuss why these genes are enriched in the X chromosome. From an evolutionary point of view, we discuss the role of X chromosome inactivation mechanisms for gaining new functional genes.

2 Biological Functions of LTR Retrotransposon-Derived *SIRH* Genes in Placenta and Brain in Eutherian Mammals

(1) *PEG10/SIRH1*

PEG10 is a therian-specific gene, that is, it is conserved in both marsupials and eutherians but absent from monotremes and other vertebrates, such as fish, frogs, reptiles, and birds (Suzuki et al. 2007). *PEG10* encodes two open reading frames (ORF), PEG10-ORF1 and ORF2, that exhibit homologies to the Gag and Pol proteins of a sushi-ichi retrotransposon, respectively (Ono et al. 2006; Shigemoto et al. 2001) (Fig. 2). In addition to the ORF1 protein, an ORF1-2 fusion protein is translated via a −1 frameshift mechanism similar to that observed with LTR

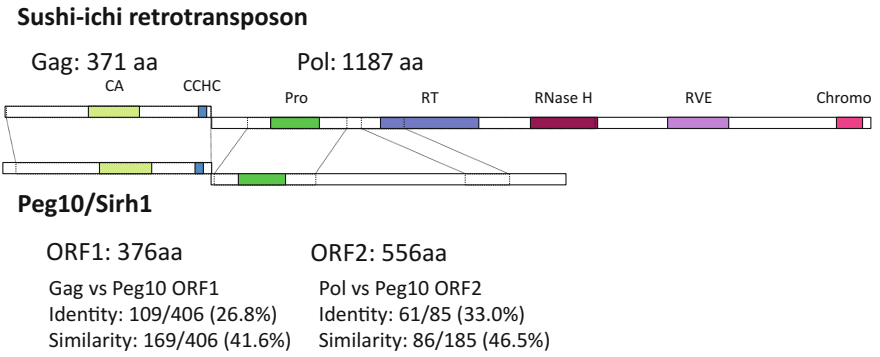


Fig. 2 Mouse Peg10 and sushi-ichi retrotransposon Gag and Pol. Amino acid sequence identities and similarities between Gag and Peg10 ORF1 and between Pol and the Pol-like parts of Peg10 ORF1-2 are shown, respectively. *CA* capsid domain, *CCHC* RNA-binding site, *Pro* aspartic protease, *RT* reverse transcriptase, *RVE* integrase core domain, *Chomo* integrase chromo domain

retrotransposons and retroviruses (Shigemoto et al. 2001). The amino acid sequence identities between the sushi-ichi retrotransposon Gag and the mouse Peg10 ORF1 protein and between sushi-ichi Pol and the Pol-like region of the mouse ORF1-2 protein are 26.8 and 33.0%, respectively (41.6 and 46.5% similarity, respectively), suggesting that numerous mutations were necessary to generate the present *PEG10* from an LTR retrotransposon that was originally integrated into the genome of a common therian ancestor. Among eutherian and marsupial species, the amino acid sequence of the PEG10 protein is highly conserved, especially, within the CCHC RNA-binding domain of the Gag protein and the DSG protease active site of the Pol protein. It is highly likely that these two domains are essential components that contribute to PEG10 function.

PEG10 is expressed in both embryos and placentas. Marsupials and eutherians are viviparous mammals but have different types of placentas (Renfree 2010). In the tammer wallaby, an Australian marsupial species, *PEG10,* is expressed in the yolk sac placenta (Suzuki et al. 2007), whereas human and mouse *PEG10*/*Peg10* is expressed in chorioallantoic placentas and yolk sacs (Ono et al. 2001, 2006). We demonstrated that *Peg10* KO mice exhibit early embryonic lethality and have poorly developed placentas (Ono et al. 2006). Their placentas lack essential components, such as labyrinth and spongiotrophoblast layers. In the labyrinth layer, nutrient and gas exchange occurs between fetal and maternal blood cells; therefore, *Peg10* KO embryos cannot develop beyond embryonic day 9.5 (d9.5) (see also Fig. 6). Given that ectoplacental cone (EPC) growth was markedly affected in the KO placentas (Ono et al. 2006), we assume that *Peg10* plays an important role in differentiating trophoblast cells in the labyrinth and spongiotrophoblast layers from the EPC. It is likely that deletion of human *PEG10* also causes abortion at very early stages of development. Although there are no reports on the biological function of marsupial *PEG10* to date, it is likely that *PEG10* functions in yolk sac placentas in marsupials (Suzuki et al. 2007; Renfree et al. 2013). The fact that the therian-specific *PEG10* gene plays an essential role in placental development implies that the domestication of *PEG10* had a significant impact on the establishment of current viviparous reproduction systems (Kaneko-Ishino and Ishino 2010, 2012, 2015). However, *PEG10* is also expressed in embryos, adults, and many types of cancers (Okabe et al. 2003; Akamatsu et al. 2015; Deng et al. 2014); therefore, it is necessary to determine whether *PEG10* plays other roles in embryonic development and growth beyond placental function.

(2) *PEG11/RTL1/SIRH2*

PEG11/RTL1 is a eutherian-specific gene, that is, it is absent from marsupials (Cornelis et al. 2015). Therefore, it is highly probable that *PEG11/RTL1* was domesticated in a common eutherian ancestor. Mouse *Peg11/Rtl1* encodes one large ORF comprising 1744 amino acids (aa) with homologies to the Gag and Pol of the sushi-ichi retrotransposon (Fig. 3). The amino acid sequence identities between the sushi-ichi retrotransposon Gag and the Gag-like region of the mouse Peg11/Rtl1 protein and between sushi-ichi Pol and the Pol-like region of the mouse Peg11/Rtl1

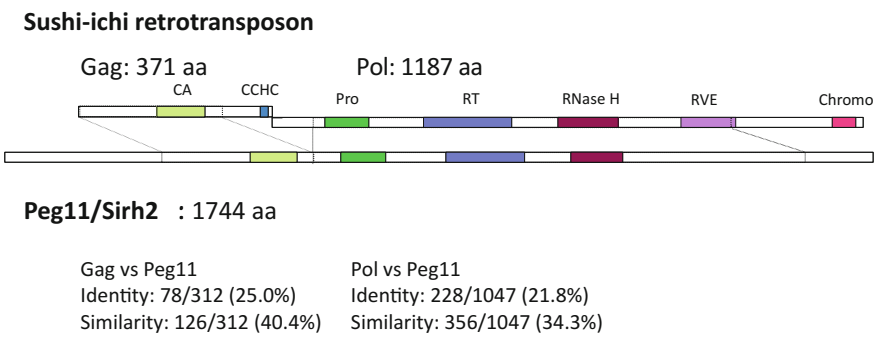


Fig. 3 Mouse Peg11/Rtl1 and sushi-ichi retrotransposon Gag and Pol. Amino acid sequence identities and similarities between Gag and the Gag-like region of Peg11 and between Pol and the Pol-like region of Peg11 are shown, respectively

protein are 25.0 and 21.8%, respectively (40.4 and 34.3% similarity, respectively). The DSG protease active site in the Pol-like part of the PEG11/RTL1 protein is highly conserved among eutherians, suggesting that this is an important functional domain.

PEG11/RTL1 is also expressed in both the embryos and placenta, similar to *PEG10*. As mentioned, the labyrinth layer is an essential part of the eutherian chorioallantoic placenta because it contains numerous fetal capillaries, where the nutrient and gas exchange actually occurs between the fetal and maternal blood cells (Renfree et al. 2013). In the placenta, the PEG11/RTL1 protein is specifically located in fetal capillary endothelial cells which are of an extraembryonic mesoderm origin (Sekita et al. 2008). The endothelial cells are surrounded by two layers of syncytiotrophoblast cells which are of an extraembryonic ectoderm origin similar to all the other trophoblast cells in the placentas, such as spongiotrophoblast (SpTs) and trophoblast giant cells (TGCs). Therefore, the localization of *PEG11/RTL1* expression is different from that of *PEG10*.

We demonstrated that half of the *Peg11/Rtl1* KO mice died at a late fetal stage (Sekita et al. 2008). By contrast, the other half exhibited neonatal lethality within 24 h of birth, and this was associated with severe growth retardation. In the *Peg11/Rtl1* KO placenta, the fetal capillaries in the labyrinth layer were severely affected. The capillaries were clogged at numerous sites because the endothelial cells had been phagocytosed by the surrounding trophoblast cells, indicating that *Peg11/Rtl1* plays an essential role in the maintenance of the feto-maternal interface of the placenta during gestation (see also Fig. 6). The fetal capillary network of the labyrinth layer is basically completed by d12.5 (Watson and Cross 2005). The network becomes larger and more extensively branched until birth (d18.5–19.5), although the placental weight peaks at d16.5. Severe damage to the basal region of the fetal capillary network by d14.5 led to mid-fetal lethality, whereas KO embryos exhibiting late fetal lethality were associated with small placentas resulting from poor expansion of the labyrinth layer (Kitazawa et al. 2017).

In humans, *PEG11/RTL1* is one of the major genes responsible for uniparental disomy of maternal human chromosome 14 (upd(14)mat: Temple syndrome) (Sekita et al. 2008; Kotzot 2004; Ioannides et al. 2014; Kagami et al. 2008) and uniparental disomy of paternal chromosome 14 (upd(14)pat: Kagami-Ogata syndrome) (Sekita et al. 2008; Kagami et al. 2008, 2015). The imprinted region on human chromosome 14 comprises three paternally expressed coding genes, *DLK1/PEG9, PEG11/RTL1,* and *DIO3,* and four maternally expressed noncoding RNAs, *MEG3/GTL2*, *antiPEG11/antiRTL1*, *MEG8,* and *MEG9*. In upd(14)mat patients, loss of expression of all three paternally expressed genes and a double dose of all four maternally expressed genes are observed. Loss of *PEG11/RTL1* causes pre- and postnatal growth retardation that is similar to that observed in *Peg11/Rtl1* KO mice, and additional loss of *DLK1/PEG9* increases the severity of growth retardation (Sekita et al. 2008; Kagami et al. 2008). In upd(14)pat patients, loss of expression of all four maternally expressed genes is associated with a double dose of *DLK1/PEG9* and *DIO3* and a four- to sixfold increase in *PEG11/RTL1* (Kagami et al. 2008) because *antiPEG11/antiRTL1* degrades *PEG11/RTL1* mRNA via its encoded siRNAs (Seitz et al. 2003; Davis et al. 2005). The severity of upd(14)pat syndrome, for example, having a bell-shaped thorax associated with respiratory problems, placentomegaly, abdominal wall defects, postnatal growth retardation, and mental retardation, is well correlated with the degree of *PEG11/RLT1* overproduction (Kagami et al. 2008, 2015). Thus, it will be very important to elucidate *PEG11/RLT1* function in embryonic/neonatal development and growth that is related to such muscle- and bone-related abnormalities, in addition to the maintenance of placental fetal capillaries.

(3) *SIRH7/LDOC1*

SIRH7/Leucine zipper, downregulated in cancer 1 (*LDOC1* (Nagasaki et al. 1999), also called *Mart7*) is conserved in all eutherian species and encodes a small Gag-like protein comprising 151 aa corresponding to the central part of the Gag protein (Fig. 4). The *SIRH7/LDOC1* protein has an additional leucine-zipper motif at the N-terminus. The mouse *Sirh7/Ldoc1* protein exhibits 28.3% identity (40.4% similarity) to the sushi-ichi Gag protein, and this is similar to that of other SIRH4-11 proteins. In mice, *Sirh7/Ldoc1* is predominantly expressed in the early stages of the placenta (Kagami et al. 2015; Naruse et al. 2014). A high expression

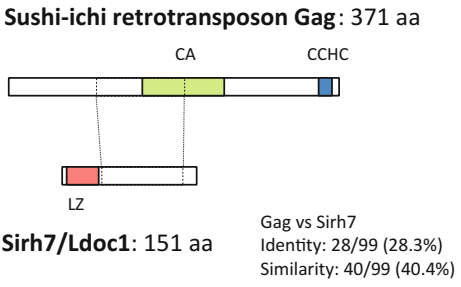


Fig. 4 Mouse Sirh7/Ldoc1 and sushi-ichi retrotransposon Gag. Amino acid sequence identity and similarity between Gag and the Gag-like region of Sirh7 are shown. Sirh7 has an additional leucine-zipper motif (LZ) at the N-terminus

level was observed in all placental cells, including TGCs and the EPC cells that subsequently produce SpTs, glycogen trophoblast (GlyTs), and various TGC subtypes at d9.5. However, its expression gradually became restricted to GlyTs, and it ultimately disappeared after the mid-stage of gestation.

Sirh7/Ldoc1 KO placentas have several structural problems, including delayed development of SpT cells and an irregular boundary between the labyrinth and spongiotrophoblast layers (Naruse et al. 2014). These placentas were associated with several endocrinological problems, such as progesterone (P4) overproduction, a delayed switch from placental lactogen I (PL1) to placental lactogen II (PL2) and alterations in the levels of several prolactin-like proteins (PRLs), given that a variety of placental cells produce placental hormones (P4, PL1, and 2 from TGCs and several PRLs from SpTs and cytotrophoblast cells). Consequently, *Sirh7/Ldoc1* KO females exhibited delayed parturition and a low pup weaning rate. As mentioned, the placenta is a major endocrine organ during gestation. However, it had been long thought that mouse (rodent) placentas do not produce P4, an essential hormone for maintaining pregnancy in mammals, and that ovaries are the only source of P4 during gestation (Malassine et al. 2003). However, in the course of the *Sirh7/Ldoc1* KO mouse study, we first demonstrated that mouse TGCs produce P4 during the mid-stage of development when ovarian P4 production exhibits a temporary reduction (Naruse et al. 2014). Thus, *Sirh7/Ldoc1* is another gene that is essential for placental development and reproduction (see also Fig. 6). Given that the reproductive advantage it conferred, it is likely that *SIRH7/LDOC1* increased fitness of its carriers and that it was positively selected during eutherian evolution (Kaneko-Ishino and Ishino 2015; Malassine et al. 2003).

The name *Leucine zipper, downregulated in cancer 1* (*LDOC1*), is derived from the observation that this gene was downregulated in human pancreas cancers (Nagasaki et al. 1999). *LDOC1* is actually highly expressed in the human pancreas but not in the mouse pancreas. Therefore, *SIRH7/LDOC1* might have species-specific functions in other organs and tissues. *SIRH7/LDOC1* is also highly expressed in the adult human brain. Interestingly, *SIRH7/LDOC1* has several human-specific gene interaction networks in the brain, in contrast to those of chimpanzees, suggesting that it might play an essential role in the evolution of brain function in humans (Oldham et al. 2006). This finding suggests that, addition to its conserved role in the placenta, *SIRH7/LDOC1* has several important roles in other organs and tissues and contributed to the diversification of eutherian species in a variety of ways.

(4) *SIRH11/ZCCHC16*

SIRH11/zinc finger CCHC domain containing 16 (*ZCCHC16,* also called *MART4*) encodes a Gag-like protein comprising approximately 300–310 aa with a CCHC RNA-binding domain in its C-terminal region (Irie et al. 2015) (Fig. 5). Mouse Sirh11/Zcchc16 exhibits 21.0% identity (38.5% similarity) with the entire sushi-ichi retrotransposon Gag, consisting of 371 aa (without the N-terminus). *SIRH11/ZCCHC16* is not conserved in all eutherian species, in contrast to *PEG10*,

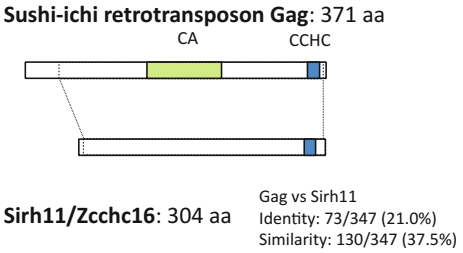


Fig. 5 Mouse Sirh11/Zcchc16 and sushi-ichi retrotransposon Gag. Amino acid sequence identity and similarity between Gag and Sirh11 are shown

PEG11/RTL1, and *SIRH7/LDOC1*. In xenarthrans, such as sloths and armadillos, *SIRH11/ZCCHC16* became a pseudogene through gaining multiple mutations; however, the location of armadillo and sloth *pseudoSIRH11/ZCCHC16* is the same as in other eutherian mammals, suggesting that it contributes to the evolution of boreoeutherian (euarchontoglires and laurasiatherian) and afrotherian mammals (Irie et al. 2015).

Mouse *Sirh11/Zcchc16* is unique among the *Sirh* genes because it does not exhibit any placental expression in mouse development. *Sirh11/Zcchc16* is expressed in the brain, testis, ovary, and kidney in embryos and adults (Irie et al. 2015). *Sirh11/Zcchc16* KO mice exhibited abnormal behaviors related to cognition, including attention, impulsivity, and working memory, probably due to abnormal control of the locus coeruleus-noradrenaline (LC-NA) system (Irie et al. 2015) (see also Fig. 6). Microdialysis analysis after perfusion treatment demonstrates that the prefrontal cortex of *Sirh11/Zcchc16* KO mice exhibited a lower NA recovery rate. It is proposed that phasic activation of the NA neurons of the LC occurs in concert with the cognitive shifts that facilitate dynamic reorganization of target neural networks (Bouret and Sara 2005). This feature permits rapid behavioral adaptation to the demands of changing environmental imperatives. Therefore, it is likely that this effect confers a critically important advantage in the competition of daily life and that this advantage was also present in eutherian evolution. All these results suggest that *SIRH11/ZCCHC16* potentially contributed to the evolution and diversification (see Sect. 2) of eutherian brain functions (Kaneko-Ishino and Ishino 2015; Irie et al. 2015). In humans, *SIRH11/ZCCHC16* may be a good candidate for involvement in X-linked intellectual disability (XLID), attention-deficit/hyperactivity disorder (ADHD) and/or other emotional diseases because it has also been proposed that the phasic LC activity participates in certain critically important behavioral functions and severe mental problems (Bouret and Sara 2005; Aston-Jones et al. 1999; Berridge and Waterhouse 2003; Sara 2009).

(5) Other *SIRH* genes

In addition to *SIRH11/ZCCHC16,* we confirmed other brain-related phenotypes in *Sirh3/Ldoc1-like* (*Ldoc1l,* also called *Mart6*) and *Sirh8/Retrotransposon gag domain containing 4* (*Rgag4,* also called *Mart5*) KO mice, such as reduction of daily activity and abnormal startle reaction, respectively (in preparation) (Fig. 6). It is possible that *SIRH3/LDOC1L* and *SIRH8/RGAG4* were generated by gene duplication because the

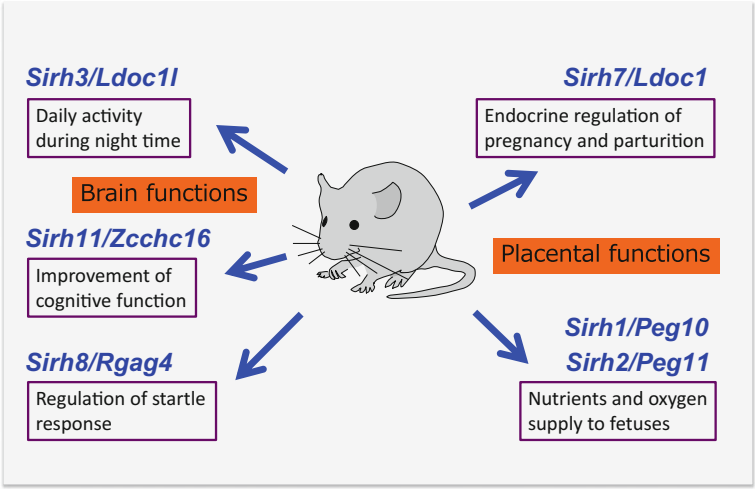


Fig. 6 Biological functions of *SIRH* genes. The biological functions deduced from six KO mouse experiments on *Peg10*, *Peg11/Rtl1*, *Sirh7/Ldoc1*, *Sirh3/Ldoc1l*, *Sirh11/Zcchc16,* and *Sirh8/Rgag4* are presented. Only the major phenotypes of each KO mouse are presented

amino acid homology between these two genes is very high, although the size of the entire ORF is quite different (mouse *Sirh3/Ldoc1l*: 243 aa and *Sirh8/Rgag4*: 599 aa). Interestingly, *SIRH3/LDOC1L* is the most conserved gene among *SIRH* genes, whereas *SIRH8/RGAG4* became a pseudogene in a lineage-specific manner similar to *SIRH11/ZCCHC16* (see Sect. 2). At present, the biological functions of *SIRH9/ZCCHC5* (also called *MART3*) and *SIRH10/RGAG1* (also called *MART9*) and of the triplet genes *SIRH4, 5,* and *6/CAAX box 1A, B,* and *C* (*CXX1A, B,* and *C*, also called *family with sequence similarity 127 member A, B,* and *C* (*FAM127A, B,* and *C*) or *MART8C, A,* and *B*) are not known. Comprehensive KO mouse studies are now in progress to elucidate their biological functions.

3 Origin of *SIRH* Genes and Their Distribution in Eutherians

Of the *SIRH* genes, only *PEG10/SIRH1* was acquired in the common therian ancestor; therefore, this gene is evolutionarily the oldest gene (Suzuki et al. 2007). Comparative genome analyses among three mammalian groups (monotremes, marsupials, and eutherians) and other vertebrate groups demonstrated that *PEG10/SIRH1* is only conserved in both marsupials and eutherians, suggesting that *PEG10* was acquired after the split of the therians from the monotremes, 166 or 186 million years ago (Ma), and before the eutherian/marsupial split, 160 Ma (Suzuki et al. 2007). Using similar comparative genome analyses, Edwards et al. demonstrated that

PEG11/RTL1/SIRH2 is conserved only in eutherians but absent from marsupials (Edwards et al. 2008), suggesting that it was acquired after the eutherian/marsupial split, 160 Ma, but before the split of the three major eutherian lineages, boreoeutheria (including euarchontoglires and laurasiatheria), afrotheria, and xenarthra, 120 Ma (Kaneko-Ishino and Ishino 2012, 2015). The *SIRH3-SIRH11* genes are also eutherian-specific genes like *PEG11/RTL1/SIRH2*. However, some of these genes (*SIRH3-7*) are conserved in most of the eutherian species, whereas the others (*SIRH8-11*) became pseudogenes or possess large structural changes in a species- or lineage-specific manner (Irie et al. 2015).

It is not easy to determine how this group of *SIRH* genes emerged and expanded during eutherian evolution. With the exception of the triplet (*SIRH4-6*) genes, the size of each *SIRH* gene is quite variable (Fig. 1). All *SIRH* genes exhibit a similar degree of identity with the sushi-ichi retrotransposon (20–35%); however, their amino acid sequences also differ from each other except among the triplet (*SIRH4-6*) genes and between *SIRH3/LDOC1L* and *SIRH8/RGAG4.* One possibility is that these eutherian-specific *SIRH* genes were acquired from the sushi-ichi-related retrotransposon independently from *PEG10/SIRH1* upon each domestication event. Another possibility is that these genes originated from *PEG10/SIRH1* by gene duplication or from cDNA retrotransposition of *PEG10/SIRH1* transcripts exhibiting a variety of lengths. Alternatively, some genes might be derived from *PEG10,* whereas others might be derived independently. It is also possible that some *SIRH* genes are derived from other *SIRH* genes. Although we do not know which possibility is correct at present, it seems certain that it took considerable time to generate each *SIRH* gene via multiple mutations after its integration into the present locus, because no direct candidate sequences (or precursor sequences) exist in the genome.

The distribution of *SIRH* genes in eutherians is of considerable interest. As mentioned, *SIRH11/ZCCHC16* became a pseudogene through acquiring multiple mutations in xenarthrans, such as sloths and armadillos (Irie et al. 2015). There is a common nonsense mutation immediately N-terminal to the functionally important C-terminal CCHC RNA-binding domain of *SIRH11/ZCCHC16* in the two extant orders in xenathrans, Pilosa (sloths), and Cingulata (armadillos) (Irie et al. 2016). This finding suggests that *SIRH11/ZCCHC16* was present in the common eutherian ancestor, similar to other *SIRH* genes, but lost in a common xenarthran ancestor. Furthermore, two types of significant mutations were observed in boreoeutherians (euarchontoglires and laurasiatherians) in a species- or lineage-specific manner (Irie et al. 2016). One mutation is a nonsense mutation leading to the loss of the CCHC RNA-binding domain (5 species: the white-cheeked gibbon, Chinese tree shrew, Amur tiger, and two flying foxes), and the other mutation is also a nonsense mutation resulting in the loss of the N-terminal half of the *SIRH11/ZCCHC16* ORF (3 lineages, Platyrrhini (the New World monkeys), Hystricognathi (the New World and African rodents), and species belonging to Cetacea and Ruminantia). In both cases, causative nonsense mutations occur independently of each other; therefore, these mutations are species- or lineage-specific. However, in the latter cases, the resulting putative C-terminal half comprising 167 aa is conserved. Extensive dN/dS

analysis suggests that such truncated SIRH11/ZCCHC16 ORFs are functionally diversified even within the same lineages. Thus, *SIRH11/ZCCHC16* might contribute to the diversification of eutherians by species- or lineage-specific structural changes after domestication in the common eutherian ancestor followed by putative species-specific functional changes that enhanced fitness and occurred as a consequence of complex natural selection events (Kaneko-Ishino and Ishino 2015; Irie et al. 2016).

Genomic data also suggest that (1) *SIRH8/RGAG4* became a pseudogene in most afrotherian and xenarthran species, (2) *SIRH9/ZCCHC5* is lost in many eutherian species, and (3) a large deletion or insertion is present in *SIRH10/RGAG1* in many eutherian species. In conclusion, *PEG10/SIRH1, PEG11/RTL1/SIRH2, SIRH3/LDOC1L, SIRH4-6/CXX1A-C,* and *SIRH7/LDOC1* are highly conserved in eutherians, whereas *SIRH8/RGAG4, SIRH9/ZCCHC5, SIRH10/RGAG1,* and *SIRH11/ZCCHC16* are lost in a species- and/or lineage-specific manner. It is likely that the former play essential and fundamental roles in the development and behavioral systems, whereas the latter could act as critical determinants in the process of diversification via their brain-related functions or other lineage- and/or species-specific characters in eutherians (Kaneko-Ishino and Ishino 2015). Such changes may be dependent on a variety of environmental factors, such as ecological niches, lifestyle dynamics, and the evolutionary history of the species, including geological events (Irie et al. 2015, 2016). Therefore, these findings will provide valuable information regarding how eutherian evolution and diversification occurred. It should be mentioned that even the essential placental *SIRH* genes might also play some roles in other organs and tissues across all species or in a species- and/or lineage-specific manner.

4 Why are *SIRH* Genes Enriched in the X Chromosome?

As discussed in Sect. 2, the origin of each *SIRH* gene is not clear at present. However, regardless of its origin, it is certain that each *SIRH* gene was somehow fixed in the genome via natural selection and/or random genetic drift in the common eutherian ancestor. The domestication of retrotransposons seems likely to be a very rare event because the integrated retrotransposons or retrotransposon derivatives are typically harmful rather than advantageous. We previously proposed a hypothesis that in the course of retrotransposon domestication, neutral or nearly neutral evolution preceded Darwinian evolution and helped supply novel materials for novel functional genes from integrated retrotransposons (Kaneko-Ishino and Ishino 2012, 2015). According to the neutral (Kimura 1968, 1983) and nearly neutral theories (Ohta 2002) of molecular evolution proposed by Motoo Kimura and Tomoko Ohta, respectively, neutral or nearly neutral (less harmful) mutations could be fixed in a population by random drift. Then, we noted the importance of epigenetic mechanisms, such as DNA methylation and histone modifications, that could silence the integrated retrotransposons transcriptionally because the silent genes can behave

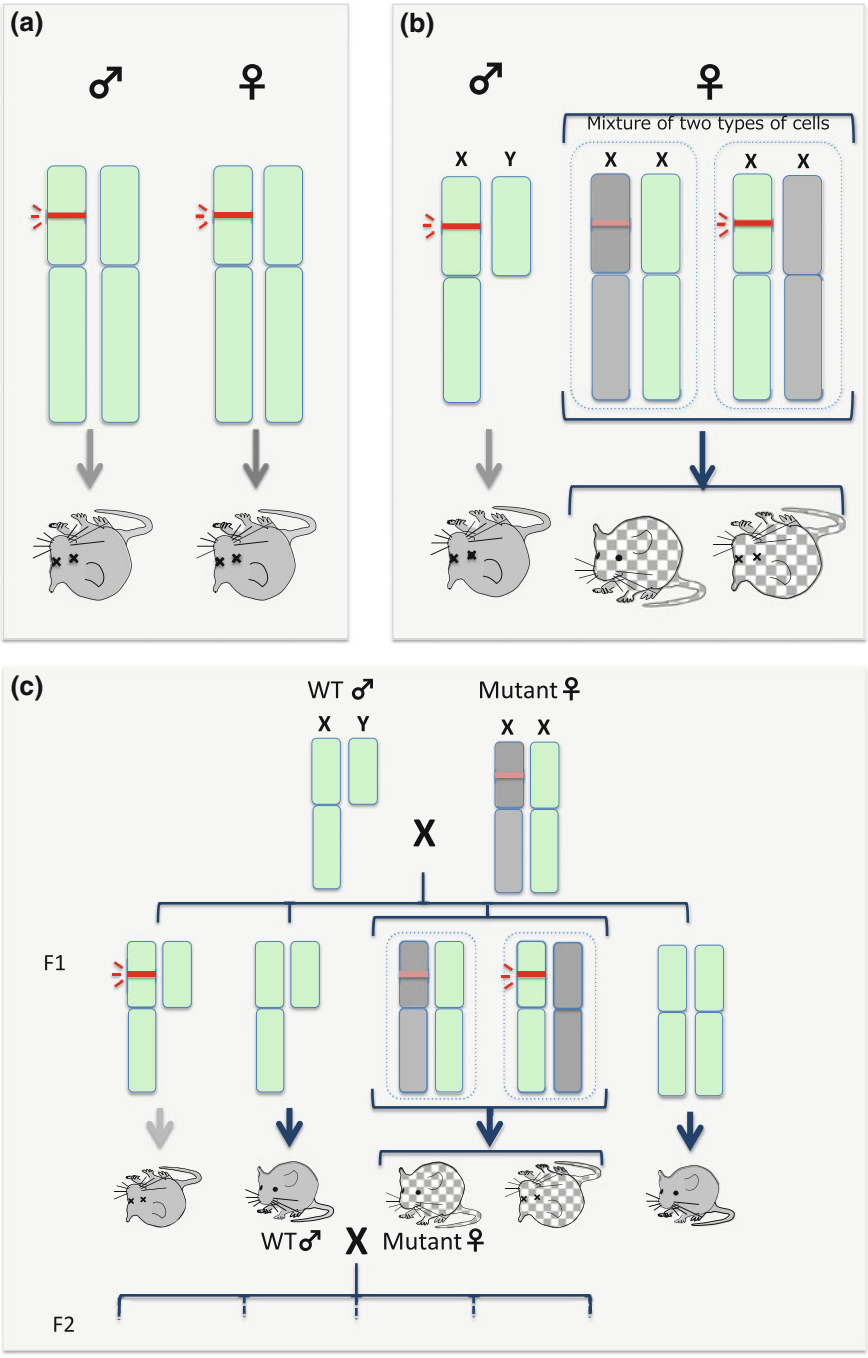
like neutral genes (Kaneko-Ishino and Ishino 2012, 2015). We hypothesize that gradual conversion from the silenced (potentially) harmful genes to the slightly advantageous genes occurred as a result of multiple mutations. Then, Darwinian selection shaped such slightly advantageous genes to be more advantageous and more functional for the host organisms (Kaneko-Ishino and Ishino 2012, 2015).

As shown in Fig. 1, eight out of 11 *SIRH* genes (*SIRH4-11*) are located on the X chromosome (Fig. 1). Is there any additional advantage for the acquisition of the LTR retrotransposon-derived genes on the X chromosome? We would like to propose the importance of another epigenetic mechanism, X chromosome inactivation (X-inactivation) (Lyon 1986), to explain the biased X chromosome location of the *SIRH* genes. In females, X-linked genes are typically subjected to random X-inactivation for gene dosage compensation, therefore, exhibiting monoallelic expression. As a result, most *SIRH* genes, except autosomal *SIRH3,* exhibit monoallelic expression similar to the imprinted *PEG10/SIRH1* and *PEG11/RTL1/SIRH2* genes which exhibit paternal-specific monoallelic expression.

It is reasonable to hypothesize that the integrated retrotransposons or retrotransposon derivatives would be present on only one of two homologous chromosomes and that they would be harmful and behave like dominant negative genes (Fig. 7). Then, in the case of autosomal integration, both males and females would be considerably affected and would exhibit lethality (Fig. 7a). In the case of X chromosomal integration, males would also be lethal. However, some females would have a chance to survive because X-inactivation could make such dangerous inserts silent and harmless, whereas others would be lethal because females comprise mixture of two types of the cells in terms of X-inactivation (Fig. 7b) and viability of individual would depend on which parts of somatic cells were rescued by random X-inactivation. As paradoxical as it may seem, it may be also important that all mutant males would die and only normal wild-type males could survive. Then, mutant females would always mate with normal healthy males and would reproduce some viable female mutants and wild-type male offspring from generation to generation (Fig. 7c).

In this scenario, even harmful DNA sequences could be stably maintained in a population by transmission through the heterogenous mutant females over a long period, allowing the accumulation of multiple mutations that is needed to generate advantageous genes. Thus, the X-linked genes may have some advantage to be selected. Once viable males with such slightly advantageous genes appeared, such genes would be propagated rapidly in both females and males and would finally be fixed in the population. Thus, we propose that X-inactivation in females could increase the chance of survival of the X-linked *SIRH* genes compared with the autosomal *SIRH* genes.

It is known that the mammalian X chromosome is of interest in terms of retroposition. Retroposition, reintegration of reverse-transcribed mRNAs into the genome, is an important mechanism of gene copying, giving rise approximately one-tenth of duplicated genes as retrogenes (Emerson et al. 2004; Khil et al. 2005). Emerson et al. reported that both the human and mouse X chromosomes harbor a substantial excess of genes that generate new retrocopies in the autosomes as well as


(a)
♂
♀
(b)
♂
♀
Mixture of two types of cells
X
Y
X
X
X
X
(c)
WT ♂
Mutant ♀
X
Y
X
X
X
F1
WT♂
X
Mutant ♀
F2

◂**Fig. 7** Hypothetical scenario for domestication of *SIRH* genes on the X chromosome. How did X-inactivation work in the domestication of X-linked *SIRH* genes? Comparison of the results of retrotransposon integration in an autosome (**a**) and the X chromosome (**b**). Harmful effects of the integrated retrotransposon would be reduced by X-inactivation (*shaded*). According to the nearly neutral theory of molecular evolution, a less harmful mutation can be fixed in a population by a random drift mechanism. The integrated retrotransposon would be maintained by transmission from heterogenous mutant females in each generation (**c**)

recruit an excess of functional copies from the autosomes (Emerson et al. 2004). They proposed two hypotheses to explain the biased X recruitment, one is a mechanical bias and the other is natural selection favoring the fixation and maintenance of retrogenes in the X chromosome (Emerson et al. 2004). As mentioned, some *SIRH* genes might be generated as retrogenes, thereby, were enriched on the X chromosome. Although the situations and mechanisms of retroposition and the tetrotransposon integration seem different, it is probable that the X-inactivation mechanism may have played an important role in the biased X recruitment of both the retrogenes and tetrotransposons as one of the mechanical biases as discussed above.

Acknowledgements This work was supported by the funding program for Next Generation World-Leading Researchers (NEXT Program) from the Japan Society for the Promotion of Science (JSPS) and the Asahi Glass Foundation to T.K.-I., Grants-in-Aid for Scientific Research (S) (23221010) and (A) (16H02478) from JSPS and Joint Usage/Research Program of Medical Research Institute Tokyo Medical and Dental University (TMDU) grants to F.I. and T.K.-I.

References

Akamatsu S et al (2015) The placental gene PEG10 promotes progression of neuroendocrine prostate cancer. Cell Rep 12:922–936

Aston-Jones G, Rajkowski J, Cohen J (1999) Role of locus coeruleus in attention and behavioral flexibility. Biol Psychiatry 46:1309–1320

Berridge CW, Waterhouse BD (2003) The locus coeruleus-noradrenergic system: modulation of behavioral state and state-dependent cognitive processes. Brain Res Rev 42:33–84

Bouret S, Sara SJ (2005) Network reset: a simplified overarching theory of locus coeruleus noradrenaline function. Trends Neurosci 28:574–582

Brandt J et al (2005) Transposable elements as a source of genetic innovation: expression and evolution of a family of retrotransposon-derived neogenes in mammals. Gene 345:101–111

Brosius J, Gould SJ (1992) On, "genomenclature": a comprehensive (and respectful) taxonomy for pseudo-genes and other "junk DNA". Proc Natl Acad Sci USA 89:10706–10710

Campillos M, Doerks T, Shah PK, Bork P (2006) Computational characterization of multiple Gag-like human proteins. Trends Genet 22:585–589

Cattanach BM, Kirk M (1985) Differential activity of maternally and paternally derived chromosome regions in mice. Nature 315:496–498

Cattanach BM, Beechey CV (1990) Autosomal and X-chromosome imprinting. Develop. Suppl 63–72

Charlier C et al (2001) Human-ovine comparative sequencing of a 250-kb imprinted domain encompassing the callipyge (clpg) locus and identification of six imprinted transcripts: DLK1, DAT, GTL2, PEG11, antiPEG11, and MEG8. Genome Res 11:850–862

Cornelis G et al (2015) Retrovial envelope gene captures and syncitin exaptation for placentation in marsupials. Proc Natl Acad Sci USA 112:E487–E496
Davis E et al (2005) RNAi-mediated allelic trans-interaction at the imprinted Rtl1/Peg11 locus. Curr Biol 15:743–749
Deng X et al (2014) PEG10 plays a crucial role in human lung cancer proliferation, progression, prognosis and metastasis. Oncol Rep 32:2159–2167
Dupressoir A et al (2005) Syncytin-A and syncytin-B, two fusogenic placenta-specific murine envelope genes of retroviral origin conserved in Muridae. Proc Natl Acad Sci USA 102:725–730
Edwards CA et al (2008) The evolution of the *DLK1-DIO3* imprinted domain in mammals. PLoS Biol 6:e135
Emerson JJ et al (2004) Extensive gene traffic on the mammalian X chromosome. Science 303:537–540
Georgiades P et al (2000) Parental origin-specific developmental defects in mice with uniparental disomy for chromosome 12. Development 127:4719–4728
Gould SJ, Vrba ES (1982) Exaptation; a missing term in the science of form. Paleobiology 8:4–15
Heidmann O et al. (2009) Identification of an endogenous retroviral envelope gene with fusogenic activity and placenta-specific expression in the rabbit: a new "*syncytin*" in a third order of mammals. Retrovirology 6: 107
Ioannides Y et al (2014) Temple syndrome: improving the recognition of an underdiagnosed chromosome 14 imprinting disorder: an analysis of 51 published cases. J Med Genet 51:495–501
Irie M et al (2015) Cognitive function related to the *Sirh11/Zcchc16* gene acquired from an LTR retrotransposon in eutherians. PLoS Genet 11:e1005521
Irie M et al (2016) An LTR retrotransposon-derived gene displays lineage-specific structural and putative species-specific functional variations in eutherians. Front Chem 4:26
Kagami M et al (2008) Deletions and epimutations affecting the human chromosome 14q32.2 imprinted region in individuals with paternal and maternal upd(14)-like phenotypes. Nat Genet 40:237–242
Kagami M et al (2015) Comprehensive clinical studies in 34 patients with molecularly defined UPD(14)pat and related conditions (Kagami-Ogata syndrome). Eur J Hum Genet 23:1488–1498
Kaneko-Ishino T, Ishino F (2010) Retrotransposon silencing by DNA methylation contributed to the evolution of placentation and genomic imprinting in mammals. Develop Growth Differ 52:533–543
Kaneko-Ishino T, Ishino F (2012) The role of genes domesticated from LTR retrotransposons and retroviruses in mammals. Front Microbiol 3:262
Kaneko-Ishino T, Ishino F (2015) Mammalian-specific genomic functions: newly acquired traits generated by genomic imprinting and LTR retrotransposon-derived genes in mammals. Proc Jpn Acad Ser B Phys Biol Sci 91: 511–538
Khil PP, Oliver B, Camerini-Otero RD (2005) X for intersection: retrotransposition both on and off the X chromosome is more frequent. Trends Genet 21:3–7
Kimura M (1968) Evolutionary rate at the molecular level. Nature 217:624–626
Kimura M (1983) The neutral theory of molecular evolution. Cambridge University Press, Cambridge, pp 305–327
Kitazawa et al. (2017) Severe damage to the placental fetal capillary network causes mid- to late fetal lethality and reduction in placental size in Peg11/Rtl1 KO mice. Genes Cells 22: 174–188
Kotzot D (2004) Maternal uniparental disomy 14 dissection of the phenotype with respect to rare autosomal recessively inherited traits, trisomy mosaicism, and genomic imprinting. Ann Genet 47:251–260
Lavialle C et al (2013) Paleovirology of 'syncytins', retroviral env genes exapted for a role in placentation. Philos Trans R Soc Lond B Biol Sci 368:20120507
Lyon MF (1986) X chromosomes and dosage compensation. Nature 320:313

Malassine A, Frendo J-L, Evain-Brion D (2003) A comparison of placental development and endocrine functions between the human and mouse model. Hum Reprod Update 9: 531–539
Mann JR, Lovell-Badge RH (1984) Inviability of parthenogenones is determined by pronuclei, not egg cytoplasm. Nature 310:66–67
McGrath J, Solter D (1984) Completion of mouse embryogenesis requires both the maternal and paternal genomes. Cell 37:179–183
Mi S et al (2000) Syncytin is a captive retroviral envelope protein involved in human placental morphogenesis. Nature 403:785–789
Nagasaki K et al (1999) Identification of a novel gene, LDOC1, down-regulated in cancer cell lines. Cancer Lett 140:227–234
Nakaya Y et al (2013) Fematrin-1 is involved in fetomaternal cell-to-cell fusion in Bovinae placenta and has contributed to diversity of ruminant placentation. J Virol 87:10563–10572
Naruse M et al (2014) *Sirh7/Ldoc1* knockout mice exhibit placental P4 overproduction and delayed parturition. Development 141:4763–4771
Ohta T (2002) Near-neutrality in evolution of genes and gene regulation. Proc Natl Acad Sci USA 99:16134–16137
Oldham MC, Horvath S, Geschwind DH (2006) Conservation and evolution of gene coexpression networks in human and chimpanzee brains. Proc Natl Acad Sci USA 103:17973–17878
Ono R et al (2001) A retrotransposon-derived gene, PEG10, is a novel imprinted gene located on human chromosome 7q21. Genomics 73:232–237
Ono R et al (2006) Deletion of *Peg10,* an imprinted gene acquired from a retrotransposon, causes early embryonic lethality. Nat Genet 38:101–106
Okabe H et al (2003) Involvement of PEG10 in human hepatocellular carcinogenesis through interaction with SIAH1. Cancer Res 63:3043–3048
Ono R et al (2011) Identification of tammar wallaby *SIRH12*, derived from a marsupial-specific retrotransposition event. DNA Res 18:211–219
Poulter R, Butler M (1998) A retrotransposon family from the pufferfish (*Fugu*) *Fugu rubripes*. Gene 215:241–249
Renfree MB (2010) Marsupials: placental mammals with a difference. Placenta 31(Suppl):S21–S26
Renfree MB, Suzuki S, Kaneko-Ishino T (2013) The origin and evolution of genomic imprinting and viviparity in mammals. Philos Trans R Soc Lond B Biol Sci 368: 20120151
Sara SJ (2009) The locus coeruleus and noradrenergic modulation of cognition. Nat Rev Neurosci 10:211–223
Seitz H et al (2003) Imprinted microRNA genes transcribed antisense to a reciprocally imprinted retrotransposon-like gene. Nat Genet 34:261–262
Sekita Y et al (2008) Role of retrotransposon-derived imprinted gene, *Rtl1,* in the feto-maternal interface of mouse placenta. Nat Genet 40:243–248
Shigemoto K et al (2001) Identification and characterisation of a developmentally regulated mammalian gene that utilises-1 programmed ribosomal frameshifting. Nucleic Acids Res 29:4079–4088
Surani MA, Barton SC, Norris ML (1984) Development of reconstituted mouse eggs suggests imprinting of the genome during gametogenesis. Nature 308:548–550
Suzuki S et al (2007) Retrotransposon silencing by DNA methylation can drive mammalian genomic imprinting. PLoS Genet 3:e55
Volff J, Kfrting C, Schartl M (2001) Ty3/Gypsy retrotransposon fossils in mammalian genomes: did they evolve into new cellular functions? Mol Biol Evol 18:266–270
Watson ED, Cross JC (2005) Development of structures and transport functions in the mouse placenta. Physiology 20:180–193
Youngson NA et al (2005) A small family of sushi-class retrotransposon-derived genes in mammals and their relation to genomic imprinting. J Mol Evol 61:481–490

The Life History of Domesticated Genes Illuminates the Evolution of Novel Mammalian Genes

Dušan Kordiš

Abstract Molecular domestications of transposable elements have occurred repeatedly during the evolution of eukaryotes. Mammals possess numerous single copy domesticated genes that have originated from the intronless multicopy transposable elements. The genesis and regulatory wiring of the Metaviridae-derived domesticated genes have been explained through phylogenomic analysis of more than 90 chordate genomes. Phylogenomic analysis has demonstrated that major diversification of these domesticated genes occurred in the ancestor of placental mammals. Mammalian domesticated genes have originated in several steps by independent domestication events. The analysis of active Metaviridae lineages in amniotes has demonstrated that domesticated genes originated from retroelement remains. The analysis of syntenic loci has shown that diverse domesticated genes and their chromosomal positions were fully established in the ancestor of placental mammals. During the domestication process, de novo acquisition of regulatory regions was crucial for the survival of the novel domesticated genes. The origin and evolution of de novo acquired promoters and untranslated regions in diverse mammalian domesticated genes have been explained by comparative analysis of orthologous gene loci. The origin of placental mammal-specific innovations and adaptations, such as placenta and newly evolved brain functions, was most probably connected to the regulatory wiring of domesticated genes and their rapid fixation in the ancestor of placental mammals.

1 Introduction

Mammals possess numerous single copy domesticated genes that have originated from intronless multicopy retroelements (Mi et al. 2000; Llorens and Marin 2001; Lynch and Tristem 2003; Gorinšek et al. 2004, 2005; de Parseval and Heidmann

D. Kordiš (✉)
Department of Molecular and Biomedical Sciences,
Josef Stefan Institute, Ljubljana, Slovenia
e-mail: dusan.kordis@ijs.si

P. Pontarotti (ed.), *Evolutionary Biology: Self/Nonself Evolution, Species and Complex Traits Evolution, Methods and Concepts*,
DOI 10.1007/978-3-319-61569-1_8

2005; Brandt et al. 2005a, b; Kordiš 2005, 2009, 2011; Zdobnov et al. 2005; Campillos et al. 2006; Volff 2006; Warren et al. 2015; Naville et al. 2016) and DNA transposons (Volff 2006; Feschotte and Pritham 2007; Sinzelle et al. 2009; Kordiš 2011; Mateo and Gonzalez 2014; Huang et al. 2016; Duan et al. 2017). Domestication may require additional mutations that modify expression of the gene and the specificity of interaction of the recruited protein with nucleotide sequences or other proteins (Volff 2006). During the domestication process, de novo acquisition of the regulatory regions is a prerequisite for the survival of domesticated genes. Molecular domestication of transposases, integrases, reverse transcriptases, and envelope proteins has occurred repeatedly during the evolution of diverse major eukaryote lineages, and, during neofunctionalization, some of the newly obtained functions have become essential for the survival of the organism (Miller et al. 1999; Volff 2006). Although the functions of the majority of domesticated genes are still unknown (Campillos et al. 2006; Volff 2006; Feschotte and Pritham 2007; Sinzelle et al. 2009; Mita and Boeke 2016; Chuong et al. 2017), some may protect against infections, some are necessary for reproduction, whereas others enable the replication of chromosomes and the control of cell proliferation and apoptosis (Naruse et al. 2014; Irie et al. 2015; Ito et al. 2015; Kitazawa et al. 2017).

During evolution, many cellular protein-coding genes originated from genes carried by long terminal repeat (LTR) retroelements (retroviruses and LTR retrotransposons). LTR retroelements have contributed different types of coding regions to the gene repertoire of their host, including gag, envelope, integrase, and protease genes (Mi et al. 2000; de Parseval and Heidmann 2005; Campillos et al. 2006; Volff 2006; Lavialle et al. 2013; Warren et al. 2015; Naville et al. 2016). Numerous Metaviridae (Ty3/Gypsy)-derived genes have been discovered in the human genome and classified into five distinct families: SASPase (ASPRV1), Sushi (=Mart), SCAN, Paraneoplastic (PNMA), and ARC (Brandt et al. 2005a; Campillos et al. 2006; Emerson and Thomas 2011) (Fig. 1). Large amounts of data concerning the mammalian retroelement-derived domesticated genes have been generated, such as gene structures, chromosome locations, potential biological functions, potential

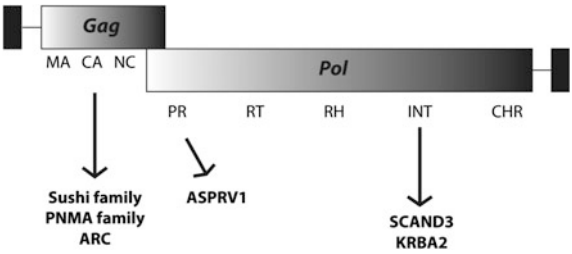


Fig. 1 Families of Metaviridae-derived domesticated genes. Analyzed domesticated genes originated from the Metaviridae (Ty3/Gypsy) group of LTR retrotransposons. Protein domains that are present in the Metaviridae retroelements are the following: matrix (*MA*), capsid (*CA*), nucleocapsid (*NC*), protease (*PR*), reverse transcriptase (*RT*), ribonuclease H (*RH*), integrase (*INT*), and chromodomain (*CHR*)

interacting partners, preliminary developmental expression analysis of a Mart family, and a first insight into their origin and evolution (Llorens and Marin 2001; Lynch and Tristem 2003; Gorinšek et al. 2004; Brandt et al. 2005a, b; Zdobnov et al. 2005; Campillos et al. 2006; Naruse et al. 2014; Irie et al. 2015; Ito et al. 2015; Kitazawa et al. 2017).

However, the evolutionary history and dynamics of domesticated genes have been only partially explored, due to the absence of genome data or due to the limited analysis of a single family of domesticated genes (Llorens and Marin 2001; Lynch and Tristem 2003; Gorinšek et al. 2004; Brandt et al. 2005a, b; Zdobnov et al. 2005; Campillos et al. 2006). The genesis and evolution of domesticated genes have been recently explained through comparative genomic and phylogenomic analyses (Kokošar and Kordiš 2013). This study has provided crucial information as to where and when Metaviridae gag, retroelement protease, and integrase domains were transformed into domesticated genes. The analysis of diverse domesticated genes in mammals has clarified their origins and evolution, and provided key insights into their regulatory and functional diversification. Our study has demonstrated that the regulatory wiring of domesticated genes and their rapid fixation in the ancestor of placental mammals have played an important role in the origin of their innovations and adaptations, such as placenta and newly evolved brain functions. We have mapped the life history of domesticated genes, from birth, their fixation in the genome, gain of regulatory elements, and structural complexity to complete integration into the functional network of the cell. Our study has demonstrated the utility of molecular domestication as a good model for understanding the origination and functional evolution of novel genes. This chapter aims to cover the most exciting insights obtained from our study about the origin, distribution, diversity, and evolution of domesticated genes in mammals (Kokošar and Kordiš 2013).

2 Domesticated Genes Originated from Retroelement Remains

The origin of domesticated genes and the transition point from Metaviridae to domesticated genes have been elucidated through the analysis of diverse Metaviridae lineages in Deuterostomia (Kokošar and Kordiš 2013). Numerous active Metaviridae lineages (represented in the genome by the full-length elements) are still present in diverse reptilian genomes (e.g., in Anolis and turtles), but not in any of bird or mammalian genomes. Since the active Metaviridae lineages are present in reptiles (sauropsids), the sister group of synapsids, they were present also in the ancestor of Amniota (Kokošar and Kordiš 2013; Kordiš et al. 2006; Kordiš 2009). Synapsids, with the mammals as the only surviving lineage of this large taxonomic group, have evidently lost all the active Metaviridae elements. In the ancestors of modern mammals, only rare Metaviridae remains (in the form of highly

fragmented molecular fossils) have persisted (Kokošar and Kordiš 2013). It has been demonstrated which Metaviridae clades are still present and active in reptiles, the sister group of mammals. Using this approach, it was possible to find the progenitors of domesticated genes and the transition point from Metaviridae to domesticated genes in amniotes. Gag- and integrase-derived domesticated genes originated from Metaviridae remains (molecular fossils), because no active Chromovirus or Barthez lineages of Metaviridae are present in any mammalian genome (Kokošar and Kordiš 2013).

3 Numerous Complex Mechanisms Were Involved in the Process of Neofunctionalization

The time frame from 250 million years ago (end-Permian mass extinction) (Benton and Twitchett 2003) to 160 million years ago (the origin of placental mammals—when progenitors of domesticated genes started to diversify) (Meredith et al. 2011) is very important for explaining the origin of domesticated genes. It demonstrates that a very long time (90–100 million years) was necessary for establishing the first domesticated genes. Why was this process so slow and complex? In the transition phase from the retroelement remains to the first domesticated genes, many nucleotide changes were necessary for neofunctionalization (Lynch and Conery 2000; Long et al. 2003; Krull et al. 2007); such a process could be quite rapid, due to the initial functional diversification by adaptive evolution. The analysis of domesticated genes in monotreme, marsupial, and placental genomes has demonstrated that these genes were fixed in the ancestor of placental mammals (Kokošar and Kordiš 2013) and evolved under strong purifying selection (Brandt et al. 2005a) to preserve the important newly gained functions. One of the crucial steps in the process of neofunctionalization was the exonization (Sorek 2007; Schmitz and Brosius 2011) of retroelement domains (gag, protease, and integrase), which produced ready-to-use modules (Fig. 2). Such a process was probably quite slow. As retroelement remains lacked regulatory regions, their acquisition (Castillo-Davis 2004; Kaessmann et al. 2009; Kaessmann 2010), such as the acquisition of 5′-untranslated regions (UTRs) and 3′-UTRs, has been very important for the survival of exapted sequences. Even more important was the simultaneous intron gain into 5′-UTRs and promoter acquisition, which enabled the regulatory wiring of diverse domesticated genes (Kordiš 2011; Kordiš and Kokošar 2012). It is well documented that domesticated genes exhibit highly restricted and specialized tissue-specific expression in brain, testis, and placenta (Brandt et al. 2005a; Schüller et al. 2005; Takaji et al. 2009; Kaneko-Ishino and Ishino 2012), which was only possible through the *cis*-regulatory evolution. Subsequent gene duplications and chromosomal gene movements have further diversified domesticated genes and enabled the acquisition of novel (more specialized or more diversified) biological functions.

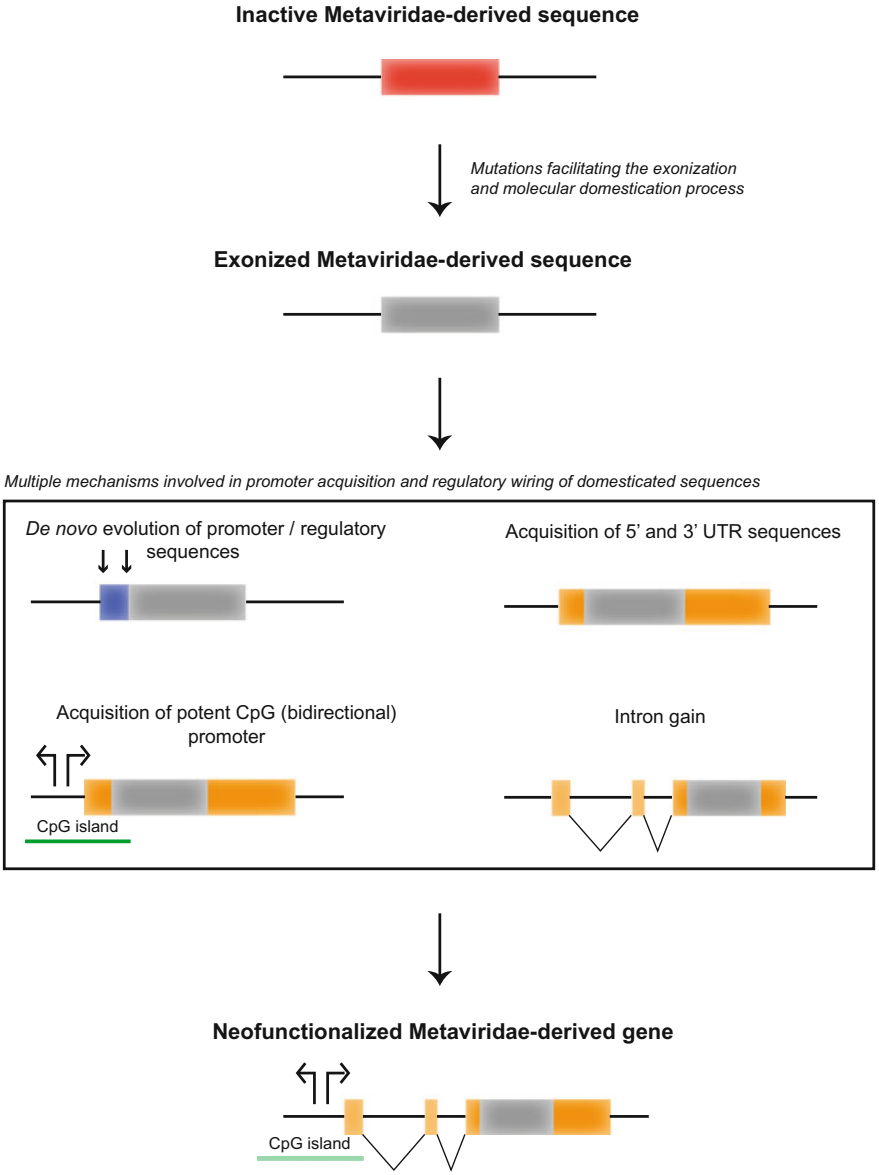


Fig. 2 Mechanisms involved in the process of neofunctionalization of domesticated genes. In the transition phase from retroelement remains to the first domesticated genes, many nucleotide changes were necessary for the neofunctionalization. One of the crucial steps in the process of neofunctionalization was the exonization of retroelement domains (gag, protease, and integrase), which produced ready-to-use modules. Retroelement remains in mammalian genomes will normally turn into pseudogenes, due to lack of a promoter, and they can survive as a functional gene only if they recruit a new promoter sequence. To become expressed at a significant level and in the tissues where it can exert a selectively beneficial function, a new gene needs to acquire a core promoter and other structural elements that regulate its expression. Exons and introns are shown as *orange* (5′ and 3′ UTR regions) or *gray* (coding part of the exons) *boxes* and *connecting lines*. A de novo acquired promoter is shown in *blue*

4 Quick Fixation of Domesticated Genes in the Ancestor of Placental Mammals

Phylogenomic analysis (Eisen 1998; Delsuc et al. 2005) has enabled characterization of domesticated genes from numerous mammalian (>50 different species), key tetrapod (amphibians and reptiles), and the remaining chordate genomes. In total, more than 90 chordate genomes were analyzed. Phylogenomic analysis of domesticated genes has shown for the first time which domesticated genes are present in the genomes of monotremes, marsupials, and all three placental superorders, and which of the domesticated genes are also present in other vertebrate genomes (Kokošar and Kordiš 2013). In addition to the mammalian genomes, all other available vertebrate and chordate genomes were analyzed to find the transition point from retroelements to domesticated genes. Phylogenomic analysis of all available domesticated genes in mammalian, vertebrate, and chordate genomes has revealed unequivocal data about their origins (when and in which taxonomic group they originated), together with numerous gene-related information (exon/intron structure, genome location, chromosome position, etc.) (Kokošar and Kordiš 2013).

Phylogenomic analysis of the domesticated genes in Tetrapoda has shown few originations in the ancestor of tetrapods (ARC), the ancestor of mammals (ASPRV1), and the ancestor of Theria (PNMA progenitor, PEG10, SIRH12, NYNRIN, KRBA2, and SCAND3). Only a few cases of bursts in functional diversification of domesticated genes have occurred, a major one being in the ancestor of placental mammals (Sushi-derived domesticated genes and PNMA-derived domesticated genes) (Kokošar and Kordiš 2013) (Fig. 3). First diversification of Sushi and PNMA gene families has occurred in the ancestor of placental mammals, when all 20 orthologous genes emerged. They differ in their expression profiles and tissue specificities (Brandt et al. 2005a; Takaji et al. 2009). Phylogenetic and sequence analysis of Chromovirus-derived domesticated genes has provided strong evidence that they originated independently several times in the ancestor of placentals. The greatest number of orthologous domesticated genes is present in the genomes of placental mammals which possess at least 27 orthologous domesticated genes. These orthologous genes have remained conserved throughout the placental mammals. Within mammals, a large difference between placentals and ancestral mammalian lineages (Prototheria and Metatheria) is clearly evident, as the latter possess only 3–10 orthologous domesticated genes (Kokošar and Kordiš 2013).

The phylogenomic analysis of domesticated genes in all extant mammalian lineages has provided a definitive answer to the timing of retroelement domestication (Kokošar and Kordiš 2013; Kordiš 2011). Domesticated genes originated stepwise by independent domestication events and later diversified through gene duplications. The analysis of all domesticated genes in chordates and mammals has shown that the greatest number of domesticated genes was fixed in the ancestor of placentals, as demonstrated by their presence and sequence conservation in all placental superorders. This kind of analysis has provided a temporal component,

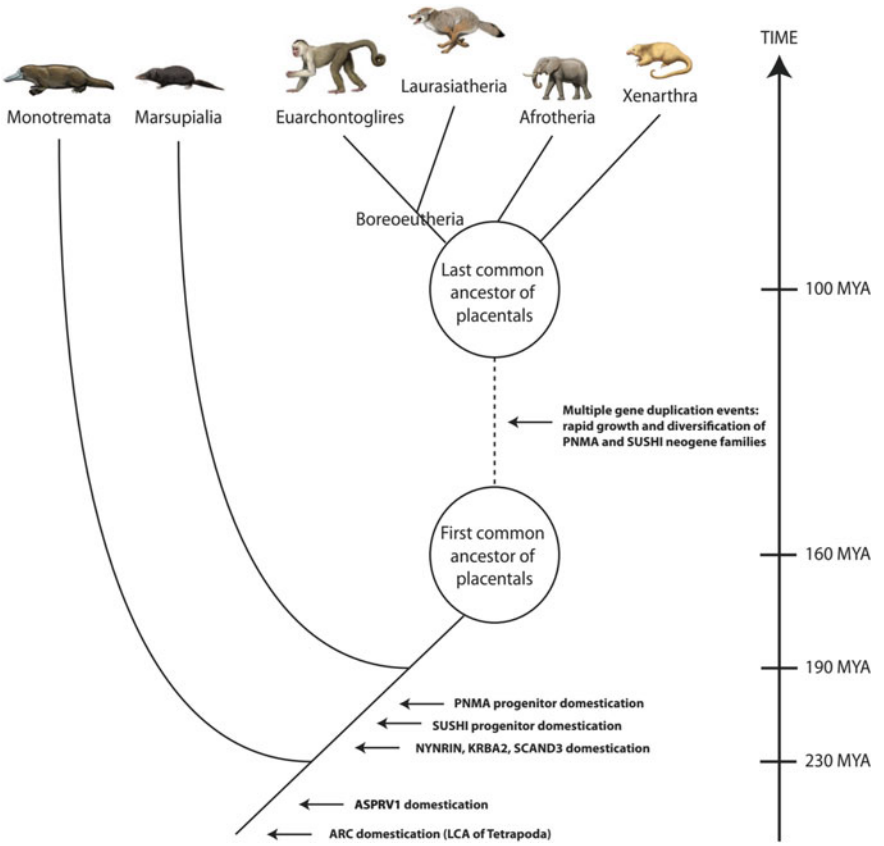


Fig. 3 Burst of domesticated gene originations in the ancestor of placental mammals. The phylogenomic analysis of domesticated genes in all extant mammalian lineages has provided a definitive answer to the timing of domestication of Metaviridae retroelements. Domesticated genes originated in several steps by independent domestication events and were later diversified by gene duplications. Phylogenomic analysis of domesticated genes has shown which genes are present in the genomes of monotremes, marsupials, and all three placental superorders, and which of the domesticated genes are also present in other vertebrate genomes. *MYA* million years ago

because it has been determined precisely when a particular domesticated gene or domesticated gene family originated. From the comparison of syntenic positions between multiple mammalian lineages, the ancestral states of the chromosomal positions of domesticated genes have been reconstructed (Kokošar and Kordiš 2013). Analysis of conserved syntenies has shown clearly that the initial emergence and subsequent diversification of numerous domesticated genes occurred in the ancestor of placental mammals. Conserved synteny has thus demonstrated that diverse domesticated genes and their chromosomal positions were fully established in the ancestor of placental mammals (Kokošar and Kordiš 2013) (Fig. 4).

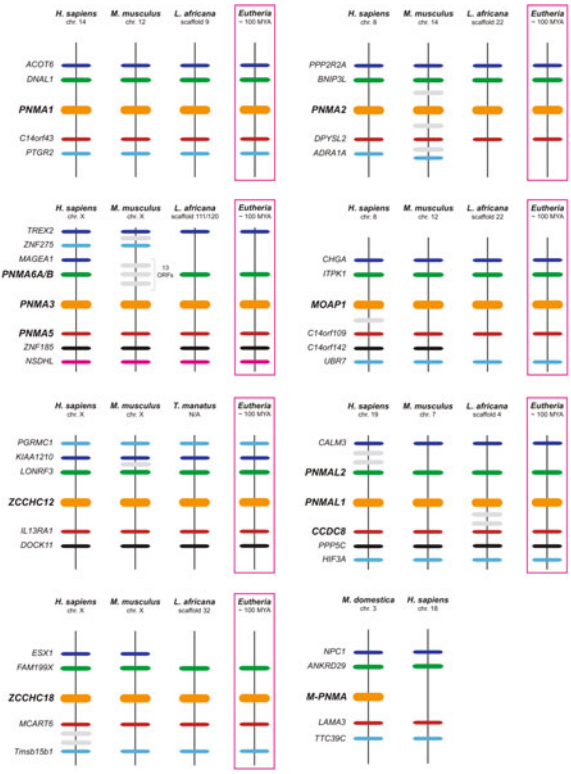


Fig. 4 Conserved synteny in the PNMA family of domesticated genes. Chromosomal regions carrying all domesticated genes in the species considered were compared, and neighboring genes with conserved synteny were identified. Horizontal lines denote orthologous relationships. Each domesticated gene is represented in *bold* as a *horizontal orange line* on the chromosome. Neighboring genes that are in synteny are shown with a schematic indication of their distance (not to scale). The ancestral states of the chromosomal positions of domesticated genes were reconstructed from a comparison of syntenic positions between multiple mammalian lineages

5 All Regulatory Regions in Domesticated Genes Have Been Acquired De Novo

New coding sequences can be generated by the recruitment of new regions (a new 5′-UTR, 3′-UTR, promoter, new introns) and gene fusions (Long et al. 2003; Fablet et al. 2009; Kaessmann et al. 2009; Kaessmann 2010; Long et al. 2013; Chen et al. 2013). The resulting gene can remain in a genome and gain a new function. Retroelement remains in mammalian genomes will normally turn into pseudogenes, due to lack of a promoter, and they can survive as a functional gene only if they recruit a new promoter sequence. Because of the very small likelihood that retroelement remains (consisting either from the protease, gag, or integrase

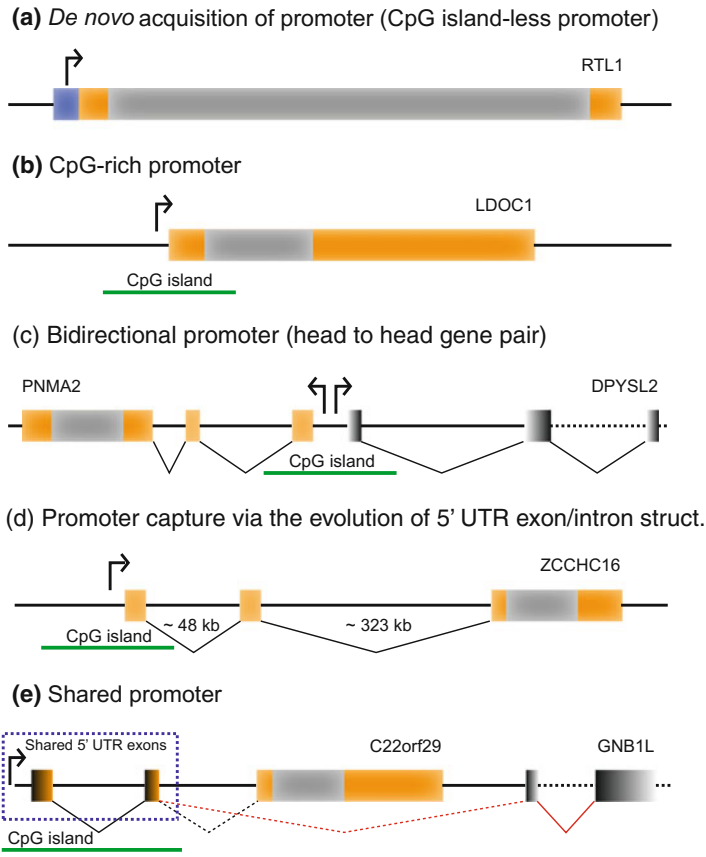


Fig. 5 Diverse sources of domesticated gene promoters. Various scenarios that lead to the transcription of domesticated gene copies are illustrated. **a** Recruitment of proto-promoters from the CpG island-less region. **b** Recruitment of proto-promoters from the CpG-rich island. **c** Recruitment of a bidirectional (CpG-enriched) promoter from neighboring gene in the vicinity of the domesticated gene. **d** Recruitment of distant promoters in the genomic neighborhood via the acquisition of a new 5′ untranslated exon–intron structure. **e** Sharing of the unidirectional (CpG-enriched) promoter from a neighboring gene in the vicinity of the domesticated gene. Exons and introns are represented by *orange* and *gray* (domesticated genes) or *black* (neighboring genes in the case of bidirectional promoters) *boxes* and *connecting lines*. Distances between exons are not to scale

domains) will acquire a promoter sequence, either de novo or from a preexisting gene (e.g., bidirectional promoters), the possibility that they become a new functional gene is limited. A mechanism by which a promoter sequence can be obtained is therefore critical for the generation of functional domesticated genes (Long et al. 2003; Fablet et al. 2009; Kaessmann et al. 2009; Kaessmann 2010). Since in retroelements, gag, protease, and integrase domains lack promoters and UTRs,

they must have been acquired de novo in domesticated genes. The origin and evolution of de novo acquired promoters, 5′- and 3′-UTRs in diverse mammalian domesticated genes has been explained by comparative analysis of orthologous gene loci (Kokošar and Kordiš 2013). The mechanisms for the acquisition of regulatory regions (Fig. 5) were found to be very similar to those observed in retrogenes (Long et al. 2003; Vinckenbosch et al. 2006; Shiao et al. 2007; Fablet et al. 2009; Kaessmann et al. 2009; Kaessmann 2010) and are outlined below.

5.1 *De Novo Acquisition of Promoters and 5'-UTRs in Domesticated Genes*

The presence of numerous functional domesticated genes in mammals (Campillos et al. 2006; Volff 2006) immediately raises the question as to how they obtained the regulatory sequences that enable them to be transcribed—a precondition for gene functionality. To become expressed at a significant level and in the tissues where it can exert a selectively beneficial function, a new gene needs to acquire a core promoter and other structural elements that regulate its expression. Various sources of promoters and regulatory sequences exist and provide general insights into how new genes can acquire promoters and evolve new expression patterns (Fablet et al. 2009; Kaessmann et al. 2009; Kaessmann 2010). The expression of domesticated genes may benefit from preexisting regulatory machinery and expression capacities of genes in their vicinity. Transcribed domesticated genes are often located close to other genes, suggesting that their transcription could be made possible by open chromatin and/or regulatory elements of nearby genes (Kokošar and Kordiš 2013). This possibility is supported by the observations that domesticated genes may be transcribed from the bidirectional promoters of neighboring genes (Kalitsis and Safferty 2009).

5.2 *Domesticated Genes Exhibit Numerous Ways of Obtaining Regulatory Regions*

Analysis of the promoters (Table 1) has shown that only a small proportion of domesticated genes (5 genes, 18%) have captured bidirectional promoters. A large majority of domesticated genes (19 genes, 68%) have recruited proto-promoters from the CpG-rich islands in their genomic vicinity (Kokošar and Kordiš 2013). Some of the domesticated genes promoters may have evolved de novo by small substitutional changes under the influence of natural selection. In seven domesticated genes (25%), the process of promoter acquisition has involved the evolution of new 5′ untranslated exon–intron structures, which often span substantial distances between the recruited promoters and domesticated genes. By the acquisition

of new 5′-UTR structures, domesticated genes might also become transcribed from distant CpG-enriched sequences, which often have the inherent capacity to promote transcription, and were not previously associated with other genes. The primary role, and selective benefit, of newly gained 5′-UTR introns has been to span the substantial distances to potent CpG promoters, driving transcription of domesticated genes and reducing the size of the UTR exons (Kokošar and Kordiš 2013; Kordiš 2011; Kordiš and Kokošar 2012).

5.3 *Independent Cis-Regulatory Evolution of Domesticated Genes*

Analysis of transcription factor binding sites in promoters of domesticated genes has shown a large diversity between genes or between human and mouse orthologous genes, indicating that *cis*-regulatory evolution was responsible for the large differences in expression patterns of domesticated genes (Kokošar and Kordiš 2013). The frequent inheritance of CpG promoters explains why a significant number of domesticated genes evolved paternally or maternally imprinted expression (Campillos et al. 2006; Volff 2006; Kaneko-Ishino and Ishino 2012).

5.4 *De Novo Acquisition of 3′-UTRs in Domesticated Genes*

The analysis of all known domesticated genes in chordates and mammals has shown that they contain newly acquired 3′-UTRs. The availability of human and mouse RefSeq genes and numerous mammalian genomes has enabled the analysis of the length of 3′-UTRs in domesticated genes. De novo acquired 3′-UTRs in placental mammals have shown large variation in length, the shortest are present in the human SCAND3 gene (281 bp) and the longest in the mouse PEG10 gene (5,166 bp). The mean 3′-UTR length in humans is approximately 520 bp (Grillo et al. 2010), but such lengths are present only in three human domesticated genes (ZCCHC12, ZCCHC18, and ASPRV1) and only two domesticated genes are shorter (SCAND3 and KRBA2). The great majority of human or mouse domesticated genes have much longer 3′-UTRs, eight are shorter than 1,000 bp, seven are in the range of 1,000–2,000 bp, six in the range of 2,000–3,000 bp, one is longer than 4 kb, and two longer than 5 kb. Searching for TEs in the unusually long 3′-UTRs with RepeatMasker has shown the absence of species-specific repeats in the analyzed species (Kokošar and Kordiš 2013). What is the reason for such increased lengths of the 3′-UTRs of domesticated genes? Although housekeeping genes possess significantly shorter coding and untranslated sequences than the tissue-specific genes (She et al. 2009), the lengths of the 3′-UTRs of tissue-specific genes have drastically increased (Stark et al. 2005). Domesticated genes may be an

exception, because the longest 3′-UTRs are found in the housekeeping domesticated genes that are expressed in the majority of tissues tested. It is likely, therefore, that all domesticated genes have recruited nearby genomic regions as 5′- or 3′-UTRs. The consequence of the very long 3′-UTRs in some domesticated genes is that the lengths of the 3′-UTR exons are greatly increased (Kokošar and Kordiš 2013).

5.5 *De Novo Acquired 3′-UTRs Have Enabled Translational Control of Gene Expression*

The 3′-UTRs are important posttranscriptional regulatory regions of mRNAs that possess numerous regulatory elements and are vital for correct spatial and temporal gene expression. They have been found to be involved in diverse regulatory processes, including transcript cleavage, stability and polyadenylation, translation, and mRNA localization. RNA-binding proteins and miRNAs bind to *cis*-acting sequences within 3′-UTRs to influence mRNA stability, translation, and localization. 3′-UTRs are thus critical in determining the fate of an mRNA (Andreassi and Riccio 2009; Barrett et al. 2012). De novo recruitment of 3′-UTRs may therefore lead to the increased RNA stability and translation efficiency of domesticated genes. Because all domesticated genes have recruited adjacent sequences as their 3′-UTRs, these de novo acquired 3′-UTRs may play an important role in regulating individual mRNA stability in response to intracellular and extracellular cues (Kokošar and Kordiš 2013).

6 Domesticated Genes as Drivers of Phenotypic Evolution in Placental Mammals

The burst of domesticated gene origination took place in the ancestor of placentals (Kokošar and Kordiš 2013; Kordiš 2011). This has important implications for explaining the origin of numerous phenotypic novelties in placentals. Domesticated genes, originating from the junk DNA or from the molecular fossils of Metaviridae, have been crucially involved in, or even promoted the development of phenotypic novelties such as placenta (Kaneko-Ishino and Ishino 2012) and neocortex (Oldham 2006). It is somehow surprising that domesticated genes participated in numerous brain/central nervous system (CNS)-connected functions and in reproduction. Some of the very important functions emerged even earlier in the ancestor of Tetrapoda with the ARC gene, which plays a crucial role in the synaptic plasticity and the long-term memory (Korb and Finkbeiner 2011). The emergence of the orthologous domesticated gene families in placental mammals is most probably connected to the origin of their innovations and adaptations, such as placenta and newly evolved brain

Table 1 Newly recruited promoters of retroelement-derived domesticated genes

Retroelement progenitor	Gene name	5'-UTR introns	CpG island/ proto-promoter	bidirectional promoter	not associated with promoter CpG island
Chromovirus					
	RGAG1	**Yes**			•
	RGAG4	**No**	•		
	PEG10	**Yes**		•	
	RTL1	**No**			•
	LDOC1	**No**	•		
	LDOC1L	**Yes**	•		
	FAM127A/B/C	**No**	•		
	C22orf29	**Yes**	•		
	ZCCHC5	**Yes**			•
	ZCCHC16	**Yes**	•		
Barthez					
	PNMA1	**No**	•		
	PNMA2	**Yes**		•	
	MOAP1	**Yes**		•	
	PNMA3	**Yes**	•		
	PNMA5	**Yes**			•
	PNMA6A/B	**Yes**	•		
	ZCCHC12	**Yes**	•		
	ZCCHC18	**Yes**	•		
	PNMAL1	**Yes**	•		
	PNMAL2	**No**	•		
	CCDC8	**No**	•		
Gmr1					
	GIN1	**Yes**		•	
	GIN2	**No**		•	
	KRBA2	**Yes**			•
	SCAND3	**No**			•
Osvaldo					
	ARC	**No**	•		
Cigr2					
	ASPRV1	**No**			•
ERV					
	NYNRIN	**Yes**	•		

Note The type of promoter is marked with the black dot

functions (Campillos et al. 2006; Oldham 2006; Volff 2006; Kaneko-Ishino and Ishino 2012; Kokošar and Kordiš 2013; Warren et al. 2015; Naville et al. 2016). In the majority of orthologous domesticated genes, the prevalent trend was loss of the ancestral activity and acquisition of a novel function (neofunctionalization). Although some of these orthologous domesticated genes still possess the conserved gag, integrase, and protease domains, they have lost ancestral activity due to mutations in structurally important regions. The number of newly gained functions in the domesticated genes indicates that the gag, integrase, and protease domains are highly versatile protein–protein interaction modules that can readily interact with novel targets (Campillos et al. 2006; Volff 2006; Feschotte 2008; Kokošar and Kordiš 2013). Newly emerged domesticated genes may evolve new functional roles through adaptive evolution of encoded proteins and/or by developing new spatial or

temporal expression patterns (Kokošar and Kordiš 2013). There is growing evidence that some domesticated genes (e.g., LDOC1) are involved in the gradual growth of CNS interaction networks in the particularly active regions of brain (neocortex)—not only during the evolution of placentals, but also in very recent times, that is, after the split of Homo and chimpanzee lineages (Oldham et al. 2006).

Acknowledgements This work was supported by grant P1-0207 from the Slovenian Research Agency.

References

Andreassi C, Riccio A (2009) To localize or not to localize: mRNA fate is in 3′UTR ends. Trends Cell Biol 19:465–474

Barrett LW, Fletcher S, Wilton SD (2012) Regulation of eukaryotic gene expression by the untranslated gene regions and other non-coding elements. Cell Mol Life Sci 69:3613–3634

Benton MJ, Twitchett RJ (2003) How to kill (almost) all life: the end-Permian extinction event. Trends Ecol Evol 18:358–365

Brandt J, Schrauth S, Veith AM, Froschauer A, Haneke T, Schultheis C, Gessler M, Leimeister C, Volff JN (2005a) Transposable elements as a source of genetic innovation: expression and evolution of a family of retrotransposon-derived neogenes in mammals. Gene 345:101–111

Brandt J, Veith AM, Volff JN (2005b) A family of neofunctionalized Ty3/*Gypsy* retrotransposon genes in mammalian genomes. Cytogenet Genome Res 110:307–317

Campillos M, Doerks T, Shah PK, Bork P (2006) Computational characterization of multiple Gag-like human proteins. Trends Genet 22:585–589

Castillo-Davis CI, Hartl DL, Achaz G (2004) *cis*-Regulatory and protein evolution in orthologous and duplicate genes. Genome Res 14:1530–1536

Chen S, Krinsky BH, Long M (2013) New genes as drivers of phenotypic evolution. Nat Rev Genet 14:645–660

Chuong EB, Elde NC, Feschotte C (2017) Regulatory activities of transposable elements: from conflicts to benefits. Nat Rev Genet 18:71–86

de Parseval N, Heidmann T (2005) Human endogenous retroviruses: from infectious elements to human genes. Cytogenet Genome Res 110:318–332

Delsuc F, Brinkmann H, Philippe H (2005) Phylogenomics and the reconstruction of the tree of life. Nat Rev Genet 6:361–375

Duan CG, Wang X, Xie S, Pan L, Miki D, Tang K, Hsu CC, Lei M, Zhong Y, Hou YJ, Wang Z, Zhang Z, Mangrauthia SK, Xu H, Zhang H, Dilkes B, Tao WA, Zhu JK (2017) A pair of transposon-derived proteins function in a histone acetyltransferase complex for active DNA demethylation. Cell Res 27:226–240

Eisen JA (1998) Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis. Genome Res 8:163–167

Emerson RO, Thomas JH (2011) *Gypsy* and the birth of the SCAN domain. J Virol 85:12043–12052

Fablet M, Bueno M, Potrzebowski L, Kaessmann H (2009) Evolutionary origin and functions of retrogene introns. Mol Biol Evol 26:2147–2156

Feschotte C (2008) Transposable elements and the evolution of regulatory networks. Nat Rev Genet 9:397–405

Feschotte C, Pritham EJ (2007) DNA transposons and the evolution of eukaryotic genomes. Annu Rev Genet 41:331–368

Gorinšek B, Gubenšek F, Kordiš D (2004) Evolutionary genomics of chromoviruses in eukaryotes. Mol Biol Evol 21:781–798

Gorinšek B, Gubenšek F, Kordiš D (2005) Phylogenomic analysis of chromoviruses. Cytogenet Genome Res 110:543–552

Grillo G, Turi A, Licciulli F, Mignone F, Liuni S, Banfi S, Gennarino VA, Horner DS, Pavesi G, Picardi E, Pesole G (2010) UTRdb and UTRsite (RELEASE 2010): a collection of sequences and regulatory motifs of the untranslated regions of eukaryotic mRNAs. Nucleic Acids Res 38 (Database issue):D75–D80

Huang S, Tao X, Yuan S, Zhang Y, Li P, Beilinson HA, Zhang Y, Yu W, Pontarotti P, Escriva H, Le Petillon Y, Liu X, Chen S, Schatz DG, Xu A (2016) Discovery of an active RAG transposon illuminates the origins of V(D)J recombination. Cell 166:102–114

Irie M, Yoshikawa M, Ono R, Iwafune H, Furuse T, Yamada I, Wakana S, Yamashita Y, Abe T, Ishino F, Kaneko-Ishino T (2015) Cognitive function related to the Sirh11/Zcchc16 gene acquired from an LTR retrotransposon in eutherians. PLoS Genet 11:e1005521

Ito M, Sferruzzi-Perri AN, Edwards CA, Adalsteinsson BT, Allen SE, Loo TH, Kitazawa M, Kaneko-Ishino T, Ishino F, Stewart CL, Ferguson-Smith AC (2015) A trans-homologue interaction between reciprocally imprinted miR-127 and Rtl1 regulates placenta development. Development 142:2425–2430

Kaessmann H (2010) Origins, evolution, and phenotypic impact of new genes. Genome Res 20:1313–1326

Kaessmann H, Vinckenbosch N, Long M (2009) RNA-based gene duplication: mechanistic and evolutionary insights. Nat Rev Genet 10:19–31

Kalitsis P, Saffery R (2009) Inherent promoter bidirectionality facilitates maintenance of sequence integrity and transcription of parasitic DNA in mammalian genomes. BMC Genom 10:498

Kaneko-Ishino T, Ishino F (2012) The role of genes domesticated from LTR retrotransposons and retroviruses in mammals. Front Microbiol 3:262

Kitazawa M, Tamura M, Kaneko-Ishino T, Ishino F (2017) Severe damage to the placental fetal capillary network causes mid- to late fetal lethality and reduction in placental size in Peg11/Rtl1 KO mice. Genes Cells 22:174–188

Kokošar J, Kordiš D (2013) Genesis and regulatory wiring of retroelement-derived domesticated genes: a phylogenomic perspective. Mol Biol Evol 30:1015–1031

Korb E, Finkbeiner S (2011) Arc in synaptic plasticity: from gene to behavior. Trends Neurosci 34:591–598

Kordiš D (2005) A genomic perspective on the chromodomain-containing retrotransposons: chromoviruses. Gene 347:161–173

Kordiš D (2009) Transposable elements in reptilian and avian (Sauropsida) genomes. Cytogenet Genome Res 127:94–111

Kordiš D (2011) Extensive intron gain in the ancestor of placental mammals. Biol Direct 6:59

Kordiš D, Kokošar J (2012) What can domesticated genes tell us about the intron gain in mammals? Int J Evol Biol 2012:27898

Kordiš D, Lovšin N, Gubenšek F (2006) Phylogenomic analysis of the L1 retrotransposons in Deuterostomia. Syst Biol 55:886–901

Krull M, Petrusma M, Makalowski W, Brosius J, Schmitz J (2007) Functional persistence of exonized mammalian-wide interspersed repeat elements (MIRs). Genome Res 17:1139–1145

Lavialle C, Cornelis G, Dupressoir A, Esnault C, Heidmann O, Vernochet C, Heidmann T (2013) Paleovirology of 'syncytins', retroviral env genes exapted for a role in placentation. Philos Trans R Soc Lond B Biol Sci 368:20120507

Llorens C, Marin I (2001) A mammalian gene evolved from the integrase domain of an LTR retrotransposon. Mol Biol Evol 18:1597–1600

Long M, Betrán E, Thornton K, Wang W (2003) The origin of new genes: glimpses from the young and old. Nat Rev Genet 4:865–875

Long M, VanKuren NW, Chen S, Vibranovski MD (2013) New gene evolution: little did we know. Annu Rev Genet 47:307–333

Lynch C, Tristem M (2003) A co-opted gypsy-type LTR-retrotransposon is conserved in the genomes of humans, sheep, mice, and rats. Curr Biol 13:1518–1523

Lynch M, Conery JS (2000) The evolutionary fate and consequences of duplicate genes. Science 290:1151–1155

Mateo L, González J (2014) Pogo-like transposases have been repeatedly domesticated into CENP-B-related proteins. Genome Biol Evol 6:2008–2016

Meredith RW, Janečka JE, Gatesy J, Ryder OA, Fisher CA, Teeling EC, Goodbla A, Eizirik E, Simão TL, Stadler T, Rabosky DL, Honeycutt RL, Flynn JJ, Ingram CM, Steiner C, Williams TL, Robinson TJ, Burk-Herrick A, Westerman M, Ayoub NA, Springer MS, Murphy WJ (2011) Impacts of the cretaceous terrestrial revolution and KPg extinction on mammal diversification. Science 334:521–524

Mi S, Lee X, Li X, Veldman GM, Finnerty H, Racie L, LaVallie E, Tang XY, Edouard P, Howes S, Keith JC Jr, McCoy JM (2000) Syncytin is a captive retroviral envelope protein involved in human placental morphogenesis. Nature 403:785–789

Miller WJ, McDonald JF, Nouaud D, Anxolabéhère D (1999) Molecular domestication—more than a sporadic episode in evolution. Genetica 107:197–207

Mita P, Boeke JD (2016) How retrotransposons shape genome regulation. Curr Opin Genet Dev 37:90–100

Naruse M, Ono R, Irie M, Nakamura K, Furuse T, Hino T, Oda K, Kashimura M, Yamada I, Wakana S, Yokoyama M, Ishino F, Kaneko-Ishino T (2014) Sirh7/Ldoc1 knockout mice exhibit placental P4 overproduction and delayed parturition. Development 141:4763–4771

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 long terminal repeat retrotransposons as a source of new genes in vertebrates. Clin Microbiol Infect 22:312–323

Oldham MC, Horvath S, Geschwind DH (2006) Conservation and evolution of gene coexpression networks in human and chimpanzee brains. Proc Natl Acad Sci USA 103:17973–17978

Schmitz J, Brosius J (2011) Exonization of transposed elements: a challenge and opportunity for evolution. Biochimie 93:1928–1934

Schüller M, Jenne D, Voltz R (2005) The human PNMA family: novel neuronal proteins implicated in paraneoplastic neurological disease. J Neuroimmunol 169:172–176

She X, Rohl CA, Castle JC, Kulkarni AV, Johnson JM, Chen R (2009) Definition, conservation and epigenetics of housekeeping and tissue-enriched genes. BMC Genom 10:269

Shiao MS, Khil P, Camerini-Otero RD, Shiroishi T, Moriwaki K, Yu HT, Long M (2007) Origins of new male germ-line functions from X-derived autosomal retrogenes in the mouse. Mol Biol Evol 24:2242–2253

Sinzelle L, Izsvák Z, Ivics Z (2009) Molecular domestication of transposable elements: from detrimental parasites to useful host genes. Cell Mol Life Sci 66:1073–1093

Sorek R (2007) The birth of new exons: mechanisms and evolutionary consequences. RNA 13:1603–1608

Stark A, Brennecke J, Bushati N, Russell RB, Cohen SM (2005) Animal microRNAs confer robustness to gene expression and have a significant impact on 3′UTR evolution. Cell 123:1133–1146

Takaji M, Komatsu Y, Watakabe A, Hashikawa T, Yamamori T (2009) Paraneoplastic antigen-like 5 gene (PNMA5) is preferentially expressed in the association areas in a primate specific manner. Cereb Cortex 19:2865–2879

Vinckenbosch N, Dupanloup I, Kaessmann H (2006) Evolutionary fate of retroposed gene copies in the human genome. Proc Natl Acad Sci USA 103:3220–3225

Volff JN (2006) Turning junk into gold: domestication of transposable elements and the creation of new genes in eukaryotes. BioEssays 28:913–922

Warren IA, Naville M, Chalopin D, Levin P, Berger CS, Galiana D, Volff JN (2015) Evolutionary impact of transposable elements on genomic diversity and lineage-specific innovation in vertebrates. Chromosome Res 23:505–531

Zdobnov EM, Campillos M, Harrington ED, Torrents D, Bork P (2005) Protein coding potential of retroviruses and other transposable elements in vertebrate genomes. Nucleic Acids Res 33:946–954

Part II
Species Evolution and Evolution of Complex Traits

Evolution of Complex Traits in Human Populations

Carolina Medina-Gomez, Oscar Lao and Fernando Rivadeneira

Abstract With the completion of the Human Genome Project and the increasing availability of dense panels with millions of genetic variants during the last decade, the identification of the fingerprint of classical positive selection events in the human genome is living a golden age of scrutiny. Nowadays, there is an ever-increasing number of methods developed to detect hard selective sweeps acting on de novo mutations within species. Despite the intrinsic problems underlying these methods (such as the lack of reproducibility, the influence of complex demographic history, and the presence of a large number of confounding factors), among different human populations, several genomic regions have been shown to be undergoing selective pressures. These discoveries are providing further understanding of the microevolution and microadaptation of our species. However, most phenotypes are of complex nature and arising from the frequently intricate demographic history of humankind. Therefore, most of the genetic variants that could have played an adaptive role in the past can be expected to have intermediate frequencies in the present. Under these circumstances, tests for detecting hard selective sweeps acting on variants determining complex human traits are underpowered, requiring new approaches for identifying positive polygenic adaptation. Here, we provide a succinct review of the current status of the field of evolutionary selection and describe the methodology we used to study the evolution of bone mineral density (BMD), and human stature as illustration of polygenic adaptation in humans.

C. Medina-Gomez · F. Rivadeneira
Erasmus MC, Rotterdam, The Netherlands

O. Lao (✉)
CNAG, Barcelona, Spain
e-mail: oscar.lao@cnag.crg.eu

P. Pontarotti (ed.), *Evolutionary Biology: Self/Nonself Evolution, Species and Complex Traits Evolution, Methods and Concepts*,
DOI 10.1007/978-3-319-61569-1_9

1 Positive Selection in the Human Genome

Natural selection is evidenced as individuals from a given generation differentially contribute to the genetic pool of the next generation depending on the genetic variants that they carry. The identification of the genetic signature of a positive selective event in a given locus (i.e., when an allele becomes more common in a population than what is expected under neutrality) is a recurrent topic of study in population genetics (Vitti, Grossman, and Sabeti 2013).

Neutrality tests can be classified according to different temporal and methodological criteria. In terms of the temporal scale, neutrality tests can address "macroevolution"—i.e., the genetic changes associated with speciation and adaptation of the species—or "microevolution"—i.e., the genetic changes associated with the recent differential adaptation of the populations (or local populations) of a species to their environment. Without neglecting the capital importance of macroevolution for the speciation process, the present chapter will focus on the evaluation of microevolution in humans. Microevolution tests can be classified according to the feature of the genetic variation used for the detection of the selection signal footprint: I) *Frequency-based tests* are designed to detect deviations from the expected allele frequency at a locus in a population; II) *Gene genealogy-based tests* analyze different aspects of the topology of the genealogy of the locus; III) *Haplotype (or linkage disequilibrium (LD))-based tests* assess the pattern of decay of LD around the locus of interest; IV) *Population differentiation-based tests* address the amount of genetic differentiation between local populations of a species; and V) *Composite tests* integrate information from multiple selection signals. A summary listing these tests can be found in Table 1. However, this strict classification is somewhat artificial as different authors classify the same tests into different categories (i.e., Vitti, Grossman, and Sabeti 2013; Pybus et al. 2015, 2013). Furthermore, meta-scores for detecting positive selection can also be generated by combining more than one of the proposed strategies for detecting selection (Vitti et al. 2013; Pybus et al. 2015; see Fig. 1).

The application of these tests has several limitations considering the underlying complexity that a selective process represents (see Sects. 1.2 and 1.3) and the inherent difficulty for distinguishing complex demographic patterns from selective events (see Sect. 1.1; Ramirez-Soriano et al. 2008). This derives: (*i*) a plethora of tests for detecting positive selection rather than "*one prime test*" that can embrace all different components; (*ii*) different tests withholding different power for rejecting the null hypothesis of neutrality (Akey 2009); and (*iii*) potential lack of concordance among the results of the different methods (Pybus et al. 2015).

1.1 Confounder Factors

Distinguishing the fingerprint of a selective event from the expected fingerprint occurring under neutrality (i.e., as a result of demographic factors) is difficult due to the complex demographic history of the human species (Fig. 2).

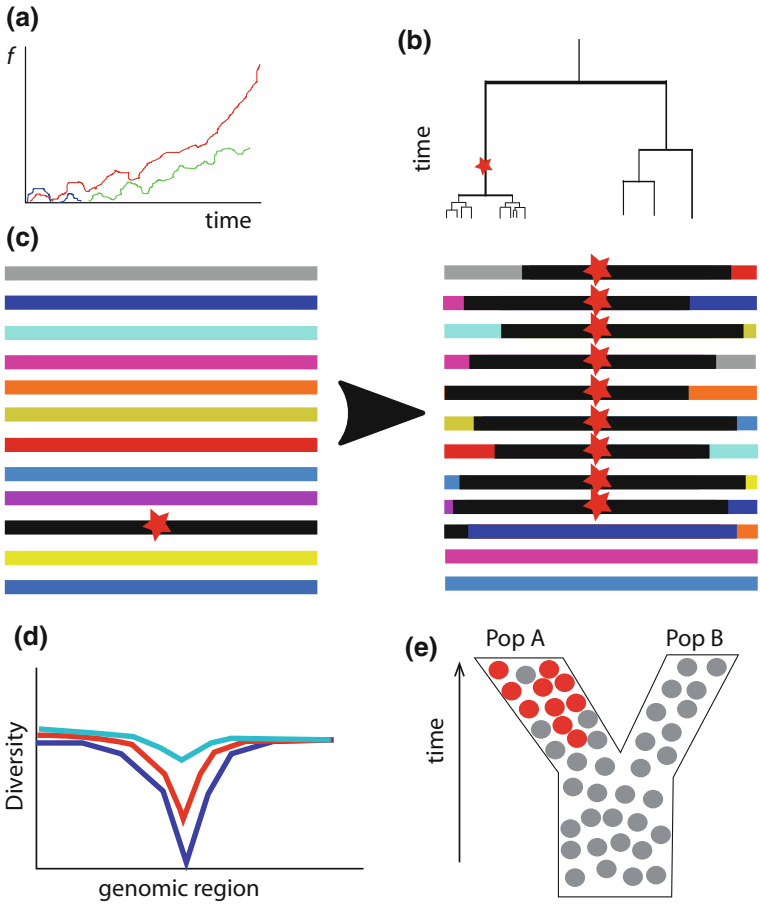
Table 1 List of statistical methods testing the null hypothesis of neutrality. Categories are broadly classified into the type of the approach that is favored by each method. Classification remains somewhat artificial, and some of the tests can be classified into more than one of the proposed categories depending on which parameter of the test is used. Adapted from (Pybus et al. 2015; Vitti et al. 2013)

Approach	Test	References
Frequency-based	*Tajima's D*	Tajima (1989)
	*Fu and Li's F**	Fu and Li (1993)
	*Fu and Li's D**	Fu and Li (1993)
	R2	Ramos-Onsins and Rozas (2002)
	Ewens–Watterson test	Ewens (1972)
	Fay & Wu's H	Fu and Li (1993), Fay and Wu (2000)
	SDS	Field et al. (2016)
Linkage disequilibrium	*Long-range haplotype (LRH) test*	Mode and Sleeman (2012)
	Long-range haplotype similarity test	Hanchard et al. (2006)
	Integrated haplotype score (iHS)	Voight et al. (2006)
	Cross-population extended haplotype homozygosity (XP-EHH)	Voight et al. (2006), Sabeti et al. (2007)
	Linkage disequilibrium decay (LDD)	Voight et al. (2006), Sabeti et al. (2007), Wang et al. (2006)
	HAF	Ronen et al. (2015b)
	nSL	Ferrer-Admetlla et al. (2014)
	Wall's B	Wall (1999)
	Wall's Q	Wall (2000)
	Fu's F	Fu (1997)
	Dh	Nei (1987)
	Za	Rozas et al. (2001)
	ZnS	Kelly (1997)
	ZZ	Rozas et al. (2001)
Population differentiation methods	*Lewontin–Krakauer test (LKT)*	Bonhomme et al. (2010)
	Locus-specific branch length (LSBL)	Shriver et al. (2004)
	Fst and related distances	Holsinger and Weir (2009)
	Qx	Berg and Coop (2014)
	hapFLK	Fariello et al. (2013)

(continued)

Table 1 (continued)

Approach	Test	References
	PCA	Duforet-Frebourg et al. (2015)
	DDAF (standard and absolute)	Hofer et al. (2009)
Composite methods: Each of these tests is further discussed in the main text	*XP-CLR*	Chen et al. (2010)
	CRL	Nielsen et al. (2005)
	CMS	Grossman et al. (2013)
	Hierarchical boosting	Pybus et al. (2015)
	SFselect	Ronen et al. (2013)



◂**Fig. 1** The signature of a microevolutionary hard positive ongoing event can be observed in the genome at different levels: **a** *Frequency increase in the population with time*: Red line represents the fate over time of a new mutation appearing in a given population conferring selective advantage to the carriers over the non-carriers, systematically increasing in frequency until fixation is reached. Blue line represents two neutral mutations appearing and disappearing by chance (genetic drift) from the population. Green line represents a neutral mutation appearing in the population and increasing frequency by chance. **b** *Shaped by the ancestral tree*: a selective event appearing along a branch of the ancestral tree that has not reached fixation. Each branch represents a copy of a genomic fragment. The ancestral tree models the coalescence events between these copies. In one of the branches of the tree, a mutation (*red star*) occurs in the genomic fragment that is under positive selection. Because of that event, the topology of the ancestral tree changes. All the eight copies carrying the mutation share a more recent common ancestor compared to the three copies that do not carry the mutation under selection (*right branch*). **c** *Shaped by the Linkage disequilibrium/haplotype structure*: in a chromosome with a particular diverse genomic background (exemplified by different colors), a mutation (*red star*) conferring selective advantage occurs (*left*). After several generations, the mutation frequency increases in the population (*right*). However, because of such rapid increase, recombination has had not enough time to break the association of the mutation under selection with the genetic background where it appeared, so all copies carrying the mutation will also share the same genetic background (*black*) around the mutation. **d** *Reduced genetic diversity*: In such region, genetic diversity around the mutation will be reduced as compared to the diversity expected across the genome. Depending on the strength of the selective event and the time since it occurred, the decay in diversity with regard to that observed on average across the genome can be more (*dark blue line*) or less (*light blue line*) evident. **e** *Population differentiation after a simple population split*: when a selective ongoing event is present after a simple two-population split. Adapted from (Vitti et al. 2013)

Nowadays, the most accepted theory of human evolution indicates that anatomically modern humans appeared in Africa ~200 thousand years ago and spread around the world during the *Out-of-Africa* diaspora (Nielsen et al. 2017), resulting in a severe demographic bottleneck emerging out of the African continent.

Whether this diaspora was unique or recurrent is still under debate. Based on archeological and genetic data, some authors claim that some populations, like the Papuans, carry in their genome signatures of anatomically modern humans arising from an ancient and largely extinct *Out-of-Africa* migration (Tassi et al. 2015; Pagani et al. 2016). Alternatively, other authors claim that most of the identified genetic discrepancies can rather be explained by differential admixture between anatomically modern humans and archaic hominins, such as Neanderthals and Denisovans, who were already present in Europe and Asia at the time of their arrival (Malaspinas et al. 2016). In fact, such anatomically modern humans encountered outside of the African continent possess some degree of Neanderthal ancestry, implying that any Neanderthal admixture occurred after the *Out-of-Africa* event (Vattathil and Akey 2015; Simonti et al. 2016; Sankararaman et al. 2014, 2016; Qin and Stoneking 2015). Similar degrees of Denisovan ancestry have also been traced in modern humans encountered in Oceania.

The effect of this archaic introgression in our genome is just starting to be elucidated. As Neanderthals and Denisovans had survived for thousands of years in Eurasia before the *Out-of-Africa* event, they were better adapted anatomically to the local conditions than the newcomer modern humans. Therefore, interbreeding of

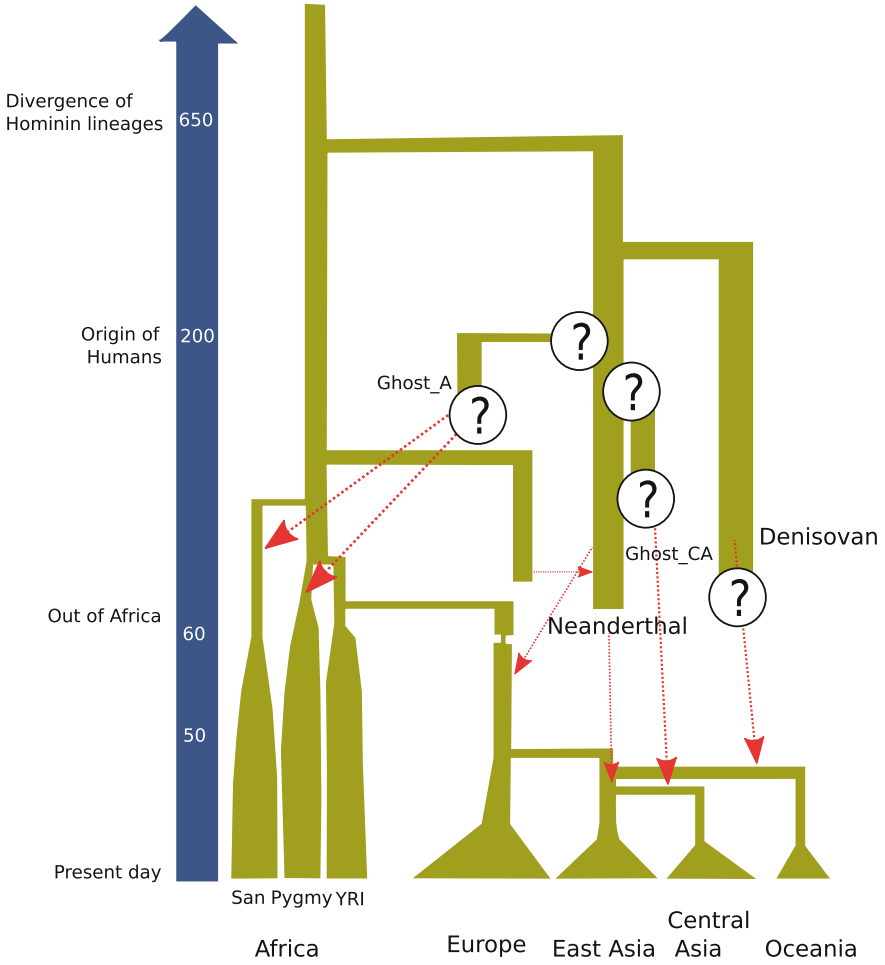


Fig. 2 Demographic history of the human species (adapted from Malaspinas et al. 2016; Nielsen et al. 2017; Mondal et al. 2016; Hsieh et al. 2016). The current model of human evolution proposes the existence of admixture with known archaic species such as Neanderthal or Denisovan, as well as with other unknown "ghost" archaic species

anatomically modern humans with archaic human species could have facilitated their adaptation to particular environments. An enrichment of archaic ancestry compatible with positive selection in some particular loci has also been demonstrated (Sankararaman et al. 2014; Vattathil and Akey 2015). Many complex traits have been proposed as targets for adaptive introgression, such as skin pigmentation, defense against pathogens (Racimo et al. 2015) or adaptation to high altitude (i.e., *EPAS1* gene haplotype) (Racimo et al. 2015). However, as a general rule, the fingerprint of archaic introgression is depressed in functional elements, conforming to a general purifying selection acting on the hybrid genome (Vattathil and Akey

2015; Sankararaman et al. 2014). In line with this contention, archaic introgression has been associated with increased risk for common multifactorial diseases like obesity, depression, and skin lesions by sun exposure (Simonti et al. 2016).

1.2 Hard Selective Sweeps

Many of the proposed methods for the detection of selection signals assume the presence of a hard or classic selective sweep, in which a de novo mutation rapidly increases in frequency in the population due to a selective advantage of the carriers of the mutation, ultimately reaching fixation (Wollstein and Stephan 2015; Fig. 1). Yet, selective sweeps in the human genome are rare and that other types of selective processes, such as background or purifying selection, can mimic the fingerprint of a selective sweep (Hernandez et al. 2011). Nevertheless, several genomic regions harboring functional variants have been shown to be under positive selection by different methods.

A classic example in humans of such strong selective sweep is lactose persistence in adulthood emerging in some pastoralist/herder human populations. Such adaptation results from selected polymorphisms in the *LCT* gene (Tishkoff et al. 2007). The fingerprint of selection at the *LCT* gene is so strong that its detection has become the gold standard for any new proposed statistic (Sabeti et al. 2002). Other examples of hard selective sweeps include the selection of the polymorphism V370A at the *EDAR* gene in East Asian populations (Sabeti et al. 2007; Grossman et al. 2010), implicated in the development of sweat glands and skin appendages (Lu and Fuchs 2014), or the fixation of polymorphisms at *SLC24A5* in European populations related to skin pigmentation (Norton et al. 2007). In mice, the *EDAR*-V370A polymorphism presents with particular tooth morphology, hair thickness, and increased sweat gland density (Kamberov et al. 2013); in humans, this polymorphism associates with hair straightness (Wu et al. 2016). Mutations in *SLC24A5* confer an albino phenotype in humans (Wei et al. 2013) as well as in zebra fish (Lamason et al. 2005).

Despite that the evidence arising from these genomic regions (Sabeti et al. 2007; Grossman et al. 2010) supports the hard selective sweep model, some investigators suggest alternative explanations. It has been claimed that given the recent origin and the relatively small historic effective population size of humankind, there has been little time for a new beneficial mutation to occur (Lewin and Foley 2004). Thus, it is more realistic to assume that selective events occur at genetic variants that were already present in the population (standing variation). Furthermore, temporal environmental variation in selective pressures (such as the alternation between ice and temperate ages or the transition to new diets) has been proposed as an important factor shaping functional genetic variants in humans (Gillespie 1994). In such variable scenarios, adaptive response to environmental shifts could be occurring faster on standing variants (Hancock and Rienzo 2008). In this, so-called *soft selective sweep* scenario, tests for detecting hard selective sweeps have low power

for rejecting the null hypothesis of neutrality (Hancock and Rienzo 2008). A general critique to the assumptions underlying the soft selective sweep/polygenic adaptation model approach is described elsewhere (Jensen 2014).

1.3 Soft Selective Sweeps and Polygenic Adaptation

The great majority of phenotypic traits are polygenic by nature, meaning they are controlled by many genes, each of which contributes small amounts to the overall phenotype variation (McEvoy et al. 2006; Chen et al. 2012; Turchin et al. 2012; Corona et al. 2013; Estrada et al. 2012; Ahn et al. 2010; Perry et al. 2013). In contrast to the pattern observed in strong directional selective sweeps, the departure from neutrality in polygenic traits is usually not confined to a given locus; rather, it embodies a range of soft selective sweeps, i.e., small allele frequency shifts that spread across many loci at the same time (Messer and Petrov 2013). This occurs either because the alleles were already present as standing genetic variation or because they arose independently by recurrent de novo mutations (see Fig. 3).

Identifying signatures of polygenic adaptation remains challenging, and to date, only few methods have been proposed for testing systematically departure of neutrality in polygenic traits (Berg and Coop 2014; Ferrer-Admetlla et al. 2014; Ronen et al. 2015; Duforet-Frebourg et al. 2016; Field et al. 2016). A fundamental difference of the tests for detecting polygenic adaptation in contrast to hard selective sweep approaches is the need of prior knowledge about the functional background in the surveyed loci. This is reflected in the use of functional pathway analyses (i.e. Engelken et al. 2016) (rather than focusing on single genes or loci) or the use of phenotype-specific information associated with the loci in the tests of polygenic adaptation. During the last decade, genome-wide association studies (GWAS) have provided a valuable source of identified genetic variants (mostly single nucleotide polymorphisms or so-called SNPs) shown to be associated with thousands of different complex phenotypes (Welter et al. 2014). GWAS perform a hypothesis-free screen of the genome to test hundreds of thousands to several million of SNPs for association with a trait of interest. Although GWAS are unbiased with respect to former biological knowledge, the cost for this freedom is the stringent multiple-testing correction needed to be applied during the analysis. In order to reach the traditionally accepted genome-wide significance threshold (i.e., p-value 5×10^{-8}), meta-analysis of studies is performed to achieve the large sample sizes needed to detect association. Hence, genetic variants identified by means of GWAS tend to be common and associated with the phenotype through linkage disequilibrium, rather than being causal for the phenotype. With the exception of quasi-Mendelian phenotypic traits such as eye color (Wollstein et al. 2017), GWAS variants explain a very small proportion of the phenotypic variation (Wray et al. 2013). Despite these limitations, GWAS discoveries have provided a big leap in the understanding of human biology and now constitute a fertile ground for analyzing the patterns of selection underlying complex traits and diseases.

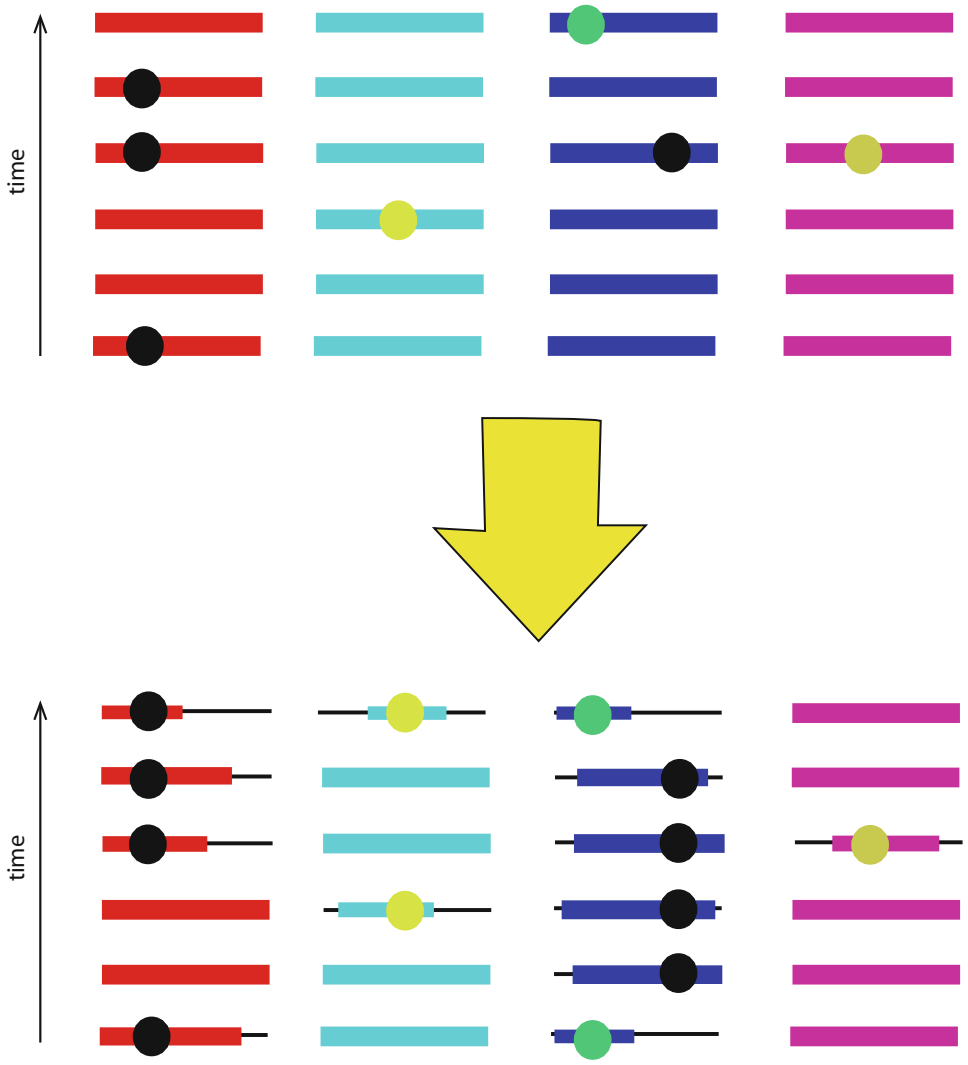


Fig. 3 A polygenic model of adaptation (modified from Pritchard et al. 2010). Each dot represents a de novo mutation and each line a chromosome. All mutations contribute to a complex phenotype, which is under selective pressures. After some generations, few mutations show the pattern expected under a strong selective sweep (i.e., almost reaching fixation or showing long-range haplotypes such as at the *dark blue* chromosomes). All the other loci show patterns that can be expected under neutrality

2 Acting Selection on Complex Traits

There is increasing evidence that the spatial distribution of complex traits and multifactorial diseases such as skin pigmentation (McEvoy et al. 2006), type 2 diabetes (Chen et al. 2012), biliary liver cirrhosis, or ulcerative colitis (Corona et al. 2013) have been shaped by soft selective pressures. Nevertheless, in most of the cases, the evidence is scarce and subject to the conclusions derived from the preferred neutrality test used for the analysis. In this review, we focus on two skeletal traits holding evidence to stand as examples of selection acting on human traits.

2.1 Genetics of Human Stature (Body Height)

Body height is a highly heritable ($h^2 \sim 80$–90%) polygenic trait widely recognized as being shaped by selective pressures. To date, approximately 700 common variants have been identified through GWAS as associated with height ($n \sim 250{,}000$), the great majority of them exerting a modest effect on the phenotype (Wood et al. 2014). Moreover, recently rare variants with large phenotypic effect contributing to height variation in the general population have been identified in a larger meta-analysis ($n \sim 700{,}000$) scrutinizing exonic variation (Marouli et al. 2017). There are several lines of evidence showing that human height is under differential selective pressures among populations within Europe (between northern and southern European populations) and between European and other worldwide populations. First, the frequencies of alleles associated with increased height are systematically elevated in northern Europeans as compared with southern Europeans, and this distribution cannot be explained by genetic drift (Turchin et al. 2012). Second, by using a weighted genetic score (GS, Dudbridge 2013) computed for each individual using the estimated effect size at each SNP:

$$GS_i = \sum_{snp}^{N} b_{snp} G_{snp,i}$$

where G is the genotype and b is the effect size of the SNP for the phenotype, the genetically determined height has been assessed. This score show differences across European populations, and these differences cannot be explained by neutral evolution (Robinson et al. 2015). Similarly, Berg and Coop found differences in the height-GS across populations worldwide (Berg and Coop 2014). Third, the analysis of the alleles associated with height in ancient samples from Europe at different time periods shows a temporal trend incompatible with a neutral model shaping this trait (Mathieson et al. 2015). Interestingly, the authors of this last study identified two independent geographic signals of height adaptation: one for decreased (genetic) height in Neolithic Iberia and one for increased (genetic) height in Bronze Age steppe populations. Finally, a recently developed test for detecting extremely

recent (<2000 years) signals of positive selection (Field et al. 2016) showed that the trend toward increased height in the UK was under selective pressures. Notably, despite these pieces of evidence in favor of height being under selective pressures, the underlying factor(s) shaping the genetic variation associated with height remain largerly unknown. The main reason for such uncertainty is that many genetic variants associated with height are pleiotropic and also play a role on other complex traits (Pickrell et al. 2015) or disease phenotypes that in turn can also be under differential pressures. Yet, assortative mating has been recently postulated as another important driving force for explaining the genomic architecture of height (Robinson et al. 2017). Body mass index (BMI), the ratio of the body mass by the square of the body height, is another phenotype that is recurrently suggested as a candidate for positive and sexual selection in humans (Wells 2010).

2.2 *Genetics of Bone Mineral Density*

In this section, we turn our attention to additional aspects of the human skeleton. The musculoskeletal system as a whole has been shaped in humans by evolutionary forces of natural selection. One obvious example is the bipedalism of humans, where the ability to walk and run upright on two feet was made possible by virtue of specialized adaptations of the skeleton and muscles (Wu and Zhang 2010). Also, studies comparing the limb bones of different species show that bone material can change through evolution to help meet the diverse demands that animals place on their skeletons (Blob et al. 2014). Our bones are made up from materials and structural properties that meet the opposing needs of strength and lightness, stiffness, and flexibility (Seeman 2002). Archeological and paleontological records clearly illustrate how the human skeleton has changed over the past 2 million years. Further, unraveling this phenomenon could shed some light on the understanding of osteoporosis, a disease characterized by reduced bone mass and disrupted bone architecture resulting in increased fracture risk (Fonseca et al. 2014).

Bones play an important role in the overall function of the human body. Besides being a highly specialized supporting framework of the body, the skeleton has many other functions. Bones protect vital organs, provide an environment for marrow (both for blood forming and fat storage), act as a mineral reservoir for calcium homeostasis, and also serve as a storehouse for growth factors and cytokines (Kini and Nandeesh 2012). Bone mineral density (BMD) expressed as bone mineral content (bone mass) per area unit is a proxy of bone strength used in clinical practice to assess bone health and risk of fracture (particularly in the elderly). Beyond BMD, other bone properties (not captured by the BMD measurement) influence bone strength independently from BMD, but are out of the scope of this chapter. Large studies aiming to study genetic factors influencing osteoporosis have mainly focused on BMD as it constitutes a precise and widely available trait. BMD is under strong genetic influence, and both family-based and

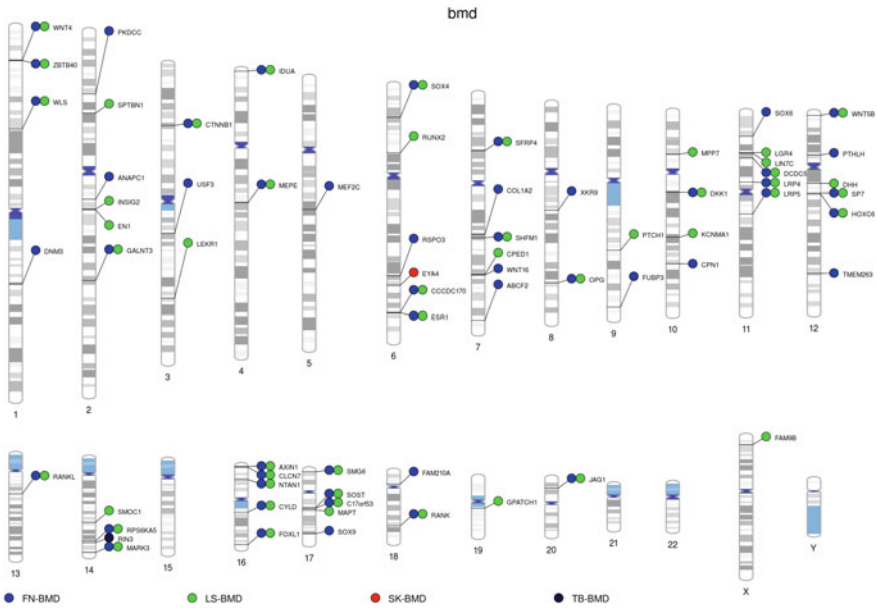


Fig. 4 Summary of BMD loci identified by GWAS to date. Each locus is labeled according to the candidate gene reported in the corresponding publication. Different colors identify the skeletal sites for the association with BMD: femoral neck (FN-BMD), lumbar spine (LS-BMD), skull BMD (SK-BMD), and total body (TB-BMD). *Note,* the latter two phenotypes have been evaluated only in children, loci are included only if associations were specific to pediatric populations

population-based studies have shown that it is a highly heritable trait, with heritability estimates ranging from 50 to 85% (Kemp et al. 2016).

Presently, the precise makeup of the genetic architecture underpinning BMD in the general population is not fully understood, although it is widely thought to be the product of hundreds if not thousands of genetic variants, each of small effect size and potentially interacting with environmental exposures (e.g., physical activity, nutrition, and lifestyle). The advent of genome-wide association studies (GWAS) has resulted in the robust identification of more than 70 loci associated with BMD (Styrkarsdottir et al. 2016a, b, 2013; Zheng et al. 2015; Rivadeneira et al. 2009; Medina-Gomez et al. 2012; Richards et al. 2008; Kemp et al. 2014; Estrada et al. 2012; Koller et al. 2013; Zheng et al. 2012; Duncan et al. 2011; Zhang et al. 2014) (Fig. 4). Most of the variants described so far are common in the population (MAF > 0.05), although recent studies have shown rare variants in *EN1*, *COL1A2*, *CPED1*, *PTCH1*, and *LGR4* playing an important role in the BMD variability (Zheng et al. 2015; Styrkarsdottir et al. 2016a, b). It is worth noting that, as for other traits, the majority of GWAS for BMD have investigated European populations and thus, it is possible that variants segregating at low frequency in European populations or private to other ethnic groups have not yet been properly surveyed in relation to their association with this trait.

BMD GWAS have traditionally focused on skeletal sites of high fracture risk (i.e., femoral neck, lumbar spine, and forearm), and the gathered results indicate that many genetic variants exert site-specific effects (Estrada et al. 2012; Kemp et al. 2014). The exact reason underlying these site differences is not fully elucidated. Nevertheless, it has been proposed that they could reflect the differential exposure of each skeletal site to varying environmental stimuli or be the consequence of differences in the type of bone at the site of measurement. For instance, the lumbar spine site comprises mostly trabecular bone (a network of thin interconnecting plates in vertebrae and long bones), while the femoral neck constitutes mostly cortical (compact) bone (Kemp et al. 2016). Therefore, it is plausible to think that genetic variants influencing different skeletal sites experienced different selective pressures.

Many of the genes discovered by GWAS have been shown to play an important role in bone metabolism as confirmed by knockout animal models and thus shed a light on the basis of bone biology. However, the discovered variants by GWAS currently explain less than 10% of BMD variability (Estrada et al. 2012), in other words, there should be a large number of genetic variants that we have not discovered yet, which influence BMD and as such contribute explaining the BMD variability. With the costs of genotyping steadily decreasing and the availability of larger reference panels comprising deeper sequenced samples from diverse populations, it is expected that an increasing number of genetic variants will be discovered, helping to characterize further the genetic architecture of BMD variation.

3 Evolution of Bone Strength in Humans

Bone strength is determined by bone material properties, bone mass, and bone structural geometry. Bone material properties remain fairly constant during an individual's life course. In contrast, both bone mass and its distribution (geometry) adapt to the mechanical forces applied on bone. As a consequence, there is high variation on bone strength through the life course of one individual and across individuals as result of changes in both bone mass and bone architecture. In general, age-related bone deterioration ultimately leads to increased fracture risk.

From a broader perspective, ethnic differences in BMD and fracture risk are well documented (Shin et al. 2014; Horlick et al. 2000; Bulathsinhala et al. 2017; Curtis et al. 2016). As a general rule, individuals of African ancestry possess higher BMD than individuals from European and Asian ancestry, and differences between these two latter groups are rather small (Finkelstein et al. 2002, Marshall et al. 2008). Similarly, fracture risk is lower in individuals of African ancestry (Barrett-Connor et al. 2005, Putman et al. 2017). Even though environmental factors (like nutrition, physical activity to name a few) definitively play a role, the genetic background underlying ancestry is thought to be the main explanation of the observed differences. How these ethnic differences arose is still unknown, although it is plausible that such diversity has emerged as different populations worldwide faced different demographic and environmental challenges once they left Africa.

The integration of GWAS-derived information in the study of these ethnic differences in BMD has shown that indeed a substantial fraction of these differences is of genetic origin. Recently, we investigated the role of the common variants in BMD accrual. For doing so, we generated a genetic score of 63 common variants robustly associated with BMD (BMD-GS; Estrada et al. 2012) in individuals from two independent multiethnic cohorts; and demonstrated that these alleles associated with higher BMD were systematically more common in individuals with sub-Saharan African ancestry as compared to individuals with other ethnic background (Medina-Gómez et al. 2015). The same pattern was observed at a phenotypic level, where individuals from sub-Saharan African ancestry showed higher BMD. Furthermore, when extrapolating the BMD-GS approach to worldwide populations (using publicly available data from the Human Genome Diversity Project panel (Li et al. 2008), we reached the same conclusion: sub-Saharan African populations showed the higher BMD-GS scores. Even more remarkable, the differentiation patterns of these markers worldwide suggested that BMD heterogeneity is the result of differential selective pressures in non-African populations.

We describe in detail here the procedure applied to gather evidence resulting from the combination of different tests for detecting hard selective sweeps with the assessments arising from tests to assess polygenic adaptation. First, as mentioned in the previous paragraph, the geographic pattern of the BMD-GS paralleled the geographic disparities observed in BMD worldwide (i.e., the BMD-GS of individuals from sub-Saharan African populations was in average higher than those from individual surveyed elsewhere) (Fig. 5). Demographic factors such as the *Out-of-Africa* event could, in principle, explain such differentiation. Nonetheless, the differentiation of the BMD-GS between sub-Saharan African and non-sub-Saharan African

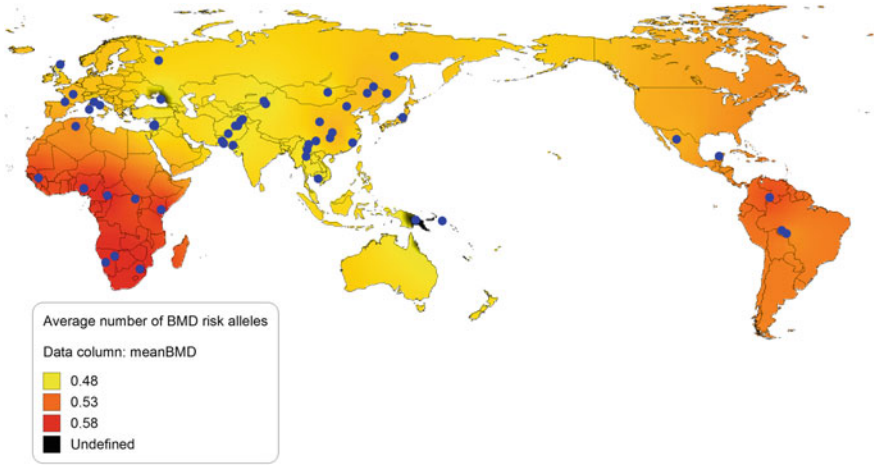


Fig. 5 Density map of the average number of BMD risk alleles (unweighted GS) using the HGDP-CEPH (Li et al. 2008) worldwide populations. The average value at each geographic location is obtained by means of inverse to power geostatistical interpolation (Isaaks and Mohan Srivastava 1989). Adapted from (Medina-Gómez et al. 2015)

populations was higher as compared with the differentiation of random sets of equal number of SNPs across the genome (sampled scores). Similar results were obtained even after matching SNPs in the sampled scores for factors related to biases of the GWAS discovery setting such as presence of background selection or the lower limit of allele frequency in European population (where variants were identified). Our results are in line with a previous study in which the authors analyzed population differentiation of candidate genes (e.g., *LRP4*, *ACAN*, *MSX2*) that had previously been proposed as influencing the skeletal system in humans (Wu and Zhang 2010). Second, we found that in 73% of the associated SNPs, the BMD-increasing allele was the ancestral allele, implying that phenotypic states related to increased BMD, such as bone robustness (Nowlan et al. 2011), represent the ancestral state in humans. It is worth noting that the examination of the skeleton of modern humans and chimpanzees, and fossils of extinct human lineages spanning several million years also disclose the relatively lightly built skeletons of modern humans (Chirchir et al. 2015). Moreover, it has been shown that low-density bones only evolved relatively recently in modern humans (Ryan and Shaw 2015), likely as a result of the shift from foraging to agriculture-based way of living. Therefore, it is tempting to hypothesize that the skeleton of modern humans has adapted to new environments, and that ancient migrations might have also played a significant role in shaping the ethnic differences observed across the skeletal characteristics of modern humans. Third, we observed that tests for detecting hard selective sweeps—under the neutrality hypothesis of no selection—applied to genomic regions associated with BMD showed a depletion of statistically significant results (at p-value < 0.05). Noteworthy, studies analyzing other complex phenotypes such as height (Mondal et al. 2016) or mental disorders (Mondal et al. 2016; Polimanti and Gelernter 2017) have indeed identified an enrichment of hard selective sweeps at their associated loci. However, an enrichment of tests supporting neutrality could be explained if tests for detecting hard selective sweeps show reduced power for detecting polygenic adaptation, as expected for BMD given all the other results. Overall, this result suggests that an enrichment of tests rejecting the null hypothesis of neutrality by hard sweep-based tests is not a *sine qua non* condition for identifying polygenic adaptation.

Altogether, the study of BMD from an evolutionary perspective sheds new light on characterizing the known, but not completely understood, ethnic differences in bone strength across populations and represents a leap in the understanding of how human polygenic adaptation can shape a complex trait through evolution.

4 Future Perspectives and Concluding Remarks

The different investigations summarized in this chapter provide *state of the art* in the field of evolution of complex traits. As larger GWAS continue to emerge, the statistical power to detect association signals will allow performing more comprehensive surveys of selection genome-wide. The dissection of the genetic architecture of adaptive traits is crucial to understand evolutionary processes and

infer causes and effects of past processes. Therefore, coupled with advances in bioinformatics, the use of GWAS is now better equipped to face the challenge of figuring out how diversity arises and is maintained by natural selection. Numerous evolutionary hypotheses have been proposed to account for the observation of specific complex phenotypes, in humans and other species. In this chapter, we have provided examples from two complex skeletal traits in humans, namely body height and BMD.

Genomic surveys in humans have identified a large amount of regions displaying recent positive selection. With regard to the human musculoskeletal system, recent investigations point to a crucial role of the increase in sedentary lifestyle on the gracilization of the modern human skeleton. Examining variation of selection signatures in loci exerting differential effects across skeletal sites, as well as the evaluation across specific bone compartments (i.e., trabecular vs cortical bone), would help to understand better the role of differential biomechanical loading (e.g., as a consequence of reduced physical activity) in the evolution of weaker bones in modern humans.

In addition, as modern humans dispersed from Africa throughout the world, they encountered and interbred with archaic hominins (see Sect. 1.1). The role of archaic introgression on BMD variability in populations out of Africa is not yet known. Nevertheless, morphological studies described a large number of skeletal differences between Neanderthal and anatomically modern humans (Sawyer and Maley 2005). Therefore, it is possible that these differences could be responsible for a fraction of the observed BMD differences between and among modern human populations based on the degree of archaic introgression among individuals.

While the evidence discussed in this chapter has been confined to genomics (as characterized by GWAS data), other *–omics* sources like epigenomics and transcriptomics will provide extra layers of information, particularly about the regulatory landscape affecting phenotypic variability alone or in interaction with environmental challenges. For instance, DNA methylation differences exist between major ethnic groups (Heyn and Esteller 2012) and the study of the epigenomic landscape of rainforest hunter-gatherers and sedentary farmers show that recent and historical changes in habitat and lifestyle have both critical impacts on DNA methylation variation (Fagny et al. 2015). Therefore, integration of additional *–omics* layers is warranted for the study and understating of how evolution has shaped modern complex traits and the occurrence of multifactorial diseases.

References

Ahn J, Kai Yu, Rachael Stolzenberg-Solomon K, Simon C, McCullough ML, Gallicchio L, Jacobs EJ et al (2010) Genome-wide association study of circulating vitamin D levels. Hum Mol Genet 19(13):2739–2745

Akey JM (2009) Constructing genomic maps of positive selection in humans: where do we go from here? Genome Res 19(5):711–722

Barrett-Connor E, Siris ES, Wehren LE, Miller PD, Abbott TA, Berger ML, Santora AC, Sherwood LM (2005) Osteoporosis and fracture risk in women of different ethnic groups. J Bone Miner Res 20(2):185–194

Berg JJ, Coop G (2014) A population genetic signal of polygenic adaptation. PLoS Genet 10(8): e1004412

Blob RW, Espinoza NR, Butcher MT, Lee AH, D'Amico AR, Baig F, Megan Sheffield K (2014) diversity of limb-bone safety factors for locomotion in terrestrial vertebrates: evolution and mixed chains. Integr Comp Biol 54(6):1058–1071

Bonhomme M, Chevalet C, Servin B, Boitard S, Abdallah JM, Blott S, San Cristobal M (2010) Detecting selection in population trees: the Lewontin and Krakauer test extended. Genetics. doi:10.1534/genetics.110.117275

Bulathsinhala L, Hughes JM, McKinnon CJ, Kardouni JR, Guerriere KI, Popp KL, Matheny Jr RW, Bouxsein ML (2017) Risk of stress fracture varies by race/ethnic origin in a cohort study of 1.3 Million U.S. army soldiers. J. Bone Mineral Res: Off J Am Soc Bone Mineral Res. doi:10.1002/jbmr.3131

Chen H, Patterson N, Reich D (2010) Population differentiation as a test for selective sweeps. Genome Res 20(3):393–402

Chen R, Corona E, Sikora M, Dudley JT, Morgan AA, Moreno-Estrada A, Nilsen GB et al (2012) Type 2 diabetes risk alleles demonstrate extreme directional differentiation among human populations, compared to other diseases. PLoS Genet 8(4):e1002621

Chirchir H, Kivell TL, Ruff CB, Hublin J-J, Carlson KJ, Zipfel B, Richmond BG (2015) Recent origin of low trabecular bone density in modern humans. Proc Natl Acad Sci USA 112(2): 366–371

Corona E, Chen R, Sikora M, Morgan AA, Patel CJ, Ramesh A, Bustamante CD, Butte AJ (2013) Analysis of the genetic basis of disease in the context of worldwide human relationships and migration. PLoS Genet 9(5):e1003447

Curtis EM, van der Velde R, Moon RJ, van den Bergh JPW, Geusens P, de Vries F, van Staa TP, Cooper C, Harvey NC (2016) Epidemiology of fractures in the United Kingdom 1988–2012: variation with age, sex, geography, ethnicity and socioeconomic status. Bone 87(June):19–26

Dudbridge F (2013) Power and predictive accuracy of polygenic risk scores. PLoS Genet 9(3): e1003348

Duforet-Frebourg N, Luu K, Laval G, Bazin E, Blum Michael G B (2015) Detecting genomic signatures of natural selection with principal component analysis: application to the 1000 genomes data. Mol Biol Evol 33(4):1082–1093

Duforet-Frebourg N, Luu K, Laval G, Bazin E, Blum Michael G B (2016) Detecting genomic signatures of natural selection with principal component analysis: application to the 1000 genomes data. Mol Biol Evol 33(4):1082–1093

Duncan EL, Danoy P, Kemp JP, Leo PJ, McCloskey E, Nicholson GC, Eastell R et al (2011) Genome-wide association study using extreme truncate selection identifies novel genes affecting bone mineral density and fracture risk. PLoS Genet 7(4):e1001372

Engelken J, Espadas G, Mancuso FM, Bonet N, Scherr A-L, Jímenez-Álvarez V, Codina-Solà M et al (2016) Signatures of evolutionary adaptation in quantitative trait loci influencing trace element homeostasis in liver. Mol Biol Evol 33(3):738–754

Estrada K, Styrkarsdottir U, Evangelou E, Hsu Y-H, Duncan EL, Ntzani EE, Oei L et al (2012) Genome-wide meta-analysis identifies 56 bone mineral density loci and reveals 14 loci associated with risk of fracture. Nat Genet 44(5):491–501

Ewens WJ (1972) The sampling theory of selectively neutral alleles. Theor Popul Biol 3(1): 87–112

Fagny M, Patin E, MacIsaac JL, Rotival M, Flutre T, Jones MJ, Siddle KJ et al (2015) The epigenomic landscape of African rainforest hunter-gatherers and farmers. Nat Commun 6:10047

Fariello MI, Boitard S, Naya H, SanCristobal M, Servin B (2013) Detecting signatures of selection through haplotype differentiation among hierarchically structured populations. Genetics 193 (3):929–941

Fay JC, Wu CI (2000) Hitchhiking under positive Darwinian selection. Genetics 155(3): 1405–1413

Ferrer-Admetlla A, Liang M, Korneliussen T, Nielsen R (2014) On detecting incomplete soft or hard selective sweeps using haplotype structure. Mol Biol Evol 31(5):1275–1291

Field Y, Boyle EA, Telis N, Gao Z, Gaulton KJ, Golan D, Yengo L et al (2016) Detection of human adaptation during the past 2,000 years. doi:10.1101/052084

Finkelstein JS, Lee ML, Sowers M, Ettinger B, Neer RM, Kelsey JL, Cauley JA, Huang MH, Greendale GA (2002) Ethnic variation in bone density in premenopausal and early perimenopausal women: effects of anthropometric and lifestyle factors. J Clin Endocrinol Metab 87(7):3057–3067

Fonseca H, Moreira-Gonçalves D, Coriolano H-JA, Duarte JA (2014) Bone quality: the determinants of bone strength and fragility. Sports Med 44(1):37–53

Fu YX (1997) Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. Genetics 147(2):915–925

Fu YX, Li WH (1993) Statistical tests of neutrality of mutations. Genetics 133(3):693–709

Gillespie JH (1994) The causes of molecular evolution. Oxford University Press

Grossman SR, Andersen KG, Shlyakhter I, Tabrizi S, Winnicki S, Yen A, Park DJ et al (2013) Identifying recent adaptations in large-scale genomic data. Cell 152(4):703–713

Grossman SR, Shlyakhter I, Shylakhter I, Karlsson EK, Byrne EH, Morales S, Frieden G et al (2010) A composite of multiple signals distinguishes causal variants in regions of positive selection. Science 327(5967):883–886

Hanchard NA, Rockett KA, Spencer C, Coop G, Pinder M, Jallow M, Kimber M, McVean G, Mott R, Kwiatkowski DP (2006) Screening for recently selected alleles by analysis of human haplotype similarity. Am J Hum Genet 78(1):153–159

Hancock AM, Di Rienzo A (2008) Detecting the genetic signature of natural selection in human populations: models, methods, and data. Annu Rev Anthropol 37:197–217

Hernandez RD, Kelley JL, Elyashiv E, Cord Melton S, Auton A, McVean G, 1000 Genomes Project, Sella G, Przeworski M (2011) Classic selective sweeps were rare in recent human evolution. Science 331(6019):920–924

Heyn H, Esteller M (2012) DNA methylation profiling in the clinic: applications and challenges. Nat Rev Genet 13(10):679–692

Hofer T, Ray N, Wegmann D, Excoffier L (2009) Large allele frequency differences between human continental groups are more likely to have occurred by drift during range expansions than by selection. Ann Hum Genet 73(1):95–108

Holsinger KE, Weir BS (2009) Genetics in geographically structured populations: defining, estimating and interpreting FST. Nat Rev Genet 10(9):639–650

Horlick M, Thornton J, Wang J, Levine LS, Fedun B, Pierson RN Jr (2000) Bone mineral in prepubertal children: gender and ethnicity. J Bone Mineral Res: Off J Am Soc Bone Mineral Res 15(7):1393–1397

Hsieh P, Woerner AE, Wall JD, Lachance J, Tishkoff SA, Gutenkunst RN, Hammer MF (2016) Model-based analyses of whole-genome data reveal a complex evolutionary history involving archaic introgression in Central African Pygmies. Genome Res 26(3):291–300

Isaaks EH, Mohan Srivastava R (1989) An introduction to applied geostatistics

Jensen JD (2014) On the unfounded enthusiasm for soft selective sweeps. Nat Commun 5 (October):5281

Kamberov YG, Wang S, Tan J, Gerbault P, Wark A, Tan L, Yang Y et al (2013) Modeling recent human evolution in mice by expression of a selected EDAR variant. Cell 152(4):691–702

Kelly JK (1997) A test of neutrality based on interlocus associations. Genetics 146(3):1197–1206

Kemp JP, Medina-Gomez C, Estrada K, Pourcain BS, Heppe Denise H M, Warrington NM, Oei L et al (2014) Phenotypic dissection of bone mineral density reveals skeletal site specificity and facilitates the identification of novel loci in the genetic regulation of bone mass attainment. PLoS Genet 10(6):e1004423

Kemp JP, Medina-Gomez C, Tobias JH, Rivadeneira F, Evans DM (2016) The case for genome-wide association studies of bone acquisition in paediatric and adolescent populations. BoneKEy Reports 5(May):796

Kini U, Nandeesh BN (2012) Physiology of bone formation, remodeling, and metabolism. In: Radionuclide and hybrid bone imaging, pp 29–57

Koller DL, Zheng H-F, Karasik D, Yerges-Armstrong L, Liu C-T, McGuigan F, Kemp JP et al (2013) Meta-analysis of genome-wide studies identifies WNT16 and ESR1 SNPs associated with bone mineral density in premenopausal women. J Bone Mineral Res: Off J Am Soc Bone Mineral Res 28(3):547–558

Lamason RL, Mohideen Manzoor-Ali P K, Mest JR, Wong AC, Norton HL, Aros MC, Jurynec MJ et al (2005) SLC24A5, a putative cation exchanger, affects pigmentation in zebrafish and humans. Science 310(5755):1782–1786

Lewin R, Foley R (2004) Principles of human evolution. Blackwell publishing, Oxford

Li JZ, Absher DM, Tang H, Southwick AM, Casto AM, Ramachandran S, Cann HM et al (2008) Worldwide human relationships inferred from genome-wide patterns of variation. Science 319 (5866):1100–1104

Lu C, Fuchs E (2014) Sweat gland progenitors in development, homeostasis, and wound repair. Cold Spring Harbor Perspect Med 4(2). doi:10.1101/cshperspect.a015222

Malaspinas A-S, Westaway MC, Muller C, Sousa VC, Lao O, Alves I, Bergström A et al (2016) A genomic history of aboriginal Australia. Nature 538(7624):207–214

Marouli E, Graff M, Medina-Gomez C, Lo KS, Wood AR, Kjaer TR, Fine RS et al (2017) Rare and low-frequency coding variants alter human adult height. Nature 542(7640):186–190

Marshall LM, Zmuda JM, Chan BK, Barrett-Connor E, Cauley JA, Ensrud KE, Lang TF, Orwoll ES, Group OFiMMR (2008) Race and ethnic variation in proximal femur structure and BMD among older men. J Bone Miner Res 23(1):121–130

Mathieson I, Lazaridis I, Rohland N, Mallick S, Patterson N, Roodenberg SA, Harney E et al (2015) Genome-wide patterns of selection in 230 ancient Eurasians. Nature 528(7583): 499–503

McEvoy B, Beleza S, Shriver MD (2006) The genetic architecture of normal variation in human pigmentation: an evolutionary perspective and model. Hum Mol Genet 15(Spec No 2): R176–R181

Medina-Gómez C, Chesi A, Heppe Denise H M, Zemel BS, Yin J-L, Kalkwarf HJ, Hofman A et al (2015) BMD loci contribute to ethnic and developmental differences in skeletal fragility across populations: assessment of evolutionary selection pressures. Mol Biol Evol 32(11):2961–2972

Medina-Gomez C, Kemp JP, Estrada K, Eriksson J, Liu J, Reppe S, Evans DM et al (2012) Meta-analysis of genome-wide scans for total body BMD in children and adults reveals allelic heterogeneity and age-specific effects at the WNT16 locus. PLoS Genet 8(7):e1002718

Messer PW, Petrov DA (2013) Population genomics of rapid adaptation by soft selective sweeps. Trends Ecol Evol 28(11):659–669

Mode CJ, Sleeman CK (2012) Stochastic processes in genetics and evolution: computer experiments in the quantification of mutation and selection. World Scientific

Mondal M, Casals F, Xu T, Dall'Olio GM, Pybus M, Netea MG, Comas D et al (2016) Genomic analysis of Andamanese provides insights into ancient human migration into Asia and adaptation. Nat Genet 48(9):1066–1070

Nei M (1987) Molecular evolutionary genetics. Columbia University Press

Nielsen R, Akey JM, Jakobsson M, Pritchard JK, Tishkoff S, Willerslev E (2017) Tracing the peopling of the world through genomics. Nature 541(7637):302–310

Nielsen R, Williamson S, Kim Y, Hubisz MJ, Clark AG, Bustamante C (2005) Genomic scans for selective sweeps using SNP data. Genome Res 15(11):1566–1575

Norton HL, Kittles RA, Parra E, McKeigue P, Mao X, Cheng K, Canfield VA, Bradley DG, McEvoy B, Shriver MD (2007) Genetic evidence for the convergent evolution of light skin in Europeans and East Asians. Mol Biol Evol 24(3):710–722

Nowlan NC, Jepsen KJ, Morgan EF (2011) Smaller, Weaker, and Less Stiff Bones evolve from changes in subsistence strategy. Osteoporos Int: A Journal Established as Result of

Cooperation between the European Foundation for Osteoporosis and the National Osteoporosis Foundation of the USA 22(6):1967–1980
Pagani L, Lawson DJ, Jagoda E, Mörseburg A, Eriksson A, Mitt M, Clemente F et al (2016) Genomic analyses inform on migration events during the peopling of Eurasia. Nature 538 (7624):238–242
Perry JRB, Corre T, Esko T, Chasman DI, Fischer K, Franceschini N, He C et al (2013) A genome-wide association study of early menopause and the combined impact of identified variants. Hum Mol Genet 22(7):1465–1472
Pickrell J, Berisa T, Segurel L, Tung JY, Hinds D (2015) Detection and interpretation of shared genetic influences on 40 human traits. doi:10.1101/019885
Polimanti R, Gelernter J (2017) Widespread signatures of positive selection in common risk alleles associated to autism spectrum disorder. PLoS Genet 13(2):e1006618
Pritchard JK, Pickrell JK, Coop G (2010) The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. Curr Biol: CB 20(4):R208–R215
Putman MS, Yu EW, Lin D, Darakananda K, Finkelstein JS, Bouxsein ML (2017) Differences in trabecular microstructure between black and white women assessed by individual trabecular segmentation analysis of HR-pQCT images. J Bone Miner Res 32(5):1100–1108
Pybus M, Dall'Olio GM, Luisi P, Uzkudun M, Carreño-Torres A, Pavlidis P, Laayouni H, Bertranpetit J, Engelken J (2013) 1000 genomes selection browser 1.0: a genome browser dedicated to signatures of natural selection in modern humans. Nucleic Acids Res 42(D1): D903–D909
Pybus M, Luisi P, Dall'Olio GM, Uzkudun M, Laayouni H, Bertranpetit J, Engelken J (2015) Hierarchical boosting: a machine-learning framework to detect and classify hard selective sweeps in human populations. Bioinformatics 31(24):3946–3952
Qin P, Stoneking M (2015) Denisovan ancestry in East Eurasian and Native American populations. Mol Biol Evol 32(10):2665–2674
Racimo F, Sankararaman S, Nielsen R, Huerta-Sánchez E (2015) Evidence for archaic adaptive introgression in humans. Nat Rev Genet 16(6):359–371
Ramirez-Soriano A, Ramos-Onsins SE, Rozas J, Calafell F, Navarro A (2008) Statistical power analysis of neutrality tests under demographic expansions, contractions and bottlenecks with recombination. Genetics 179(1):555–567
Ramos-Onsins SE, Rozas J (2002) Statistical properties of new neutrality tests against population growth. Mol Biol Evol 19(12):2092–2100
Richards JB, Rivadeneira F, Inouye M, Pastinen TM, Soranzo N, Wilson SG, Andrew T et al (2008) Bone mineral density, osteoporosis, and osteoporotic fractures: a genome-wide association study. Lancet 371(9623):1505–1512
Rivadeneira F, Styrkársdottir U, Estrada K, Halldórsson BV, Hsu Y-H, Richards JB, Zillikens MC et al (2009) Twenty bone-mineral-density loci identified by large-scale meta-analysis of genome-wide association studies. Nat Genet 41(11):1199–1206
Robinson MR, Hemani G, Medina-Gomez C, Mezzavilla M, Esko T, Shakhbazov K, Powell JE et al (2015) Population genetic differentiation of height and body mass index across Europe. Nat Genet 47(11):1357–1362
Robinson MR, Kleinman A, Graff M, Vinkhuyzen AAE, Couper D, Miller MB, Peyrot WJ et al (2017) Genetic evidence of assortative mating in humans. Nat Hum Behav 1:0016
Ronen R, Tesler G, Akbari A, Zakov S, Rosenberg NA, Bafna V (2015) Haplotype Allele Frequency (HAF) score: predicting carriers of ongoing selective sweeps without knowledge of the adaptive allele. Lecture notes in computer science, pp 276–280
Ronen R, Tesler G, Akbari A, Zakov S, Rosenberg NA, Bafna V (2015b) Haplotype Allele Frequency (HAF) score: predicting carriers of ongoing selective sweeps without knowledge of the adaptive allele. Lecture notes in computer science, pp 276–280
Ronen R, Udpa N, Halperin E, Bafna V (2013) Learning natural selection from the site frequency spectrum. Genetics 195(1):181–193

Rozas J, Gullaud M, Blandin G, Aguadé M (2001) DNA variation at the rp49 gene region of Drosophila simulans: evolutionary inferences from an unusual haplotype structure. Genetics 158(3):1147–1155

Ryan TM, Shaw CN (2015) Gracility of the modern homo sapiens skeleton is the result of decreased biomechanical loading. Proc Natl Acad Sci USA 112(2):372–377

Sabeti PC, Reich DE, Higgins JM, Levine Haninah Z P, Richter DJ, Schaffner SF, Gabriel SB et al (2002) Detecting recent positive selection in the human genome from haplotype structure. Nature 419(6909):832–837

Sabeti PC, Varilly P, Fry B, Lohmueller J, Hostetter E, Cotsapas C, Xie X et al (2007) Genome-wide detection and characterization of positive selection in human populations. Nature 449(7164):913–918

Sankararaman S, Mallick S, Dannemann M, Prüfer K, Kelso J, Pääbo S, Patterson N, Reich D (2014) The genomic landscape of Neanderthal ancestry in present-day humans. Nature 507 (7492):354–357

Sankararaman S, Mallick S, Patterson N, Reich D (2016) The combined landscape of Denisovan and Neanderthal ancestry in present-day humans. Current Biology: CB 26(9):1241–1247

Sawyer GJ, Maley B (2005) Neanderthal reconstructed. Anat Rec Part B New Anat 283(1):23–31

Seeman E (2002) Pathogenesis of bone fragility in women and men. Lancet 359(9320):1841–1850

Shin M-H, Zmuda JM, Barrett-Connor E, Sheu Y, Patrick AL, Leung PC, Kwok A et al (2014) Race/ethnic differences in associations between bone mineral density and fracture history in older men. Osteoporos Int: Journal Established as Result of Cooperation between the European Foundation for Osteoporosis and the National Osteoporosis Foundation of the USA 25(3): 837–845

Shriver MD, Kennedy GC, Parra EJ, Lawson HA, Sonpar V, Huang J, Akey JM, Jones KW (2004) The genomic distribution of population substructure in four populations using 8,525 Autosomal SNPs. Hum Genom 1(4):274–286

Simonti CN, Vernot B, Bastarache L, Bottinger E, Carrell DS, Chisholm RL, Crosslin DR et al (2016) The phenotypic legacy of admixture between modern humans and Neandertals. Science 351(6274):737–741

Styrkarsdottir U, Thorleifsson G, Eiriksdottir B, Gudjonsson SA, Ingvarsson T, Center JR, Nguyen TV et al (2016a) Two rare mutations in the COL1A2 gene associate with low bone mineral density and fractures in Iceland. J Bone Mineral Res: Off J Am Soc Bone Mineral Res 31(1):173–179

Styrkarsdottir U, Thorleifsson G, Gudjonsson SA, Sigurdsson A, Center JR, Lee SH, Nguyen TV et al (2016b) Sequence variants in the PTCH1 gene associate with spine bone mineral density and osteoporotic fractures. Nat Commun 7:10129

Styrkarsdottir U, Thorleifsson G, Sulem P, Gudbjartsson DF, Sigurdsson A, Jonasdottir A, Jonasdottir A et al (2013) Nonsense mutation in the LGR4 gene is associated with several human diseases and other traits. Nature 497(7450):517–520

Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123(3):585–595

Tassi F, Ghirotto S, Mezzavilla M, Vilaça ST, De Santi L, Barbujani G (2015) Early modern human dispersal from Africa: genomic evidence for multiple waves of migration. Invest Genet 6(November):13

Tishkoff SA, Reed FA, Ranciaro A, Voight BF, Babbitt CC, Silverman JS, Powell K et al (2007) Convergent adaptation of human lactase persistence in Africa and Europe. Nat Genet 39(1): 31–40

Turchin, MC, Charleston WKC, Palmer CD, Sankararaman S, Reich D, Genetic Investigation of ANthropometric Traits (GIANT) Consortium, Hirschhorn JN (2012) Evidence of widespread selection on standing variation in Europe at height-associated SNPs. Nat Genet 44(9): 1015–1019

Vattathil S, Akey JM (2015) Small amounts of archaic admixture provide big insights into human history. Cell 163(2):281–284

Vitti JJ, Grossman SR, Sabeti PC (2013) Detecting natural selection in genomic data. Annu Rev Genet 47(1):97–120

Voight BF, Kudaravalli S, Wen X, Pritchard JK (2006) A map of recent positive selection in the human genome. PLoS Biol 4(3):e72

Wall JD (2000) A comparison of estimators of the population recombination rate. Mol Biol Evol 17(1):156–163

Wall JD (1999) Recombination and the power of statistical tests of neutrality. Genet Res 74(1): 65–79

Wang ET, Kodama G, Baldi P, Moyzis RK (2006) Global landscape of recent inferred darwinian selection for Homo sapiens. Proc Natl Acad Sci USA 103(1):135–140

Wei A-H, Zang D-J, Zhang Z, Liu X-Z, He X, Yang L, Wang Y et al (2013) Exome sequencing identifies SLC24A5 as a candidate gene for nonsyndromic oculocutaneous albinism. J Invest Dermatol 133(7):1834–1840

Wells JCK (2010) The evolutionary biology of human body fatness: thrift and control. Cambridge University Press

Welter D, MacArthur J, Morales J, Burdett T, Hall P, Junkins H, Klemm A et al (2014) The NHGRI GWAS catalog, a curated resource of SNP-trait associations. Nucleic Acids Res 42(Database issue):D1001–D1006

Wollstein A, Stephan W (2015) inferring positive selection in humans from genomic data. Invest Genet 6(April):5

Wollstein A, Walsh S, Liu F, Chakravarthy U, Rahu M, Seland JH, Soubrane G et al (2017) Novel quantitative pigmentation phenotyping enhances genetic association, epistasis, and prediction of human eye colour. Sci Rep 7(February):43359

Wood AR, Esko T, Yang J, Vedantam S, Pers TH, Gustafsson S, Chu AY et al (2014) Defining the role of common variation in the genomic and biological architecture of adult human height. Nat Genet 46(11):1173–1186

Wray NR, Yang J, Hayes BJ, Price AL, Goddard ME, Visscher PM (2013) Pitfalls of predicting complex traits from SNPs. Nat Rev Genet 14(7):507–515

Wu D-D, Zhang Y-P (2010) positive selection drives population differentiation in the skeletal genes in modern humans. Hum Mol Genet 19(12):2341–2346

Wu S, Tan J, Yang Y, Peng Q, Zhang M, Li J, Dongsheng L et al (2016) Genome-wide scans reveal variants at EDAR predominantly affecting hair straightness in han chinese and Uyghur populations. Hum Genet 135(11):1279–1286

Zhang L, Choi HJ, Estrada K, Leo PJ, Li J, Pei Y-F, Zhang Y et al (2014) Multistage genome-wide association meta-analyses identified two new loci for bone mineral density. Hum Mol Genet 23 (7):1923–1933

Zheng H-F, Forgetta V, Hsu Y-H, Estrada K, Rosello-Diez A, Leo PJ, Dahia CL et al (2015) Whole-genome sequencing identifies EN1 as a determinant of bone density and fracture. Nature 526(7571):112–117

Zheng H-F, Tobias JH, Duncan E, Evans DM, Eriksson J, Paternoster L, Yerges-Armstrong LM et al (2012) WNT16 influences bone mineral density, cortical bone thickness, bone strength, and osteoporotic fracture risk. PLoS Genet 8(7):e1002745

The Descent of Bison

Marie-Claude Marsolier-Kergoat and Jean-Marc Elalouf

Abstract Two bison species roamed the Eurasian continent during the Middle and Upper Pleistocene, the steppe bison, *Bison priscus*, and the woodland bison, *Bison schoetensacki*. Despite the wealth of fossil remains for these species, especially for the steppe bison, their genomic characterization started only a few years ago. Even now, when complete mitochondrial genomes are available for several specimens, information about nuclear genomes is still very fragmentary, limited to a few thousands positions at best. We present here our contribution to the characterization of these ancient bison genomes and to the clarification of their phylogeny.

1 Extinct and Extant Bison Species

The earliest members of the genus *Bison* appeared at the beginning of the Pleistocene in India and China. Bison then spread from Asia to Europe and America (Kurten 1968). During the Middle and Upper Pleistocene, two bison species, the steppe bison and the woodland bison, were part of the large Bovidae present in Europe and in northern Asia, along with the aurochs, *Bos primigenius* (Bojanus 1827). The steppe bison, *Bison priscus* (Bojanus 1827), was quite abundant and exhibited a wide geographic distribution extending from western Europe, through Central Asia and Beringia, and into North America. Endowed with long horns and

M.-C. Marsolier-Kergoat (✉) · J.-M. Elalouf (✉)
Institute for Integrative Biology of the Cell (I2BC), Institut des sciences du vivant Frédéric Joliot, CEA, CNRS, Université Paris-Sud, Université Paris-Saclay, 91198 Gif-sur-Yvette cedex, France
e-mail: mcmk@cea.fr

J.-M. Elalouf
e-mail: jean-marc.elalouf@cea.fr

M.-C. Marsolier-Kergoat · J.-M. Elalouf
CNRS-UMR 7206, Eco-anthropologie et Ethnobiologie, Département Homme et environnement, MNHN, CNRS, Université Paris Diderot, Musée de l'Homme, 17 place du Trocadéro et du 11 novembre, 75016 Paris, France

P. Pontarotti (ed.), *Evolutionary Biology: Self/Nonself Evolution, Species and Complex Traits Evolution, Methods and Concepts*,
DOI 10.1007/978-3-319-61569-1_10

robust legs, *Bison priscus* stood about two meters at the withers and was almost three meters long. This impressive animal occupied cool, steppe-like grasslands. Many bison depictions featuring anatomical details compatible with the morphology of the steppe bison and dating from the Aurignacian to the Magdalenian periods were found in painted caves such as the Chauvet, Trois-Frères, Lascaux, and Pech Merle caves in France (Soubrier et al. 2016). *Bison priscus* became extinct in Europe at the end of the last Ice Age, about 12,000 years ago (Kurten 1968).

The woodland bison, *Bison schoetensacki* (Freudenberg 1910) appeared in the lower Middle Pleistocene. Its size was almost as large as that of *Bison priscus* but its leg bones and metapodials were slenderer (Guérin and Valli 2000; Vercoutère and Guérin 2010). Moreover, the horns of *Bison schoetensacki* were shorter and of a slightly different shape than those of *Bison priscus*. *Bison schoetensacki* fossil remains, which are less abundant than those of *Bison priscus*, are often associated with forest biotopes. The geographic distribution of *Bison schoetensacki* stretched from western Europe to the south of Siberia (Fig. 1), but unlike *Bison priscus*, *Bison*

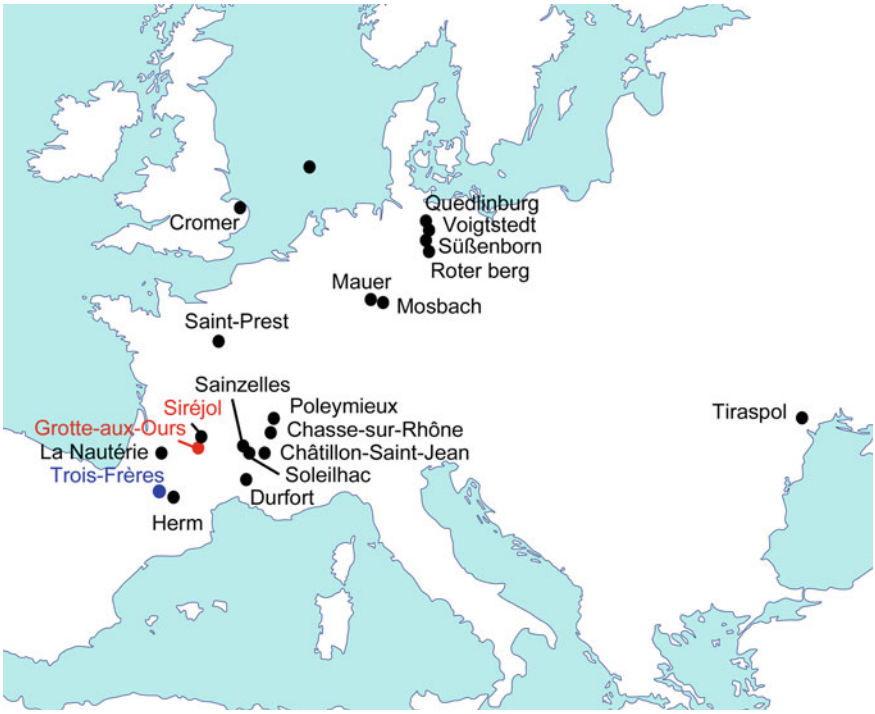


Fig. 1 European cave sites with *Bison schoetensacki* fossil records. *Black dots* indicate sites where *Bison schoetensacki* remains have been reported in previous studies. *Red characters* indicate the two cave sites that yielded the *Bison schoetensacki* bone fragment (Siréjol) and the hyena coprolite containing *Bison schoetensacki* DNA (Grotte-aux-Ours). *Blue characters* identify the site where the *Bison priscus* SGE2 bone fragment was collected (Trois-Frères)

schoetensacki was absent from Beringia and America. The species was first described in Germany where it has been found in several early Middle Pleistocene sites, but it has also been recorded in England, Moldova and Russia. In France, *Bison schoetensacki* remains have been described in many sites including the archeological site of Châtillon-Saint-Jean (Drôme) (Mourer-Chauviré 1972), the Herm cave (Ariège) and the Siréjol cave (Souillac, Lot) (Guérin and Philippe 1971).

Two bison species exist today: the American bison and the European bison. The American bison, in North America, includes two subspecies, the plains bison, *Bison bison bison*, and the wood bison, *Bison bison athabascae*. The European bison, *Bison bonasus,* is found in Europe and the Caucasus, where it has been reintroduced after its extinction as a wild species in the early twentieth century. The *Bison bonasus* genomes reflect a complex descent. Indeed, the *Bison bonasus* nuclear genome is closely related to that of *Bison bison* (Verkaar et al. 2004; Nijman et al. 2008; Hassanin et al. 2013), in agreement with morphological evidence and the fact that the two bison species can produce completely fertile hybrid offspring. However, the mitochondrial genomes of *Bison bonasus* specimens are more similar to cattle (modern *Bos primigenius*) genomes than to *Bison bison* genomes (Verkaar et al. 2004; Zeyland et al. 2012). These observations could be explained either by incomplete lineage sorting of the mitochondrial genome or by a scenario according to which the nuclear DNA of an ancient population probably related to the extinct aurochs (itself closely related to cattle) would have been changed by the systematic introgression of bison bulls (Verkaar et al. 2004; Hassanin et al. 2012; Bibi 2013).

At the time when we began our analyses, the genetic history of *Bison bison* had been extensively investigated through the analysis of mitochondrial DNA sequences (Shapiro et al. 2004; Llamas et al. 2012). The scenario emerging from these mitochondrial DNA studies was that a population of *Bison priscus* had first entered Beringia from Asia during the Middle Pleistocene, around 300 to 130 ky ago, and then moved southward into central North America approximately 130 to 75 ky ago (Shapiro et al. 2004). Genetic exchanges between bison populations in Beringia and central North America had taken place during long periods but had been recently (soon after 14 ky B.P.) limited by the establishment of spruce forest in Alberta and of peatland across northwestern Canada (Shapiro et al. 2004). In addition to these analyses on mitochondrial sequences, nuclear single nucleotide polymorphism (SNP) genotyping (Decker et al. 2009) had established that *Bison priscus* and *Bison bison* nuclear genomes were sister genomes. However, only a small portion of the *Bison priscus* mitochondrial genome, the D-loop region, representing less than 5% of the expected 17-kb complete genome, had been characterized at that time (Shapiro et al. 2004; Llamas et al. 2012), and we endeavored to provide the first complete mitochondrial genome for a *Bison priscus* specimen (Marsolier-Kergoat et al. 2015). Regarding *Bison schoetensacki*, its phylogenetic relationships with *Bison bonasus* and *Bison priscus* were largely debated. Some authors had suggested that *Bison schoetensacki* could be the ancestor of *Bison bonasus* (Kurten 1968), but others considered that *Bison bonasus* was derived from an unknown form of *Bison priscus* (Benecke 2005; Croitor 2010). The lack of genomic data for *Bison schoetensacki* had so far prevented any conclusive view on this point, which was clarified by the

mitochondrial and the nuclear genomic data that we and others have obtained these last few months (Soubrier et al. 2016; Massilani et al. 2016; Palacio et al. 2017).

2 Establishing the Mitochondrial Genome Sequence of *Bison priscus*

2.1 *Selection of a Suitable Archeological Sample in the Trois-Frères Cave*

We analyzed several bone samples that were collected from the Trois-Frères cave (Ariège, France, see Fig. 1) (Marsolier-Kergoat et al. 2015). The Trois-Frères cave consists of a series of chambers that spread over 800 m, from the Enlène cave to the Tuc d'Audoubert cave, from which it is currently disconnected. This cave was named after the three sons of the archeologist Henri Bégouën who discovered its entrance in July 1914. It contains numerous animal representations, including the drawing and engraving of at least 170 bison (Bégouën and Breuil 1958). The importance of the bison for the artists who decorated the cave is evidenced by the fact that bison are the most frequently represented animals and also by the presence of a therianthropic figure with a bison upper part. Besides, the adjacent Tuc d'Audoubert cave contains two modeled clay bison, a masterpiece of the Paleolithic period (Bégouën 1912). In this cave system, the *Salle du Grand Éboulis* (Chamber of the Large Scree) functioned as a natural trap for Pleistocene animals and was sealed before the Holocene (Bégouën et al. 2014). The bottom of the scree contains remains of the extinct cave bear (*Ursus spelaeus*), one of which was dated to 36,600–34,800 calBP (Bégouën et al. 2014). In the other layers 19,400- to 17,800-year-old remains of the steppe bison predominate (Bégouën et al. 2014). Out of the four bone samples that were collected in the *Salle du Grand Éboulis*, only one (the SGE2 sample) was devoid of contamination by modern cattle DNA, probably brought by human intrusions since the discovery of the cave. As shown in Fig. 2a, the SGE2 sample, a rib fragment, stood on a rock, several meters away from the path used by modern visitors, which probably explains that it had remained uncontaminated. SGE2 was radiocarbon dated to 15,880 ± 70 B.P. (19,390–18,940 calBP, Fig. 2b), a value consistent with the age previously obtained for *Bison priscus* bones collected in this sector (Bégouën et al. 2014).

2.2 *The Mitochondrial Genome of the SGE2* Bison priscus *Specimen from the Trois-Frères Cave*

A library of DNA fragments was produced from the SGE2 sample, and a total of 1,033,460,536 Illumina single-end reads, each at least 20 nucleotides in length,

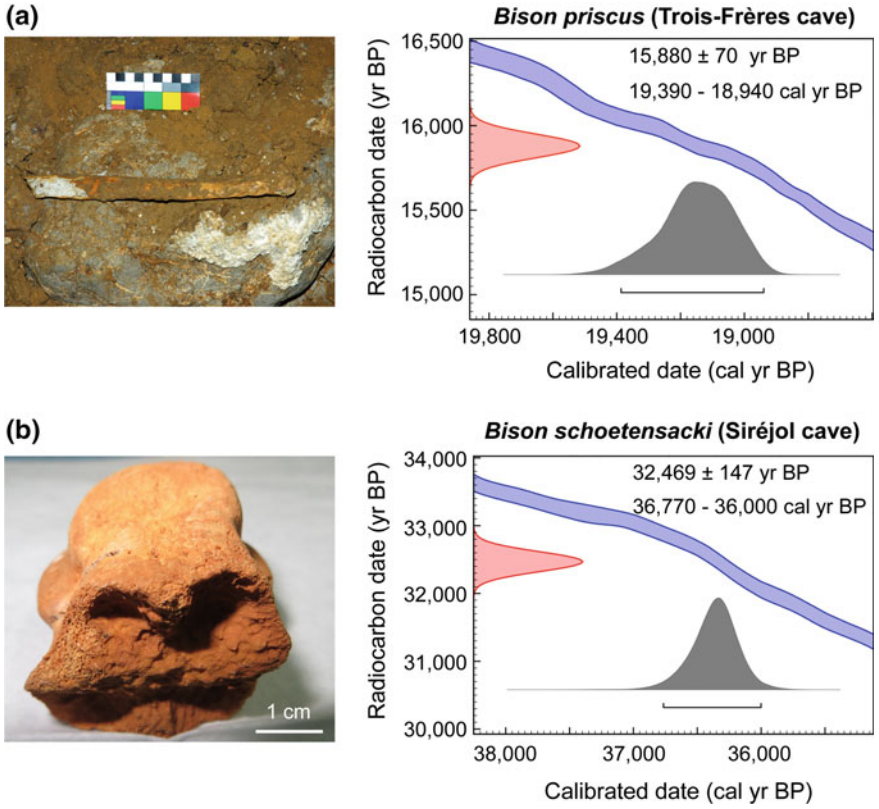


Fig. 2 Samples and radiocarbon dating analyses. **a** The *Bison priscus* SGE2 rib fragment as originally found in the Trois-Frères cave and the curves showing the uncalibrated (yr BP) and calibrated (cal yr BP) ages of the sample. **b** The *Bison schoetensacki* cannon bone sample collected in the Siréjol cave and its uncalibrated and calibrated ages. The calBP data correspond to 95.4% (2 σ) confidence interval of the sample age using IntCal13 calibration curve (Reimer et al. 2013)

were generated. We evaluated the coverage of *Bison priscus* nuclear genome by these reads by mapping them onto the nuclear genome of modern *Bos primigenius*, the closest related species of *Bison priscus* whose nuclear genome had been completely sequenced. This mapping yielded 741,848 unique reads, with a cumulative length of 26,117,728 bp, which represents a 0.01-fold coverage of the nuclear genome. This very low coverage precluded any analysis of the *Bison priscus* nuclear genome. We assembled the *Bison priscus* mitochondrial genome by aligning the reads to the mitochondrial genome of *Bison bison,* and we complemented this approach by PCR experiments to analyze the regions where less than two concordant reads were available. The final sequence, termed SGE2seq, consisted of a circular genome, for which we obtained a mean coverage of 10.4-fold from 3,851 unique Illumina reads. Only 1,443 mismatches out of 166,849 aligned

bases were observed between the 3,851 reads and the consensus sequence SGE2seq. The frequency of these mismatches, corresponding predominantly to G-to-A substitutions, increases at the 3′ end of the reads, which is considered as a hallmark authenticating sequences generated from ancient DNA fragments (Briggs et al. 2007). SGE2seq corresponds to a mitochondrial genome sequence of 16,318 bp. This length is similar to the lengths of *Bison bison* mitochondrial genomes, which range from 16,318 to 16,323 bp (Douglas et al. 2011). As expected, the genome consists of 13 protein-coding genes, 22 tDNA genes, 2 rDNA genes, and the D-loop region.

SGE2seq shows 114 differences compared to the reference *Bison bison* mitochondrial genome (GenBank accession number NC_012346.1), including two indels: a 1-bp insertion located in the D-loop region and a 2-bp deletion at the end of the tRNA serine gene. The other differences consist in 105 transitions and six

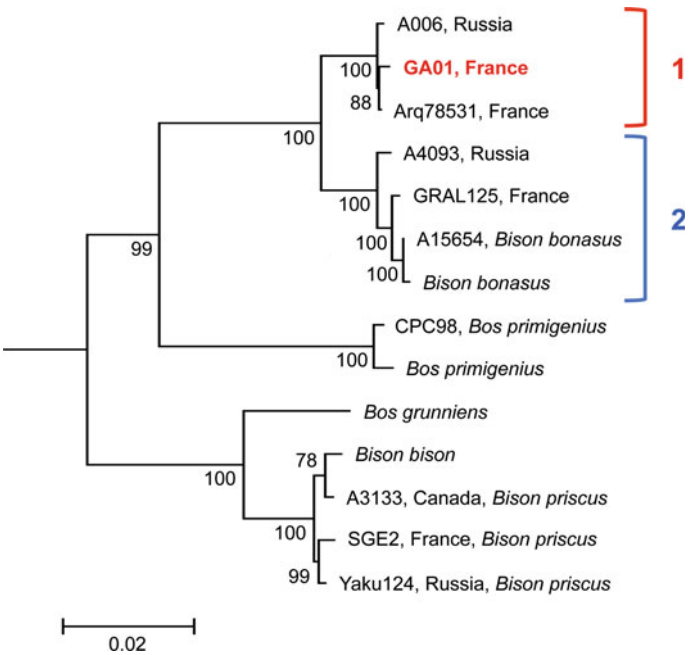


Fig. 3 Maximum Likelihood phylogenetic tree of complete mitochondrial genomes of *Bison* and *Bos* species. The tree with the highest log-likelihood is shown drawn to scale, with branch lengths established from the numbers of substitutions per site. The percentages of trees in which the associated taxa clustered together are displayed next to the branches (the bootstrap values were determined with 500 replicates). Here is the top-down list of the sequences GenBank accession numbers: KX592187, NC_033873, KX898007, KX592175, KX898009, KX592176, NC_014044, NC_013996, NC_006853, NC_006380, NC_012346, KX592174, NC_027233, and KX898020. Specimen name and country of origin are indicated for ancient samples, the other genomes indicated only by the species name correspond to the NCBI reference mitochondrial genomes. The tree was rooted using the reference mitochondrial sequence of *Bubalus bubalis* (GenBank accession number NC_006295). See main text for explanations about clades 1 and 2

transversions. A similar, strong transitional bias for mitochondrial genomes has also been reported for *Bos* species (Edwards et al. 2010). Twenty-five differences (22%) are located in the D-loop region, which therefore exhibits a mutation rate four times higher than the rest of the genome. Finally, 73 substitutions (65%) are located in protein-coding genes, with 16, 3 and 54 substitutions occurring in the first, second and third codon positions, respectively. These substitutions result in 11 amino acid differences. As shown in Fig. 3, phylogenetic analyses demonstrated that SGE2seq, along with several other *Bison priscus* mitochondrial genomes that have been reported since then, forms a distinct and well-supported clade (100% bootstrap support, 500 replicates) with the reference mitochondrial genome of the American bison *Bison bison*. These analyses, in combination with nuclear SNP genotyping (Decker et al. 2009), definitively established that *Bison priscus* and *Bison bison* are sister groups.

3 Genome Data on the Extinct *Bison schoetensacki* Establish it as a Sister Species of the European Bison (*Bison bonasus*)

3.1 Serendipitous Discovery of DNA from an Unknown Bovine Species in a Cave Hyena Coprolite

Coprolites, i.e. fossilized feces, contain DNA from both the defecator and the organisms forming its diet (Poinar et al. 1998; Gilbert et al. 2008; Bon et al. 2012), which makes them a valuable source of information on the genomes of both predators and preys. As part of a long-term effort to obtain genomic information on the diet of the cave hyena (*Crocuta crocuta*), we initiated several years ago the analysis of cave hyena coprolites (Bon et al. 2012). In this study, we analyzed an intact cave hyena coprolite that we had collected in the Grotte-aux-Ours cave (Souillac, Lot, France, Fig. 1). The name of this cave, which was discovered in 2008, refers to the abundant evidence for occupation by *Ursus spelaeus*, including bones, hibernation nests, claw marks, and footprints on the paleosurface. Intrusions by cave hyenas are attested by the presence of coprolites and a hyena skull.

DNA was extracted from the central part of the coprolite, and we produced a library of DNA fragments that was analyzed through shotgun high-throughput DNA sequencing. To confirm the identity of the coprolite producer and gain a molecular insight into its diet, a total of 601,509,879 Illumina single-end reads were aligned simultaneously to a set of 49 mitochondrial genomes, including the reference mitochondrial genomes for the cave hyena and for several species likely to be part of its diet. This analysis confirmed that a cave hyena was indeed the producer of the coprolite, since the number of reads (7,550) matching perfectly, without indel or mismatch, the cave hyena genome was by far the largest. Interestingly, *Bison bonasus* mitochondrial genome was the second best covered genome with 3,220

reads, which suggested the ingestion by the coprolite producer of a bovine specimen (hereafter referred to as "the GAO bovine") whose closest known relative was *Bison bonasus*. The alignment of the Illumina reads to *Bison bonasus* reference mitochondrial genome, complemented by PCR experiments to derive a robust sequence at positions covered by less than two independent reads, led to the assembly of a 16,325-bp bovine mitochondrial genome, termed GAOseq_Bovinae, with a 32-fold median coverage. When comparing the differences between the aligned reads and the consensus sequence GAOseq_Bovinae, we noticed again an increasing G-to-A substitution rate at the 3′ end of the reads, marking them as sequences generated from ancient, damaged DNA fragments. Dating analysis of the coprolite failed because the sample completely dissolved during the pretreatment for Atomic Mass Spectroscopy (AMS) measurement, but since the cave hyena vanished from Europe at about 30,000 cal yr BP (Stuart and Lister 2014), we can surmise that the Grotte-aux-Ours hyena coprolite is at least 30,000 years old, which is consistent with the characteristics of the reads aligning to GAOseq_Bovinae.

Phylogenetic analyses were performed using several mitochondrial genomes from Late Pleistocene/Holocene bison specimens related to *Bison bonasus* that were published at the time our manuscript describing GAOseq_Bovinae was under review (Soubrier et al. 2016; Massilani et al. 2016), along with the reference mitochondrial genomes of *Bos primigenius*, *Bison bison*, *Bos grunniens* (yak) and *Bubalus bubalis* (swamp buffalo). These analyses positioned GAOseq_Bovinae in a well-supported clade (clade 1) comprising all the ancient genomes forming CladeX in (Soubrier et al. 2016) and clade Bb1 in (Massilani et al. 2016) (Fig. 3). No paleontologically defined species name had been proposed so far for clade 1, which represents the sister group of clade 2, the set of mitochondrial genomes of all *Bison bonasus* ancient and modern specimens.

3.2 *Identification of the Coprolite Bovinae DNA Through Analysis of a* Bison schoetensacki *Bone Sample*

Because the Grotte-aux-Ours cave that had yielded the hyena coprolite was located close to Siréjol (Fig. 1), we also analyzed the DNA content of a *Bison schoetensacki* cannon bone fragment (Fig. 2b) that had been collected from the Siréjol cave during excavations carried out between 1972 and 1975. The bone sample was radiocarbon dated to 36,770–36,000 cal yr BP (Fig. 2b), in agreement with previous ^{14}C dating of bulk bone material from the same cave sector. The content of bovine DNA in the bone sample was much lower than in the hyena coprolite, so that we resorted to PCR experiments to determine *Bison schoetensacki* mitochondrial sequences. A set of 16 primer pairs allowed PCR amplification of short mitochondrial DNA fragments, yielding information on a total of 609 bp. The whole set of *Bison schoetensacki* sequences only differed at 1 and 6 positions, respectively, from the orthologous sequences of the Arq78531 and

GAOseq_Bovinae genomes, and phylogenetic analyses confirmed the position of the Siréjol *Bison schoetensacki* sequences within clade 1 (Palacio et al. 2017). These results clearly identified GAOseq_Bovinae as the mitochondrial genome of a *Bison schoetensacki* specimen.

3.3 *Analysis of* Bison schoetensacki *Nuclear SNPs*

Soubrier et al. (Soubrier et al. 2016) had analyzed a set of ~10,000 genome-wide bovine SNPs from one specimen (A006) of their *Bison bonasus*-related CladeX, from one historical (A15654) and one ancient (A4093) *Bison bonasus* specimens and from two *Bison priscus* specimens. In order to compare the *Bison schoetensacki* nuclear sequences of the coprolite to these data, we determined the genotypes of the same bovine SNPs by mapping the library reads against the taurine cattle reference genome. Multidimensional scaling analyses (Fig. 4) showed that the closest specimen to our *Bison shoetensacki* specimen was the CladeX specimen A006. This result corroborated the conclusions drawn from the mitochondrial sequence data that position the *Bison schoetensacki* sequences of the coprolite in clade 1, the sister group to *Bison bonasus* genomes.

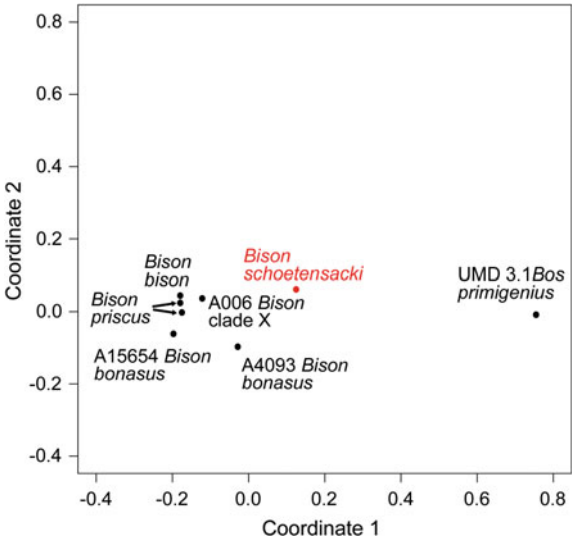


Fig. 4 Multidimensional scaling analysis of bovine nuclear SNPs. We considered the SNPs genotyped in the *Bison schoetensacki* sequences of the coprolite, in a modern *Bison bison* specimen (*Bison bison bison* isolate TAMUID 2011002044) and in the reference genome of *Bos primigenius* (taurine cattle reference UMD 3.1) as well as the SNPs genotyped by Soubrier et al. (2016) in the CladeX A006 specimen, in the historical *Bison bonasus* A15654 specimen, in the ancient *Bison bonasus* A4093 specimen, and in the two *Bison priscus* A875 and A3133 specimens

The most ancient remains of *Bison schoetensacki* specimens date back to the beginning of the Middle Pleistocene (i.e. some 750,000 years ago). Our studies, combined with the results of others (Soubrier et al. 2016; Massilani et al. 2016), have shown that this species was still present during the Upper Pleistocene in a number of Eurasian sites. We have demonstrated that our *Bison schoetensacki* specimen belongs to a bison clade previously referred to as CladeX (Soubrier et al. 2016) or clade Bb1 (Massilani et al. 2016), that should be renamed appropriately. Moreover, the emergence of *Bison bonasus* as derived from *Bison schoetensacki* lineages rather than from *Bison priscus* lineages is now clarified. The divergence date between *Bison bonasus* and *Bison schoetensacki* remains unclear since it has been estimated around 120 (152-92) kya and around 246 (283-212) kya in the aforementioned studies with different datasets (Soubrier et al. 2016; Massilani et al. 2016). Opposite conclusions have also been reached regarding the evolution of *Bison bonasus* as due either to incomplete lineage sorting (Massilani et al. 2016) or to gene introgression between *Bison priscus* males and *Bos primigenius* females (Soubrier et al. 2016). Progress toward a more precise dating of Bovinae evolution and a clarification of *Bison bonasus* descent should come in the near future from the genomic analysis of still older fossil specimens.

Acknowledgements We thank the Plateau Technique du MNHN, site du Musée de l'Homme (Plateforme Paléogénomique et Génétique Moléculaire, Musée de l'Homme, Paris) and the Service de Systématique Moléculaire (UMS 2700 OMSI CNRS-MNHN) for their contribution to this work. We also wish to thank Daniel Dalet for the authorization to use the base map (http://d-maps.com/carte.php?num_car=2224&lang=fr) used in Fig. 1.

References

Benecke N (2005) The Holocene distribution of European bison—the archaeozoological record. Munibe (Anthropologia-Arkeologia) 57: 421–428

Bégouën H (1912) Les statues d'argile préhistoriques de la caverne du Tuc d'Audoubert (Ariège). Comptes rendus des séances de l'Académie des Inscriptions et Belles-Lettres 532

Bégouën H, Breuil H (1958) Les cavernes du Volp: Trois Frères-Tuc d'Audoubert à Montesquieu-Avantès (Ariège). Arts et Métiers Graphiques, Paris

Bégouën R, Clottes J, Feruglio V, Pastoors A (2014) La caverne des Trois-Frères. Somogy éditions d'art, Paris

Bibi F (2013) A multi-calibrated mitochondrial phylogeny of extant Bovidae (Artiodactyla, Ruminantia) and the importance of the fossil record to systematics. BMC Evol Biol 13:166. doi:10.1186/1471-2148-13-166

Bon C, Berthonaud V, Maksud F, Labadie K, Poulain J, Artiguenave F, Wincker P, Aury J-M, Elalouf J-M (2012) Coprolites as a source of information on the genome and diet of the cave hyena. Proc Biol Sci 279:2825–2830. doi:10.1098/rspb.2012.0358

Briggs AW, Stenzel U, Johnson PLF, Green RE, Kelso J, Prüfer K, Meyer M, Krause J, Ronan MT, Lachmann M, Pääbo S (2007) Patterns of damage in genomic DNA sequences from a Neandertal. Proc Natl Acad Sci USA 104:14616–14621. doi:10.1073/pnas.0704665104

Croitor R (2010) Kriticheskie zamechania o bisonakh iz Pleostocena Moldovy (Bison, Bovidae, Mammalia). Revista Arheologocă V: 172–188

Decker JE, Pires JC, Conant GC, McKay SD, Heaton MP, Chen K, Cooper A, Vilkki J, Seabury CM, Caetano AR, Johnson GS, Brenneman RA, Hanotte O, Eggert LS, Wiener P, Kim J-J, Kim KS, Sonstegard TS, Van Tassell CP, Neibergs HL, McEwan JC, Brauning R, Coutinho LL, Babar ME, Wilson GA, McClure MC, Rolf MM, Kim J, Schnabel RD, Taylor JF (2009) Resolving the evolution of extant and extinct ruminants with high-throughput phylogenomics. Proc Natl Acad Sci USA 106:18644–18649. doi:10.1073/pnas.0904691106

Douglas KC, Halbert ND, Kolenda C, Childers C, Hunter DL, Derr JN (2011) Complete mitochondrial DNA sequence analysis of Bison bison and bison-cattle hybrids: function and phylogeny. Mitochondrion 11:166–175. doi:10.1016/j.mito.2010.09.005

Edwards CJ, Magee DA, Park SDE, McGettigan PA, Lohan AJ, Murphy A, Finlay EK, Shapiro B, Chamberlain AT, Richards MB, Bradley DG, Loftus BJ, MacHugh DE (2010) A complete mitochondrial genome sequence from a mesolithic wild aurochs (Bos primigenius). PLoS ONE 5:e9255. doi:10.1371/journal.pone.0009255

Gilbert MTP, Jenkins DL, Götherstrom A, Naveran N, Sanchez JJ, Hofreiter M, Thomsen PF, Binladen J, Higham TFG, Yohe RM, Parr R, Cummings LS, Willerslev E (2008) DNA from pre-Clovis human coprolites in Oregon, North America. Science 320:786–789. doi:10.1126/science.1154116

Guérin C, Philippe M (1971) Les gisements de vertébrés pléistocènes du Causse de Martel. Bull Soc Hist Archéol 93:31–46

Guérin C, Valli AMF (2000) Le gisement pléistocène supérieur de la grotte de Jaurens à Nespouls, Corrèze: les Bovidae (Mammalia, Artiodactyla). Cahiers scientifiques Mus hist nat Lyon 7–39

Hassanin A, An J, Ropiquet A, Nguyen TT, Couloux A (2013) Combining multiple autosomal introns for studying shallow phylogeny and taxonomy of Laurasiatherian mammals: Application to the tribe Bovini (Cetartiodactyla, Bovidae). Mol Phylogenet Evol 66:766–775. doi:10.1016/j.ympev.2012.11.003

Hassanin A, Delsuc F, Ropiquet A, Hammer C, Jansen van Vuuren B, Matthee C, Ruiz-Garcia M, Catzeflis F, Areskoug V, Nguyen TT, Couloux A (2012) Pattern and timing of diversification of Cetartiodactyla (Mammalia, Laurasiatheria), as revealed by a comprehensive analysis of mitochondrial genomes. C R Biol 335:32–50. doi:10.1016/j.crvi.2011.11.002

Kurten B (1968) Pleistocene mammals of Europe. Aldine Transaction, Chicago

Llamas B, Holland ML, Chen K, Cropley JE, Cooper A, Suter CM (2012) High-resolution analysis of cytosine methylation in ancient DNA. PLoS ONE 7:e30226. doi:10.1371/journal.pone.0030226

Marsolier-Kergoat M-C, Palacio P, Berthonaud V, Maksud F, Stafford T, Bégouën R, Elalouf J-M (2015) Hunting the extinct steppe bison (Bison priscus) mitochondrial genome in the Trois-Frères paleolithic painted cave. PLoS ONE 10:e0128267. doi:10.1371/journal.pone.0128267

Massilani D, Guimaraes S, Brugal J-P, Bennett EA, Tokarska M, Arbogast R-M, Baryshnikov G, Boeskorov G, Castel J-C, Davydov S, Madelaine S, Putelat O, Spasskaya NN, Uerpmann H-P, Grange T, Geigl E-M (2016) Past climate changes, population dynamics and the origin of bison in Europe. BMC Biol 14:93. doi:10.1186/s12915-016-0317-7

Mourer-Chauviré C (1972) Etude de nouveaux restes de vertébrés provenant de la carrière Fournier à Châtillon-Saint-Jean (Drôme) III - Artiodactyles, chevaux et oiseaux. Bull Ass fr Et quat 271–305

Nijman IJ, van Boxtel DCJ, van Cann LM, Marnoch Y, Cuppen E, Lenstra JA (2008) Phylogeny of Y chromosomes from bovine species. Cladistics 24:723–726

Palacio P, Berthonaud V, Guérin C, Lambourdière J, Maksud F, Philippe M, Plaire D, Stafford T, Marsolier-Kergoat M-C, Elalouf J-M (2017) Genome data on the extinct Bison schoetensacki establish it as a sister species of the extant European bison (Bison bonasus). BMC Evol Biol 17:48. doi:10.1186/s12862-017-0894-2

Poinar HN, Hofreiter M, Spaulding WG, Martin PS, Stankiewicz BA, Bland H, Evershed RP, Possnert G, Paabo S (1998) Molecular coproscopy: dung and diet of the extinct ground sloth Nothrotheriops shastensis. Science 281:402–406

Reimer PJ, Bard E, Bayliss A, Beck JW, Blackwell PG, Ramsey CB, Buck CE, Cheng H, Edwards RL, Friedrich M, Grootes PM, Guilderson TP, Haflidason H, Hajdas I, Hatte C, Heaton TJ, Hoffmann DL, Hogg AG, Hughen KA, Kaiser KF, Kromer B, Manning SW, Niu M, Reimer RW, Richards DA, Scott EM, Southon JR, Staff RA, Turney CSM, van der Plicht J (2013) Intcal13 and Marine13 radiocarbon age calibration curves 0-50,000 years cal bp. Radiocarbon 55:1869–1887

Shapiro B, Drummond AJ, Rambaut A, Wilson MC, Matheus PE, Sher AV, Pybus OG, Gilbert MTP, Barnes I, Binladen J, Willerslev E, Hansen AJ, Baryshnikov GF, Burns JA, Davydov S, Driver JC, Froese DG, Harington CR, Keddie G, Kosintsev P, Kunz ML, Martin LD, Stephenson RO, Storer J, Tedford R, Zimov S, Cooper A (2004) Rise and fall of the Beringian steppe bison. Science 306:1561–1565. doi:10.1126/science.1101074

Soubrier J, Gower G, Chen K, Richards SM, Llamas B, Mitchell KJ, Ho SYW, Kosintsev P, Lee MSY, Baryshnikov G, Bollongino R, Bover P, Burger J, Chivall D, Cregut-Bonnoure E, Decker JE, Doronichev VB, Douka K, Fordham DA, Fontana F, Fritz C, Glimmerveen J, Golovanova LV, Groves C, Guerreschi A, Haak W, Higham T, Hofman-Kamińska E, Immel A, Julien M-A, Krause J, Krotova O, Langbein F, Larson G, Rohrlach A, Scheu A, Schnabel RD, Taylor JF, Tokarska M, Tosello G, van der Plicht J, van Loenen A, Vigne J-D, Wooley O, Orlando L, Kowalczyk R, Shapiro B, Cooper A (2016) Early cave art and ancient DNA record the origin of European bison. Nat Commun 7:13158. doi:10.1038/ncomms13158

Stuart AJ, Lister AM (2014) New radiocarbon evidence on the extirpation of the spotted hyaena (Crocuta crocuta (Erxl.)) in northern Eurasia. Q Sci Rev 96:108–116

Vercoutère C, Guérin C (2010) Les Bovidae (Mammalia, Artiodactyla) du Pléistocène moyen final de l'Aven de Romain-la-Roche (Doubs, France). Rev Paléobiol 29:655–696

Verkaar ELC, Nijman IJ, Beeke M, Hanekamp E, Lenstra JA (2004) Maternal and paternal lineages in cross-breeding bovine species. Has wisent a hybrid origin? Mol Biol Evol 21:1165–1170. doi:10.1093/molbev/msh064

Zeyland J, Wolko L, Lipiński D, Woźniak A, Nowak A, Szalata M, Bocianowski J, Słomski R (2012) Tracking of wisent-bison-yak mitochondrial evolution. J Appl Genet 53:317–322. doi:10.1007/s13353-012-0090-4

Convergent and Parallel Evolution in Early Glires (Mammalia)

Łucja Fostowicz-Frelik

Abstract Glires (lagomorphs, rodents, and their kin), based on molecular and morphological evidence, form a monophyletic clade nested within Euarchontoglires, a large clade including also primates, tree shrews (Scandentia), and flying lemurs (Dermoptera). The earliest currently known Glires are represented by duplicidentate lineage (closer to lagomorphs than to rodents), which appeared shortly after the K/Pg boundary in Asia. Evolution of Glires is interspersed with instances of convergence between its two main branches: Duplicidentata and Simplicidentata (rodents, eurymylids, and their relatives), and also convergences on basal Euarchontoglires (Pseudictopidae), Anagalidae, or even some ungulate lineages. Within more closely related basal lines, parallel evolution is frequent. Homoplastic characters manifest themselves mainly in the teeth and appendicular skeleton. An important example of convergent evolution within Glires (and broader within Euarchontoglires) is the structure of the tarsal joint, and the calcaneal morphology in particular. Lagomorphs, similar to ungulates and elephant shrews (and several fossil taxa), have two articulation facets at the eminence of the calcaneus. The calcaneoastragalar facet is typical of all mammals, while the calcaneofibular one is characteristic of all true lagomorphs and probably *Mimotona*, the earliest duplicidentate, but is absent in *Gomphos* and *Mimolagus*. Further, it is known in *Rhombomylus*, a Paleogene Asian eurymylid (basal simplicidentate), and in Paleocene *Pseudictops*, a basal Euarchontoglires. The calcaneofibular facet stabilizes the tarsal joint, thus contributing to increased cursoriality. Overall, this chapter focuses on convergence and parallelism observed in the dental and pedal characters of the early Glires, with emphasis on the duplicidentates.

Ł. Fostowicz-Frelik (✉)
Institute of Paleobiology, Polish Academy of Sciences, PL 00-818 Warsaw, Poland
e-mail: lfost@twarda.pan.pl; lucja_fostowicz@yahoo.com

Ł. Fostowicz-Frelik
Institute of Vertebrate Paleontology and Paleoanthropology,
Chinese Academy of Sciences, 100044 Beijing, People's Republic of China

P. Pontarotti (ed.), *Evolutionary Biology: Self/Nonself Evolution, Species and Complex Traits Evolution, Methods and Concepts*,
DOI 10.1007/978-3-319-61569-1_11

1 Introduction

The Glires includes two extant clades, Lagomorpha and Rodentia, containing nearly half of living mammalian species; their fossil record is even more diverse. Together with Archonta (primates, flying lemurs or colugos, and tree shrews, but see below), it is one of the main placental groups (the remaining are Afrotheria, Xenarthra, and Laurasiatheria; Asher 2007). The most obvious morphological diagnostic characteristics of Glires is a pair of enlarged ever-growing incisors, both in the (pre)-maxilla and in mandible, with the enamel covering only labial surfaces (a character shared with zalambdalestids in the lower incisor; Fostowicz-Frelik and Kielan-Jaworowska 2002). The monophyly of Glires is now generally accepted (Murphy et al. 2001; Meng and Wyss 2001, 2005; Springer et al. 2005); however, many phylogenetic questions within this clade are still unresolved.

As currently understood, Rodentia and their stem clades form Simplicidentata (Wyss and Meng 1996). Among fossil outgroups closest to rodents are eurymylids, e g., *Eurymylus*, *Heomys*, *Matutinia,* and *Rhombomylus* (Sych 1971; Li 1977; Ting et al. 2002; Meng et al. 2003).

A more inclusive clade for Lagomorpha is Duplicidentata, defined as all Glires sharing a more recent common ancestor with Lagomorpha than with Rodentia. The closest to lagomorphs fossil outgroups are mimotonids, probably a paraphyletic group of five genera (*Anatolimys*, *Gomphos*, *Mimolagus*, *Mimotona,* and *Mina*; Fostowicz-Frelik et al. 2015a and references herein; Li et al. 2016a).

The Mesozoic ancestry of Glires has been connected with the concept of Anagalida, originally proposed by Szalay and McKenna (1971). The Anagalida (McKenna and Bell 1997) was considered a clade that includes Zalambdalestidae, a specialized group of late Cretaceous Eutheria, Anagalidae and Pseudictopidae, both exclusively Paleogene families endemic to Asia, Macroscelidae (elephant shrews), an endemic African group of small insectivorous mammals, and Glires; macroscelids are now considered Afrotheria (Murphy et al. 2007: Fig. 6). Archibald et al. (2001) analyzed the affinities of zalambdalestids and found that *Barunlestes* is an outgroup closest to Glires. Although the closer relationships of zalambdalestids to the Cenozoic placentals were put in doubt on the basis of some cranial characters (e.g., Wible et al. 2004), the morphological characteristics of two other families (Anagalidae and Pseudictopidae; Hu 1993 and Sulimski 1968, respectively) point to their closer affinity with Euarchontoglires (Glires + Archonta) than with any other mammalian group (Meng et al. 2003).

Basal Glires, in general, and Duplicidentata, in particular, are quite uniform in the dental and skeletal morphology (Fostowicz-Frelik and Meng 2013), which in closely related lineages may be attributed to parallel evolution. On the other hand, Glires share general adaptations and mode of life with other small herbivores, showing convergent adaptations on small ungulates (Fostowicz-Frelik et al. 2015a), and Anagalidae and Pseudictopidae (Fig. 1).

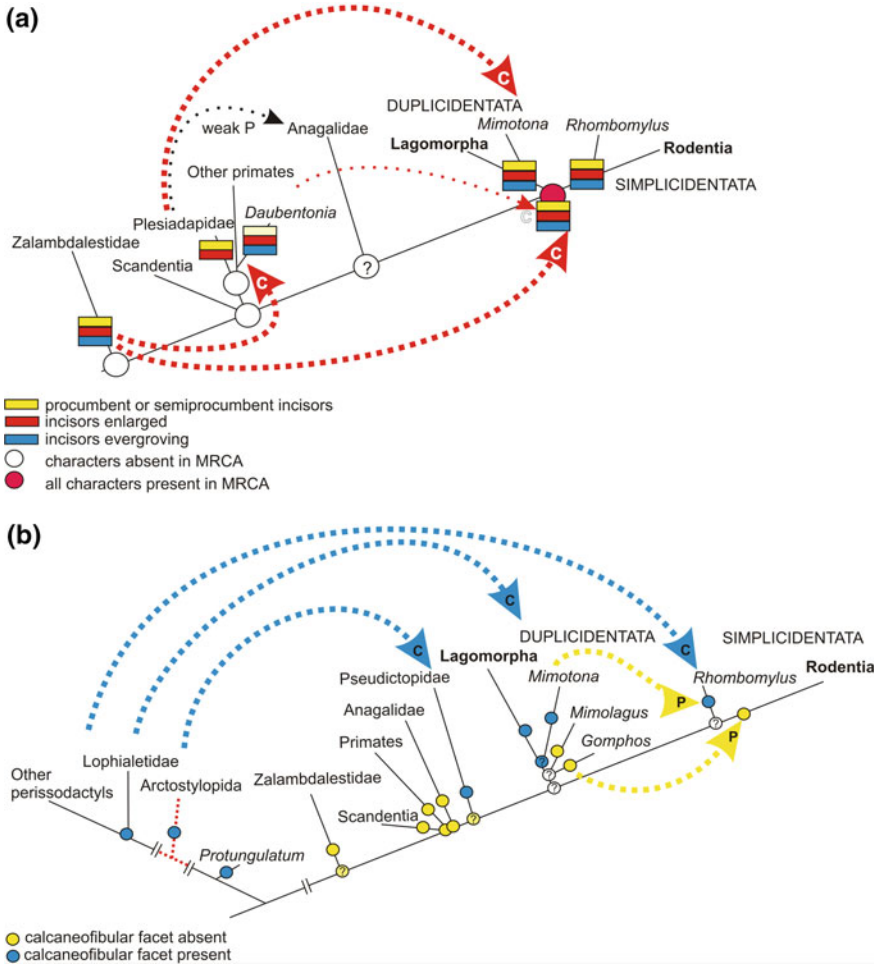


Fig. 1 Simplified phylogenetic diagram showing relationships between the discussed groups and marking the parallel and convergent characters in incisor structure and tarsal bones. **a** Distribution of enlarged and procumbent/semiprocumbent incisors within Euarchontoglires and Zalambdalestidae. **b** Distribution of the calcaneofibular facet within Glires and its convergent appearance in Arctostylopida and ungulates (primitive perissodactyls). Modified from Archibald et al. (2001), Meng et al. (2003), and Asher et al. (2005). Abbreviations: *C* convergence; *P* parallel evolution; *MRCA* most recent common ancestor

In the fossil record, especially in groups with no extant representatives (e.g., Anagalidae or Zalambdalestidae), we do not have enough information on underlying developmental mechanisms, and thus, we are unable to determine whether a homoplastic trait is due to convergent or parallel evolution. Hence, I accept the operational definition of convergence and parallelism, which links them only to topology of phylogenetic tree, that is to say that convergent and parallel evolution is

characterized by two distinct patterns of character state changes in phylogenetic tree (see Pearce 2012: Fig. 1 for more details). Consequently, a homoplastic trait that evolved by parallel evolution will be expressed in the closely related lineages. Their immediate common ancestor, however, did not itself show a changed character state. The convergence of structures arises when a homoplastic trait occurs in lineages considerably distant phylogenetically, and frequently, the trait in question evolved from different ancestral character set in each of the lineages.

2 The 'Lagomorph Heel'

One of the hallmarks of lagomorph skeleton is a specialized structure of the crurotarsal joint (CTJ) displaying a well-developed calcaneofibular articulation (Ca-Fi; Fig. 2b). In most mammalian groups, except ungulates (Fig. 2c, f) and elephant shrews, the only point of contact between the bones of the foot and the shank is the connection between the tibia and astragalar trochlea (Szalay 1985). The astragalus then lies on the calcaneus dorsally (Fig. 2d), connected by three articulation facets of the calcaneoastragalar joint (the anterior, medial, and posterior facets; the anterior one being most often the smallest; see Fostowicz-Frelik 2007: Fig. 12), and transmits animal's weight to the foot. In elephant shrews, lagomorphs (Fig. 2b, e), ungulates (Fig. 2c, f), and *Pseudictops* (Fig. 3h, i), an additional area of support is created, the calcaneofibular facet (facies articularis tibialis sensu Fostowicz-Frelik 2007), which is an immediate contact between the fibula or fibular part of the distal extremity of the tibiofibular bone (the tibia and fibula are fused in lagomorphs) and the calcaneus (Fostowicz-Frelik 2007: Fig. 12). This additional articulation surface is clearly a convergence between ungulates, elephant shrews, pseudictopids, and lagomorphs, because it evolved independently from primitive morphotype and these lineages are not closely related.

Although, strictly speaking, the calcaneofibular facet is not a synapomorphy of duplicidentates, it is shared by all lagomorphs of a modern aspect (Szalay 1985; Fostowicz-Frelik 2007; Li et al. 2007) and *Mimotona* (Szalay 1985; Li and Ting 1985; Zhang et al. 2016). This facet is also present in some genera of eurymylids (e.g., *Rhombomylus*).

The calcaneofibular articulation in mammals is related to the stabilization of the ankle joint. This articulation considerably widens the calcaneal eminence and creates the area of support for both bones (not only the tibia, but also the fibula) of the shank. Such an arrangement of the bones in the crurotarsal joint prevents from the harmful torsions and facilitates effective and powerful movements of the foot in the parasagittal plane. Thus, the calcaneofibular joint is an adaptation which increases cursorial ability and facilitates a leaping mode of locomotion (but not ricochetal, which is strictly bipedal). Modern lagomorphs, especially hares and rabbits (Leporidae), employ a 'leaping gallop' (Fostowicz-Frelik 2007 and

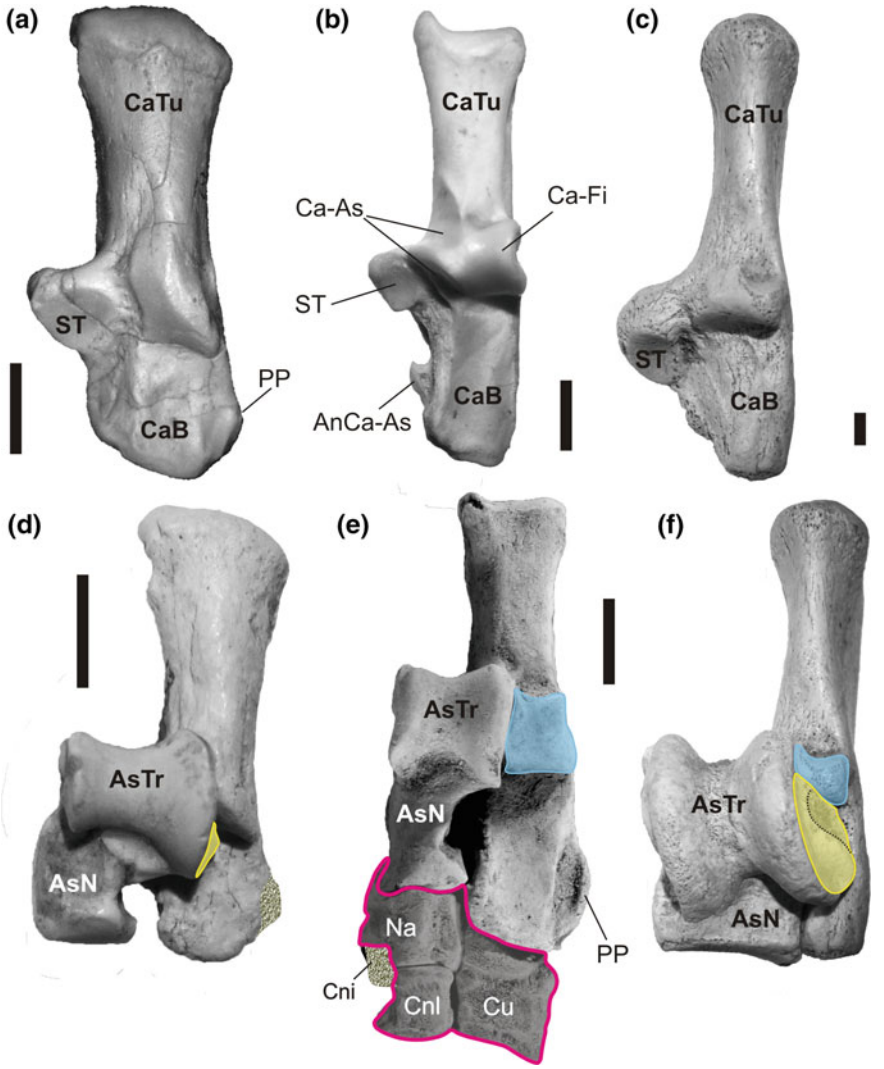


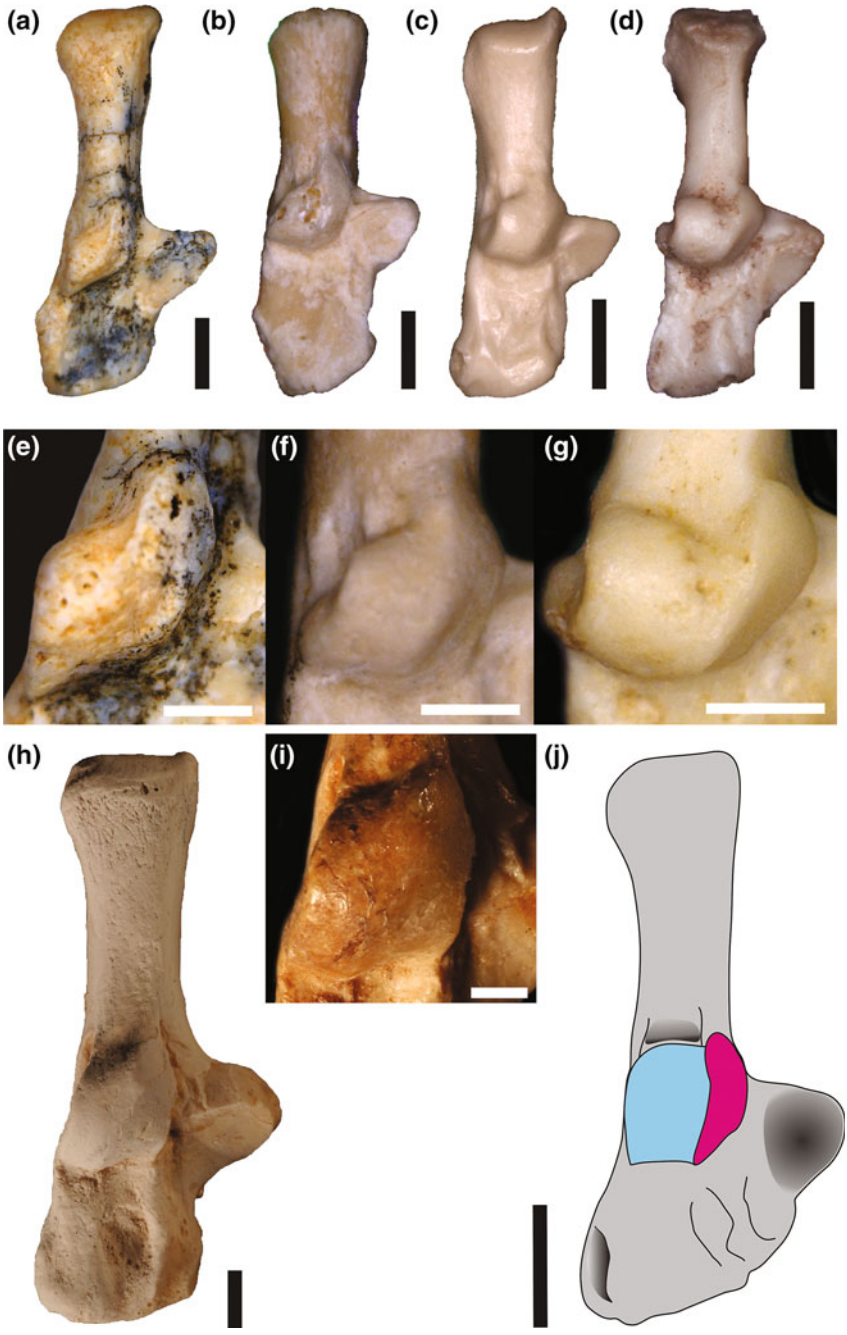
Fig. 2 Crurotarsal articulation in lagomorphs, *Gomphos*, and tapiroid perissodactyl. **a** Left calcaneus of *Gomphos elkema* (coll. IVPP), early Eocene, Inner Mongolia, China. **b** Left calcaneus of extant leporid, *Lepus* sp. (coll. IVPP). **c** Left calcaneus of tapiroid perissodactyl *Lophialetes* sp. (coll. IVPP), middle Eocene, Inner Mongolia, China. **d** Left calcaneus and astragalus of *Gomphos* sp. (coll. IVPP) in articulation, early Eocene, Inner Mongolia, China. **e** Reconstructed left tarsus of *Hypolagus beremendensis* (coll. Institute of Systematic and Evolution of Animals, Polish Academy of Sciences, Kraków, Poland), early Pliocene, Poland. **f** Left calcaneus and astragalus of *Lophialetes* sp. (coll. IVPP) in articulation, middle Eocene, Inner Mongolia, China. All figures in dorsal view; posterior direction to the top. Abbreviations: *AnCa-As* anterior calcaneoastragalar facet; *AsN* astragalar neck; *AsTr* astragalar trochlea; *Ca-As* calcaneoastragalar facet; *CaB* calcaneal body; *Ca-Fi* calcaneofibular facet; *Ca-Tu* calcaneal tuber; *Cni* intermedial cuneiform bone; *Cnl* lateral cuneiform bone; *Cu* cuboid; *Na* navicular; *PP* peroneal process; *ST* sustentaculum tali; *IVPP* Institute of Vertebrate Paleontology and Paleoanthropology, Chinese Academy of Sciences, Beijing, China. Calcaneofibular facet marked in *blue* (**e**, **f**), astragalofibular facet marked in *yellow* (**d**, **f**). *Scale bars* equal 5 mm

references therein), which differs from a typical gallop by the extension of the airborne phase, when all four feet are detached from the ground. The length of the air-suspended phase depends on several biomechanical factors, such as the angle of departure and the overall muscle strength and capacity, and is reflected in the length of the limb bones and the shape and proportions of the calcaneus itself, e.g., the elongation of the calcaneal body, which facilitates jumping in general (Fostowicz-Frelik 2007).

2.1 Paleogene Glires

In the fossils, the calcaneofibular facet is present among the earliest known Glires, from the Paleocene deposits of Qianshan (Anhui Province, China; Li 1977). A single calcaneus (IVPP V7422; Fig. 3a) with the calcaneofibular facet was described from the same beds as *Heomys* and *Mimotona* (Li and Ting 1993). It was attributed to either *Heomys* sp. (Li and Ting 1993: Fig. 11.9) or indeterminate Glires (Zhang et al. 2016: Fig. 1). In fact, its duplicidentate provenance (i.e., that of *Mimotona*) is much more plausible, because this particular lineage of Glires (except for large-sized genera *Gomphos* and *Mimolagus*) is characterized by the calcaneofibular articulation, while this feature is untypical of Simplicidentata and thus of *Heomys*. The bone from Qianshan shows a rather mixed morphology between the typical lagomorph calcaneus and that of early rodents (paramyids or ctenodactyloids). It has a rather long and slender calcaneal tuber and a relatively short calcaneal body, which suggests a rather poor jumping ability. The peroneal process is strong (Fig. 3a), although it is not as prominent as in rodents. Additionally, the sustentaculum tali is aligned with the calcaneal eminence (medial to it), which is typical of the calcanei of lagomorphs and other duplicidentate Glires, including *Gomphos* (Meng et al. 2004; Fig. 2a) and *Mimolagus* (Fostowicz-Frelik 2015a).

The calcaneal eminence of Qianshan specimen is long, which is typical of Rodentiaformes such as *Tribosphenomys*, and rodents (see Meng and Wyss 2001), but not of lagomorphs of a modern aspect, where it is shortened (Figs. 2b, 3b, c). Similarly, the calcaneoastragalar facet is larger and dominates the calcaneal eminence, forming a large and gentle semilunar articulation area with an eminent lateral edge crossing the dorsal surface of the eminence diagonally (in anterolateral to posteromedial direction). The calcaneofibular facet is smaller and rhomboidal in dorsal view, with the anterior margin reduced almost to a point; it lies diagonally at the eminence. In lagomorphs of the modern aspect and some eurymylids, it is generally as wide anteriorly as in its mid-length. The torsion and diagonal position of the calcaneoastragalar facet is more weakly expressed in *Dawsonolagus* (Fig. 3b, f), the earliest currently known lagomorph of a modern aspect (Li et al. 2007), or in the middle Eocene *Strenulagus* (see Fostowicz-Frelik et al. 2015b: Fig. 7a), although the torsion is still stronger (Fig. 3c) than in more derived taxa (Fig. 2b), where the longitudinal axes of both facets at the calcaneal eminence are closer to the parasagittal orientation (see Fostowicz-Frelik 2007: Fig. 12 for modern Leporidae).


(a)
(b)
(c)
(d)
(e)
(f)
(g)
(h)
(i)
(j)

◂**Fig. 3** Calcaneofibular articulation in Glires and closely related groups. **a** Right calcaneus from Qianshan (IVPP V7422; late Paleocene, Anhui, China). **b** Right calcaneus of *Dawsonolagus antiquus* (IVPP V7465.1) early/middle Eocene, Inner Mongolia, China. **c** Left calcaneus of *Strenulagus solaris* (IVPP 20218; mirror view), middle Eocene, Inner Mongolia, China. **d** Eurymylidae indet. (coll. IVPP), middle Eocene, Inner Mongolia, China. **e–g** Calcaneal eminence close-up in **a**, **b**, and **d**, respectively. **h** Right calcaneus of *Pseudictops lophiodon* (ZPAL MgM-II/48, Institute of Paleobiology, Polish Academy of Sciences, Warsaw, Poland). **i** Calcaneal eminence in (**h**). **j** Left calcaneus of *Paleostylops* (mirror view) showing articulation facets at the calcaneal eminence: calcaneoastragalar facet (*red*) and calcaneofibular facet (*blue*); based on Missiaen et al. (2006). *Scale bars* equal 2 mm except **e–g** (1 mm)

Compared to Qianshan bone, the calcanei of lagomorphs of a modern aspect have a wider calcaneofibular facet, which in modern lagomorphs becomes the dominant facet of the eminence (Figs. 2b, 3b, c, f).

A slightly different morphology is observed in some eurymylids. The calcaneus of *Rhombomylus turpanensis* has a shorter and more bulbous calcaneal eminence (Meng et al. 2003: Fig. 66) than the bone from Qianshan or *Strenulagus*. The facets are only slightly bent, expressing weakly an additional transversal ridge, which is developed fully in leporids and some *Desmatolagus*. *R. turpanensis* has a relatively short calcaneal body, and the swelling of the calcaneofibular facet is not as strong as in an unidentified eurymylid from the middle Eocene of Inner Mongolia (China), which is even more similar to coeval *Strenulagus solaris* in the calcaneal body elongation and a wide calcaneofibular surface typical of more advanced lagomorphs (Fig. 3d, g). Overall, the calcanei of early lagomorphs and eurymylids are quite similar morphologically and indicate shared locomotor adaptations, most probably to running in extended leaps.

Among the Glires discussed above, the appearance of the calcaneofibular facet could be seen as a parallel character for *Mimotona* (see above), lagomorphs, and *Rhombomylus*. However, this hypothesis cannot be confirmed for eurymylids, as we do not know the ancestral structure of the tarsal joint in this lineage.

2.2 Gomphos—Neither Rodent nor Lagomorph

Gomphos is a large duplicidentate Glires; however, it is peculiar in having the crurotarsal joint (including the calcaneus morphology) typical of rodents (Fig. 2a, d; see also Meng et al. 2004).

The *Gomphos* lineage shows closer similarities (especially in the calcaneus structure) and possibly the affinity, to *Mimolagus*, the largest Paleogene duplicidentate (Bohlin 1951; Fostowicz-Frelik et al. 2015a). The calcanei of *Mimolagus* are slightly bigger, generally more robust (*M. aurorae*) or having a more elongated calcaneal tuber (*M. rodens*) than those of *Gomphos*. The peroneal process is strongly reduced in *Mimolagus* (as in lagomorphs), whereas in *Gomphos*, especially in the early Eocene *G. elkema* it is relatively bigger (Meng et al. 2004), although

still less significant than the peroneal process in Paramyidae (Szalay 1985) or Ctenodactylidae. Nevertheless, the calcanei of *Gomphos* and *Mimolagus* are strikingly similar.

The presence of a typical 'rodent-like' calcaneus (displaying only the calcaneoastragalar facet) in *Gomphos* (and *Mimolagus*) is a character paralleling Rodentiaformes and rodents of the modern aspect. The only significant character differing calcanei of *Gomphos* and *Mimolagus* from those of rodents is a substantial dorsoplantar compression of the calcaneal tuber, which posterior extremity is much wider than high, unlike in most of rodent taxa. Interestingly, it may not be directly related to the size of animal, because the coypu (*Myocastor coypu*), a large amphibious rodent, has the calcaneal tuber almost isometric, with no significant compression.

2.3 *Pseudictops, a Basal Euarchontoglirid*

Pseudictops lophiodon is one of the best-known representatives of Pseudictopidae, an endemic Asian family of Paleocene origin (Sulimski 1968; Lucas 2001; Wang et al. 2007). The exact phylogenetic position of this group is uncertain. Most of phylogenetic analyses place *Pseudictops* with another endemic Paleogene Asian group, Anagalida, although the closest relationships of such a clade are vague, suggested either as a sister group to Glires (Meng and Wyss 2001; Meng et al. 2003) or even at the very base of the Euarchontoglires (Asher et al. 2005).

Despite its rather distant phylogenetic position, *Pseudictops* shares an important character with lagomorphs—a strikingly similar morphology of the calcaneus, displaying a well-developed calcaneofibular facet (Fig. 3h, i). The overall calcaneal morphology in *Pseudictops* is, however, more similar to that of basal Duplicidentata and earliest Lagomorpha (Fig. 3a–c, e, f) than to modern Leporidae (Fig. 2b). In fact, it resembles most the bone from Qianshan, also in general proportions and a relative elongation of the tuber calcanei, although it has a wider calcaneofibular facet. The articular surfaces at the calcaneal eminence are more rounded and the edges of the calcaneoastragalar facet are not distinct. Although the elongation of the calcaneal body is not significant and *Pseudictops* did not apparently display any enhanced jumping ability, its crurotarsal joint was strengthened and stabilized to movements mostly in parasagittal plane, which rules out arboreal mode of life and points to the terrestrial mode of locomotion as a moderately able cursor.

2.4 *Primitive Perissodactyls and 'Gliriform' Mammals*

In the Paleogene of Asia, also ungulates share the calcaneofibular articulation with duplicidentate Glires and *Pseudictops*, which is an obvious convergence.

The calcaneal morphology nearly identical to aforementioned groups can be found in Lophialetidae, a family of primitive perissodactyls (Fig. 2c, f), and Arctostylopida. The latter (known also from North America) are an enigmatic group of uncertain phylogenetic relationships, sometime associated with Notoungulata, but usually considered an independent order (see Cifelli et al. 1989; Zack 2004; Wang et al. 2008).

Lophialetidae, known in the fossil record since the early Eocene, were small tapiroids with body mass of about 2–7–10 kg (see Fostowicz-Frelik et al. 2015a). They match Lagomorpha in the overall architecture of the calcaneus and astragalus (Fig. 2). The calcaneus has a long and mediolaterally compressed calcaneal tuber and moderately elongated calcaneal body. The sustentaculum tali is slightly oriented anteriorly, and thus, it is not aligned transversely with the calcaneal eminence as in Glires. The bones of lophialetids differ from those of duplicidentate Glires, apart from being more strongly compressed mediolaterally, also in a significant shortening of the calcaneal eminence anteroposteriorly and a somewhat reduced calcaneofibular facet. The calcaneoastragalar facet is wider than long and dominates the eminence, while the calcaneofibular facet has triangular outline, tapering strongly anteriorly. It has a relatively deep tendon sulcus at its posterior edge, at the base of the eminence, which is deeper than that in Lagomorpha.

The calcaneus of Arctostylopida (Fig. 3j) is much more similar to that of pseudictopids and lagomorphs than to lophialetids (and thus, perissodactyls). This peculiar similarity prompted questions concerning the exact relationships of arctostylopids to Glires. The former were even sometimes termed ‘gliriform mammals’ (e.g., Missiaen et al. 2006). Overall, the calcaneus is slender, slightly compressed mediolaterally, with an elongated calcaneal tuber and moderately long calcaneal body with a poorly developed peroneal process. The reduction in the peroneal process is characteristic of most duplicidentates, including *Gomphos* (which still may have quite a prominent process) and *Mimolagus*, which shows a significant reduction in this structure (Fostowicz-Frelik et al. 2015a). The sustentaculum tali is positioned medially and in line with the calcaneal eminence. The sustentacular shelf has a more acute medial edge than in lagomorphs and *Pseudictops*, which resembles in that respect the bone from Qianshan. The latter, however, has a whole shelf directed more posteriorly (see Fig. 3a, j). The calcaneal eminence is relatively large and not as much compressed anteroposteriorly as in leporids; therefore, it is not so bulbous, but still eminent, similar to that of *Pseudictops*. The calcaneoastragalar facet is wide, oval-shaped, slightly tightening in the mid-length, and in dorsal view, it widens anteriorly. However, the calcaneofibular facet maintains its width along the whole length and is roughly rectangular in dorsal view.

Morphology of the calcaneus in Arctostylopida is clearly convergent on that of *Pseudictops* and lagomorphs, indicating a very similar mode of locomotion. Stratigraphically, arctostylopids and pseudictopids precede lagomorphs of a modern aspect, but their extinction had not been caused by the immediate competition with lagomorphs, as the latter were much smaller at the time of their origin and clearly occupied different ecological niches. Rather, the competition with large

mimotonids, particularly with *Gomphos* of similar size, may have exerted more pressure on food resources. *Gomphos* coexisted with arctostylopids in Inner Mongolia during the early Eocene (Bumbanian), although they were a very rare faunal element there, while *Gomphos* was quite abundant.

3 Dental Characters in Early Glires

One of the dental hallmarks of Glires is ever-growing semiprocumbent incisors (Fig. 4c, d). Simplicidentata (Rodentia and Eurymylidae) have only one pair of the upper and lower incisors, while duplicidentates show two upper, and one (Lagomorpha) or two (mimotonids) lower pairs of ever-growing incisors, of which the first pair is enlarged. The exact homology of these teeth has been a subject of long debate (see Meng and Wyss 2001) to the effect that the 'first pair' of incisors in Duplicidentata, as indicated by the embryological studies in extant representatives (Moss-Salentijn 1978; Ooë 1980; Simoens et al. 1995), are in fact deciduous teeth of the second pair (DI2/di2).

The enlarged and procumbent (to a different extent) incisors of the first pair (and sometimes also the second one) occur in several groups of Euarchontoglires, including a prosimian *Daubentonia* (aye-aye), plesiadapids, and some anagalids. Moreover, both characters are well expressed in Zalambdalestidae (Fig. 4a, b), a group of Cretaceous stem placentals, which display a large, ever-growing, procumbent lower incisor, of which the open-rooted portion is running ventrally beneath the tooth row, along most of the mandible body (Fostowicz-Frelik and Kielan-Jaworowska 2002; Fostowicz-Frelik 2016). There are no data indicating to which generation the anterior lower incisors in zalambdalestids belong; they are considered the permanent first incisor pair rather than the second pair of deciduous teeth. Nevertheless, the morphology of these incisors is very similar to that of Glires. They are not as strongly curved as in Glires, but their position within the mandible is the same (Fig. 4). Also, the distribution of the enamel layer on the tooth is similar, being restricted mostly to the ventral and lateral tooth wall. The exact extent of the enamel varies among the taxa and covers more of the tooth perimeter than in Glires, but overall the similarity is suggestive of the possible, although distant, phylogenetic relationships between these groups (Fostowicz-Frelik 2016). A different character set joins the enlarged incisors of plesiadapids and mimotonids. The incisors in the latter are open-rooted and ever-growing (Li et al. 2016b), while in the former they have closed roots and determinate growth. Nevertheless, their shape and formation of an elongated bladelike cutting facet, clearly distinct from the incisor shaft, evolved in convergence in these two groups (Fig. 1a), as both groups are too distantly related to consider parallel evolution.

Regardless of exact homologies and detailed structural peculiarities, the multiple instances of the incisor enlargement in Euarchontoglires indicate a certain genetic

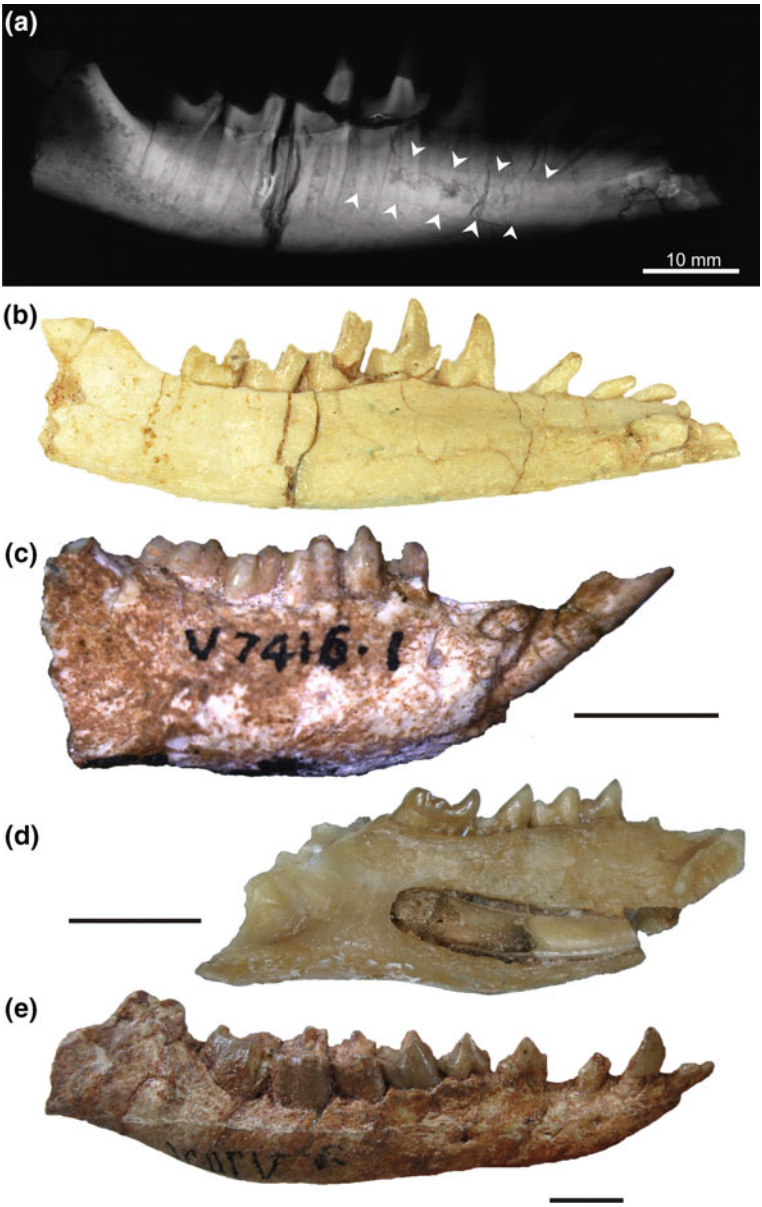


Fig. 4 Lower incisor structure among Glires, Anagalidae, and Zalambdalestidae. **a**–**b** Enlarged and procumbent ever-growing incisor in right mandible of *Zalambdalestes lechei* (ZPAL MgM-I/43), late Cretaceous, Mongolia (*white arrows* show course of open root in X-ray image, **a** from Fostowicz-Frelik 2016). **c** Right mandible with complete dentition of *Mimotona wana* (IVPP 7416.1), Paleocene, Qianshan, China, note semiprocumbent lower incisors typical of Glires. **d** Course of open root of lower incisor in right mandible of *Eurymylus laticeps* (ZPAL MgM-II/61; mirror view), Paleocene, Mongolia. **e** Left mandible of *Qipania yui* (IVPP V07426; mirror view), Paleocene, Anhui, China; note procumbent incisors in Anagalidae. *Scale bars* equal 5 mm (**c**–**e**)

component susceptible to mutations, which demonstrates the anterior dentition plasticity in the whole group.

In the terms of parallel evolution and convergence, the enlarged lower anterior incisors in Zalambdalestidae are convergent on the incisor in plesiadapids, *Daubentonia*, and Glires. Similarity between the first lower incisor in plesiadapids and that of *Zalambdalestes* may be a true homology.

In the upper and lower premolar and molar dentition of most basal Glires (Eurymylidae and Mimotonidae), as well as in Lagomorpha of a modern aspect, one feature is evident, namely dental hypsodonty. The heightening of the tooth crowns occurs very early in the lagomorph evolution, and thus, already in the Eocene the forms showing a complete lack of the buccal roots and extended crown appear (see Fostowicz-Frelik 2013). However, a unilateral (partial) hypsodonty typical of the stem forms (Fig. 5) is quite frequent among Lagomorpha until the late Miocene (Tobien 1974).

The teeth of basal Glires are relatively weakly hypsodont, and the height of the crown at the lingual side is up to three times higher than that at the buccal side (Dashzeveg and Russell 1988), whereas the teeth of the first lagomorph of a modern aspect (*Dawsonolagus*) are relatively much higher (Fig. 5). However, the heightening of the crowns of premolars and molar occurs in all lineages of duplicidentate Glires and indicates parallel evolution not only between eurymylids and mimotonids, and lagomorphs of a modern aspect, but also between some Eocene groups of rodents (e.g., Eomyidae). Apart from Glires, the dental hypsodonty is also well developed in a closely related lineage, Anagalidae (Bohlin 1951; McKenna 1963;

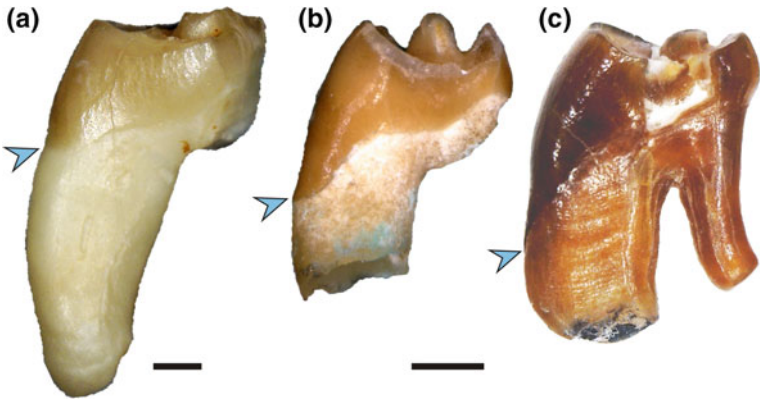


Fig. 5 Premolars of basal Glires and early lagomorphs of a modern aspect showing unilateral hypsodonty. **a** Right P3 of *Gomphos elkema* (coll. IVPP; mirror view), early Eocene, Inner Mongolia, China. **b** Left P4 of *Dawsonolagus antiquus* (IVPP V7499.3), early/middle Eocene, Inner Mongolia, China. **c** Left P4 of *Strenulagus solaris* (IVPP 20192.1), middle Eocene, Inner Mongolia, China. *Arrows* indicate crown base at the lingual side. *Scale bars* equal 1 mm

Hu 1993), most probably one of basal clade of Euarchontoglires. Not only have Anagalids well developed dental hypsodonty (Fig. 6b, c), but their crown height is frequently higher than in coeval Mimotonidae (see Li et al. 2016b). The teeth of Anagalidae share many apparently convergent features with the dentition of lagomorphs and advanced mimotonids (e.g., *Gomphos* and *Mimolagus*; Fig. 6d). The upper cheek teeth of anagalids show relatively broad occlusal surfaces, which quickly in ontogeny become worn, leaving gently concave, mostly smooth occlusal surfaces (Fig. 6c). Such morphology resembles closely that of the dentition of *Mimolagus* (Bohlin 1951; Fostowicz-Frelik et al. 2015a; Fig. 6d) and most probably has an adaptive significance related to diet. *Mimolagus*, which has exceptionally among basal duplicidentates worn teeth (Fig. 6d), was interpreted as a bulk feeder and compared to small coeval tapiroids, which matched it in size. Anagalids with their less specialized (omnivore-like) dental formula (including canines and three incisor loci) are unlike to share such type of feeding adaptations, although their molar structure may suggest a herbivorous diet.

An interesting parallel feature in the duplicidentate dentition is the presence and formation of so-called lingual enamel bridges on the lower molars. The enamel bridge is a narrow enamel lamella joining the trigonid with talonid at the buccal side of the tooth. Generally, it has been regarded as a character discriminating Leporidae (which form bridges) from Ochotonidae (which do not). This distinction holds well for the crown lagomorphs, but fails for stem taxa. In most early Paleogene stem lagomorphs, the enamel bridges are formed late in ontogeny and mostly are visible in older individuals, which show relatively worn teeth, and thus, the bridges may be overlooked. In some of late-Oligocene to early-Miocene North American species of *Palaeolagus* lineage (*P. burkei*, *P. hypsodus*, and *P. subhypsodus*), which are regarded stem lagomorphs, the enamel bridges are never formed. Thus, the lower premolars and molars (p4–m2) in these species have structures parallel to those of crown ochotonids (e.g., †*Sinolagomys* or *Ochotona*).

Finally, rapid advances in experimental biology of mammalian dentition have been applied recently also to extinct species, thus making possible the determination of developmental mechanisms lying under the morphological characters. Harjunmaa et al. (2014) reported in vivo experiments that enabled them to reproduce characters on murine molars strikingly reminiscent of those of a Paleocene rodentiaform *Tribosphenomys*, thus effectively reverting some dental characters in a very derived rodent to their purportedly basal states. Further, Tapaltsyan et al. (2015) studied evolution of hypselodonty (full hypsodonty) in the molar teeth of fossil North American rodents in the time frame of 48 million years (up to 2 Mya, the Pleistocene). They were able to simulate computationally that the increase in hypsodonty resulted rather from quantitative continuous changes than qualitative steps. Such studies give us hope that we will be eventually able to reconcile topological notions of convergence and parallelism in the fossil record with developmental approach of neontology, not only for Glires.

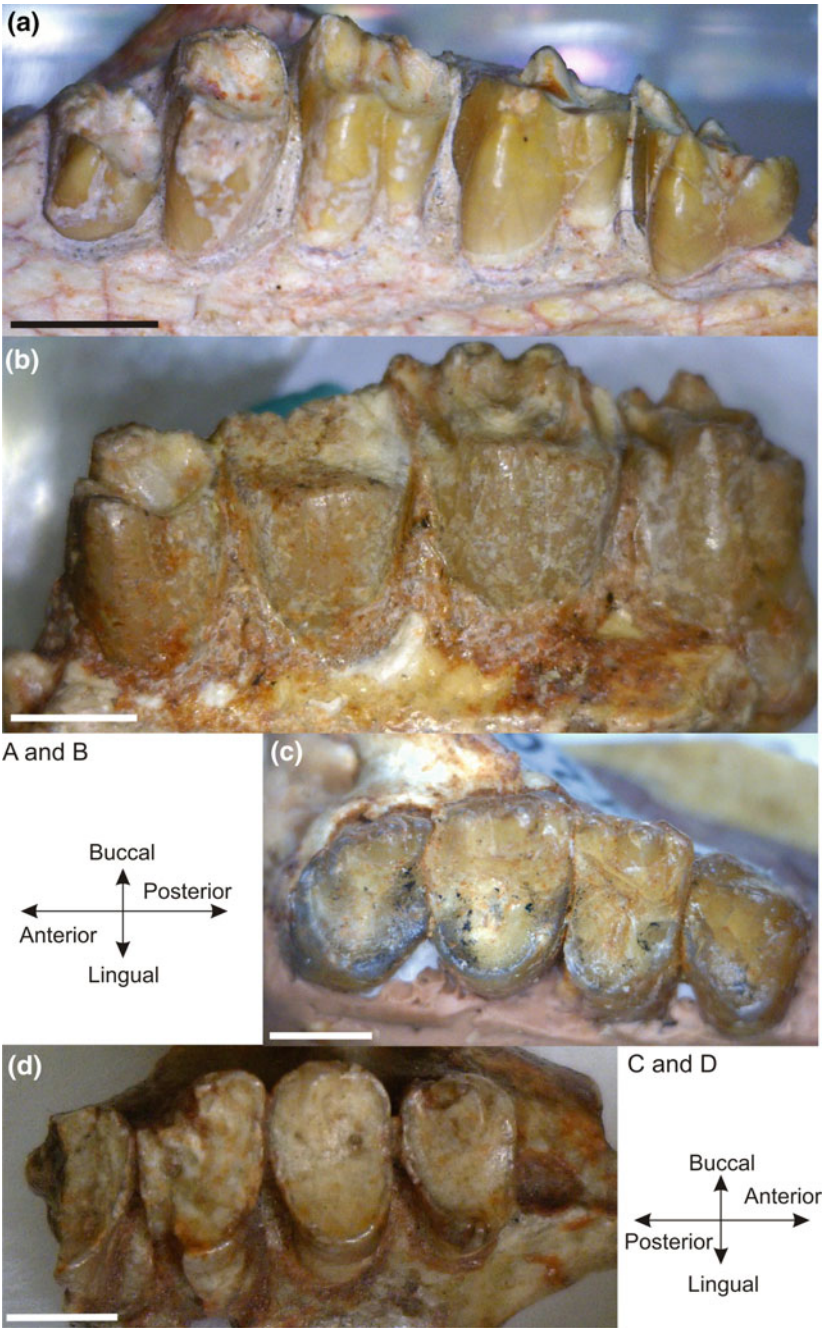


Fig. 6 Convergent occurrences of hypsodonty in dentition of basal Glires and Anagalidae. **a** Eurymylid *Rhombomylus laianensis*, left maxilla with P3–M3 (IVPP V7428), early Eocene, Hubei, China. **b** Anagalid *Hsiuannania tabiensis*, right maxilla with P4–M3 (IVPP V4274; mirror view), Paleocene, Anhui, China. **c** Anagalid *Qipania yui*, right maxilla with P4–M3 (IVPP V07426), Paleocene, Anhui, China. **d** Mimotonid *Mimolagus rodens*, right maxilla with P3–M2 (IVPP RV51002.1), late Eocene/Oligocene, Gansu, China. *Scale bars* equal 3 mm

Acknowledgements I am greatly indebted to Li Chuan-Kui, Ni Xijun, Wang Yuan-Qing, Li Qian (all Institute of Vertebrate Paleontology and Paleoanthropology CAS, Beijing, China), and Jin Meng (AMNH, New York, USA) for providing access to the specimens in their care and helpful discussions. The present work was supported by the National Science Centre (Cracow, Poland) grant No. 2015/18/E/NZ8/00637. I acknowledge insightful review by an anonymous reviewer, which improved this paper.

References

Archibald JD, Averianov AO, Ekdale EG (2001) Late cretaceous relatives of rabbits, rodents, and other extant eutherian mammals. Nature 414:62–65

Asher RJ (2007) A web-database of mammalian morphology and a reanalysis of placental phylogeny. BMC Evol Biol 7 (108)

Asher RJ, Meng J, Wible JR, McKenna MC, Rougier GW et al (2005) Stem Lagomorpha and the antiquity of Glires. Science 307:1091–1094

Bohlin B (1951) Some mammalian remains from Shih-ehr-ma-ch'eng, Hui-hui-p'u area, Western Kansu. Reports from the Scientific Expedition to the North-Western Provinces of China under Leadership of Dr. Sven Hedin. Sino-Swed Exp Publ 35, VI. Vert Paleont 5:1–48

Cifelli RL, Schaff CR, McKenna MC (1989) The relationships of the Arctostylopidae (Mammalia): new data and interpretation. Bull Mus Comp Zool 152:1–44

Dashzeveg D, Russell (1988) Palaeocene and Eocene Mixodontia (Glires) of Mongolia and China. Palaeontology 31: 129–164

Fostowicz-Frelik Ł (2007) The hind limb skeleton and cursorial adaptations of the Plio-Pleistocene rabbit *Hypolagus beremendensis*. Acta Pal Pol 52:447–476

Fostowicz-Frelik Ł (2013) Reassessment of *Chadrolagus* and *Litolagus* (Mammalia: Lagomorpha) and a new genus of North American Eocene Lagomorpha from Wyoming. Am Mus Novit 3773:1–76

Fostowicz-Frelik Ł (2016) A new zalambdalestid (Eutheria) from the Late Cretaceous of Mongolia and its implications for the origin of Glires. Paleont Pol 67:127–136

Fostowicz-Frelik Ł, Meng J (2013) Comparative morphology of premolar foramen in lagomorphs (Mammalia: Glires) and its functional and phylogenetic implications. PLoS ONE 8(11):e79794

Fostowicz-Frelik Ł, Kielan-Jaworowska Z (2002) Lower incisor in zalambdalestid mammals (Eutheria) and its phylogenetic implications. Acta Pal Pol 47:177–180

Fostowicz-Frelik Ł, Li CK, Mao FY, Meng J, Wang YQ (2015a) A large mimotonid from the Middle Eocene of China sheds light on the evolution of lagomorphs and their kin. Sci Rep 5(9394):1–9

Fostowicz-Frelik Ł, Li CK, Li Q, Meng J, Wang YQ (2015b) *Strenulagus* (Mammalia: Lagomorpha) from the Middle Eocene Irdin Manha Formation of the Erlian Basin, Nei Mongol, China. Acta Geol Sin (Eng Ed) 89:12–26

Harjunmaa E, Seidel K, Häkkinen T, Renvoisé E, Corfe IJ et al (2014) Replaying evolutionary transitions from the fossil dental record. Nature 512:44–48

Hu (1993) Two new genera of Anagalidae (Anagalida, Mammalia) from the Paleocene of Qianshan, Anhui and the phylogeny of anagalids. Vert PalAsiat 31: 153–182

Li CK (1977) Paleocene eurymyloids (Anagalida, Mammalia) of Qianshan, Anhui. Vert PalAsiat 15:103–118

Li CK, Ting SY (1985) Possible phylogenetic relationships of eurymylids and rodents, with comments on mimotonids. In: Luckett WP, Hartenberger JL (eds) Evolutionary relationships among rodents: a multidisciplinary analysis. Plenum, New York, pp 35–58

Li CK, Ting SY (1993) New cranial and postcranial evidence for the affinities of the eurymylids (Rodentia) and mimotonids (Lagomorpha). In: Szalay FS, Novacek MJ, McKenna MC (eds) Mammal phylogeny–placentals. Springer, Berlin, pp 151–158

Li CK, Meng J, Wang YQ (2007) *Dawsonolagus antiquus*, a primitive lagomorph from the Eocene Arshanto Formation, Nei Mongol, China. Bull Carnegie Mus 39:97–110

Li CK, Wang YQ, Zhang ZQ, Mao FY, Meng J (2016a) A new mimotonidan mammal *Mina hui* (Mammalia, Glires) from the Middle Paleocene of Qianshan, Anhui Province, China. Vert PalAsiat 54(2):121–136

Li Q, Wang YQ, Fostowicz-Frelik Ł (2016b) Small mammal fauna from Wulanhuxiu (Nei Mongol, China) implies faunal turnover across the Irdinmanhan-Sharamurunian boundary. Acta Pal Pol 61:759–776

Lucas SG (2001) Chinese fossil vertebrates. Columbia University Press, New York

McKenna MC (1963) New evidence against tupaioid affinities of the mammalian family Anagalidae. Am Mus Novit 2158:1–16

McKenna MC, Bell SK (1997) Classification of mammals above the species level. Columbia University Press, New York

Meng J, Wyss AR (2001) The morphology of *Tribosphenomys* (Rodentiaformes, Mammalia): phylogenetic implications for basal Glires. J Mam Evol 8:1–71

Meng J, Wyss AR (2005) Glires. In: Rose KD, Archibald JD (eds) The rise of placental mammals; origins, timing, and relationships of the major extant clades. Johns Hopkins University Press, Baltimore, pp 145–158

Meng J, Hu YM, Li CK (2003) The osteology of *Rhombomylus* (Mammalia, Glires): implications for phylogeny and evolution of Glires. Bull Am Mus Nat Hist 275:1–247

Meng J, Bowen GJ, Ye J, Koch PL, Ting SY, Li Q, Jin X (2004) *Gomphos elkema* (Glires, Mammalia) from the Erlian Basin: evidence for the early Tertiary Bumbanian Land Mammal Age in Nei-Mongol, China. Am Mus Novit 3425:1–24

Missiaen P, Smith T, Guo DY, Bloch JI, Gingerich PD (2006) Asian gliriform origin for arctostylopid mammals. Naturwissenschaften 93:407–411

Moss-Salentijn L (1978) 2. Vestigial teeth in the rabbit, rat and mouse; their relationship to the problem of lacteal dentitions. In: Butler PM, Joysey KA (eds) Development, function and evolution of teeth. Academic Press, London, pp 13–29

Murphy WJ, Eizirik E, O'Brien SJ, Madsen O, Scally M et al (2001) Resolution of the early placental mammal radiation using Bayesian phylogenetics. Science 294:2348–2351

Murphy WJ, Pringle TH, Crider TA, Springer MS, Miller W (2007) Using genomic data to unravel the root of the placental mammal phylogeny. Genome Res 17:413–421

Ooë T (1980) Développement embryonnaire des incisives chez le lapin (*Oryctolagus cuniculus* L.). Interprétation de la formule dentaire. Mammalia 44(2):259–269

Pearce T (2012) Convergence and parallelism in evolution: a neo-Gouldian account. Brit J Phil Sci 63:429–448

Simoens P, Lauwers H, Verraes W, Huysseune A (1995) On the homology of the incisor teeth in the rabbit (*Oryctolagus cuniculus*). Belg J Zool 125(2):315–327

Springer MS, Murphy WJ, Eizirik E, O'Brien SJ (2005) Molecular evidence for major placental clades. In: Rose KD, Archibald JD (eds) The rise of placental mammals; origins, timing, and relationships of the major extant clades. Johns Hopkins University Press, Baltimore, pp 37–49

Sulimski A (1968) Paleocene genus *Pseudictops* Matthew, Granger and Simpson 1929 (Mammalia) and its revision. Palaeont Pol 19:101–129

Sych L (1971) Mixodontia, a new order of mammals from the Paleocene of Mongolia. Pal Pol 25:147–158

Szalay FS (1985) Rodent and lagomorph morphotype adaptations, origins, and relationships: some postcranial attributes analyzed. In: Luckett WP, Hartenberger JL (eds) Evolutionary relationships among rodents: a multidisciplinary analysis. Plenum Press, New York, pp 63–132

Szalay FS, McKenna MC (1971) Beginning of the age of mammals in Asia: the late Paleocene Gashato fauna, Mongolia. Bull Am Mus Nat Hist 144:269–318

Tapaltsyan V, Eronen JT, Lawing AM, Sharir A, Janis C et al (2015) Continuously growing rodent molars result from a predictable quantitative evolutionary change over 50 million years. Cell Rep 11:1–8

Ting SY, Meng J, McKenna MC, Li CK (2002) The osteology of *Matutinia* (Simplicidentata, Mammalia) and its relationship to *Rhombomylus*. Am Mus Novit 3371:1–33

Tobien H (1974) Zur Gebiβstruktur, Systematik und Evolution der Genera *Amphilagus* und *Titanomys* (Lagomorpha, Mammalia) aus einigen Vorkommen im jungeren Tertiar Mittel- und Westeuropas. Mainzer geowiss Mitt 3:95–214

Wang YQ, Meng J, Ni X, Li CK (2007) Major events of Paleogene mammal radiation in China. Geol J 42:415–430

Wang YQ, Meng J, Ni XJ, Beard KC (2008) A new early Eocene Arctostylopid (Arctostylopida, Mammalia) from the Earlian Basin, Nei Mongol (Inner Mongolia), China. J Vert Pal 28:553–558

Wible J, Novacek MJ, Rougier GW (2004) New data on the skull and dentition in the Mongolian late Cretaceous eutherian mammal *Zalambdalestes*. Bull Am Mus Nat Hist 281:1–144

Wyss A, Meng J (1996) Application of phylogenetic taxonomy to poorly resolved crown clades: a stem-modified node-based definition of Rodentia. Syst Biol 45:559–568

Zack SP (2004) An Early Eocene arctostylopid (Mammalia: Arctostylopida) from the Green River Basin, Wyoming. J Vert Pal 24:498–501

Zhang ZQ, Li CK, Wang J, WangYQ Meng J (2016) Presence of the calcaneal canal in basal Glires. Vert PalAsiat 54(3):235–242

Reductive Evolution of Apicomplexan Parasites from Phototrophic Ancestors

Zoltán Füssy and Miroslav Oborník

Abstract Apicomplexans are widespread parasites of animals including humans with an interesting evolutionary history of trophic transitions from predation to photoautotrophy and later loss of photosynthesis. Comparison of extant phototrophic, predatory, and parasitic species revealed how engulfment of an alga constrained cellular biochemistry in the future parasites to a dependence on their non-photosynthetic plastid. Reconstructions of the common ancestor of Apicomplexa point out how complex this organism was as for metabolic repertoire, life cycle, and structural pre-adaptations. This ancestor was supposedly adapted to aerobic and anaerobic environments, predated on other eukaryotes using a flagellum-derived apical complex, and exhibited a complex life cycle to respond to sudden environmental changes. Rather than discovering entirely new features, therefore, apicomplexans arose mainly via reductive evolution of cellular structures and pathways existing in free-living ancestors.

1 Introduction

Apicomplexan parasites (Eukaryota; Alveolata; Apicomplexa) are microscopic single-celled protists known to live as obligate parasites of animals including humans. The most devastating human parasitosis, tropical malaria elicited by apicomplexan genus *Plasmodium* results in over 400,000 fatalities yearly. Further, toxoplasmosis caused by *Toxoplasma gondii* is less obviously harmful but highly prevalent even in developed countries (Lester 2012; Flegr 2013; Flegr et al. 2014), while yet other apicomplexan species are responsible for various diseases of domesticated and wild animals with high negative economic impact. Ultrastructural

Z. Füssy
Biology Centre CAS, Institute of Parasitology, Branišovská 31,
37005 České Budějovice, Czech Republic

M. Oborník
Faculty of Science, University of South Bohemia, Branišovská 31,
37005 České Budějovice, Czech Republic

P. Pontarotti (ed.), *Evolutionary Biology: Self/Nonself Evolution, Species and Complex Traits Evolution, Methods and Concepts*,
DOI 10.1007/978-3-319-61569-1_12

characteristics of the group include the apical complex, a set of tubular and secretory organelles at the anterior end of the cell used to penetrate cells of the host (Levine et al. 1980), and the apicoplast, a relic plastid found in most apicomplexans to have an essential metabolic role (Ralph et al. 2004).

Bearing a remnant plastid suggests that the ancestor of apicomplexans was capable of photosynthesis (McFadden and Waller 1997). However, apicomplexans are an ancient group, and direct comparison with living phototrophic relatives was not available at the time of apicoplast discovery. Initially, dinoflagellates as sister alveolates were investigated as the closest phototrophic relatives to Apicomplexa. However, both groups are extremely divergent. Dinoflagellates are known to possess a reduced plastid genome fragmented to minicircles with only 14 protein-coding genes, all associated with photosynthesis (Zhang et al. 1999). In contrast, the apicoplast genome lacks any traces of photosynthetic genes. The only genes present in both genomes are those coding for rRNAs, which are, however, also too divergent to be used for credible phylogenetic analysis. With virtually no overlap between the plastid genomes of dinoflagellates and apicomplexan, the comparison was impossible (Keeling 2008).

The discovery of chromerid algae, *Chromera velia* (Moore et al. 2008) and *Vitrella brassicaformis* (Oborník et al. 2012), filled a gap in our knowledge of early branching apicomplexans. *Chromera* and *Vitrella* contain a fully photosynthetic plastid and can be cultured without the need for an organic carbon source. They are not directly related (Janouškovec et al. 2015), but because they are both photoautotrophic, we refer to *Chromera* and *Vitrella* as the "chromerids," as opposed to closely related predatory "colpodellids." Chromerids were isolated from stony corals in Australia by an experimental procedure usually used to isolate intracellular symbionts, for instance, *Symbiodinium* dinoflagellates (Moore et al. 2008). A possible interaction of *C. velia* with corals (mutualistic, commensal, or facultatively parasitic) has not been firmly experimentally demonstrated, although cells of *C. velia* were observed inside the larvae of *Acropora* (Cumbo et al. 2013). A symbiotic tendency of *C. velia* can be relevant to the evolution to parasitism because close association with a multicellular organism may represent a starting point for the trophic transition to exploiting the host's metabolism (Lesser et al. 2013; Janouškovec and Keeling 2016). A mutualistic symbiont interacts with its partner in a fashion that is beneficial for both counterparts (such as metabolic compound exchange). The trade-off may not, however, pay back in long term, as one of the partners might start cheating and become harmful (parasitic). Such transition is currently taking place also in the *Symbiodinium* phylotype (Lesser et al. 2013). Pre-adaptations of free-living ancestors of chromerids and apicomplexans, such as the apical complex and thick-walled cysts, could have been repurposed for the benefit of the future parasite (Janouškovec and Keeling 2016).

Photoautotrophy/heterotrophy transitions are not scarce and can also be found in other lineages. The chlorophyte *Helicosporidium* spp. is a well-known example of algal parasite of insects (Weiser 1970; Tartar et al. 2002), and there are dozens of described rhodophytes parasitic on other closely or distantly related red algae,

taking advantage of intercellular pit connections developed in their kin hosts (recently reviewed by Salomaki and Lane 2014; Blouin and Lane 2015). Further, dinoflagellate peridinin plastid losses occurred several times independently (Saldarriaga et al. 2001; Gornik et al. 2015). As many as half of dinoflagellate species are heterotrophic or parasitic (Shields 1994, Nuismer and Otto 2004), including basal taxa *Perkinsus* and *Oxyrrhis* (Saldarriaga et al. 2003). In the heterotrophic stramenopile oomycetes, previous plastid presence is suspected but has not been unequivocally shown (Tyler et al. 2006). Several euglenophytes also lost photosynthesis, for instance, *Euglena longa* (Schoenborn 1949, 1952). A thorough inspection of the evolutionary history of eukaryotes, therefore, reveals many transitions from photoautotrophy to osmotrophy, predation, or parasitism.

Apicomplexan parasites represent algae that lost photosynthesis and adapted for obtaining nutrients from a multicellular host, similarly as other parasites do. Apicomplexans, therefore, passed through an astoundingly complicated evolutionary history involving fundamental switches of trophic modes, from primary heterotrophy, through photoautotrophy, to secondary heterotrophy and parasitism. As such they are excellent examples to study evolutionary processes linked to trophic transitions. Here we summarize how genetic and functional reduction and reusing of pre-adaptations, rather than the acquisition of new capabilities, drive the evolution to parasitism in apicomplexans.

2 Heterotrophy First

Apicomplexans show no obvious phylogenetic affiliation to Archaeplastida, the only major primary (cyanobacteria-to-eukaryote) phototrophic group known encompassing glaucophytes, red and green algae, and land plants. Consistently, four membranes enclose the apicoplast suggesting this organelle was acquired through an endosymbiotic relationship with a eukaryotic alga (McFadden and Waller 1997). The inner two membranes of this complex plastid arose from the (primary) plastid envelopes; the second outermost membrane supposedly derives from the cytoplasmic membrane of the endosymbiont, and the outermost membrane is probably host-derived. It is straightforward to assume that the ancestors of Apicomplexa were heterotrophic before obtaining their plastid via eukaryote-to-eukaryote endosymbiosis, but it is still debated which other eukaryotic groups shared this evolutionary event with apicomplexans. According to the chromalveolate hypothesis (Cavalier-Smith 1999), a rhodophyte-derived (secondary) plastid gain occurred at the root of the group consisting of stramenopiles, cryptophytes, haptophytes, and alveolates. Support for this theory comes from pigmentation, morphological observations, and plastid phylogenetic data (reviewed in Sanchez-Puerta and Delwiche 2008). Many works have challenged this hypothesis and changed the topology of the tree of eukaryotes since (e.g., Falkowski et al. 2004; Bodył 2005; Bodył et al. 2009; Keeling 2013; Burki et al. 2016; Waller et al. 2016). Notably, deep-branching representatives of the

"chromalveolate" crown groups are heterotrophic and probably plastid-less. Oomycota and Labyrinthulomycota are early branching stramenopiles; ciliates and *Acavomonas* (Janouškovec et al. 2013) branch close to the common ancestor of dinoflagellates and apicomplexans; katablepharids (Okamoto et al. 2009) and centrohelids (Burki et al. 2016) branch as early cryptophytes and haptophytes, respectively.

Two alternative explanations are at hand to explain topologies where the photoautotrophic crown group is nested within heterotrophic lineages. The chromalveolate hypothesis represents a "plastid-early" scenario and suggests that plastids were lost from early branching lineages, before becoming essential for the host cell survival (Oborník et al. 2009). "Plastid-late" scenario, on the other hand, implies plastid acquisition in the common ancestor of the phototrophic crown group only, while presuming early branching lineages to be primarily heterotrophic. For stramenopiles, alveolates, haptophytes, and cryptophytes, this scenario necessitates four independent endosymbioses with a rhodophyte (Falkowski et al. 2004). Under closer look, however, both these scenarios are problematic. Despite similarities seen in plastids of the "chromalveolates," the monophyly of this group is inconsistent with molecular phylogenies performed on large and genomic scales that tear the crown groups apart (e.g., Rodríguez-Ezpeleta et al. 2007; Okamoto et al. 2009; Brown et al. 2013; Burki et al. 2016; He et al. 2016). Particularly the addition of mostly heterotrophic rhizarians to the company of stramenopiles and alveolates defined the now generally acknowledged SAR supergroup to the exclusion of haptophytes and cryptophytes (Burki et al. 2007). The position of haptophytes and cryptophytes was clarified recently using comprehensive phylogenomic analyses, showing that haptophytes form a clade sister to SAR group, while cryptophytes branch with high support within Archaeplastida (Baurain et al. 2010; Burki et al. 2016; He et al. 2016). According to some data, cryptophytes branch as a sister group to chlorophytes and glaucophytes clade, with rhodophytes on the root (Burki et al. 2016). Furthermore, consistent presence of deep-branching heterotrophs favors the "plastid-late scenario."

On the other hand, not all early branching lineages are necessarily primary heterotrophs. Similarly to the spectrum of reductions seen in mitochondria and mitochondria-related organelles, plastids might be genetically and morphologically reduced to the form of an inconspicuous tiny vesicle lacking a genome, or even lost as in *Hematodinium* (Gornik et al. 2015) and *Cryptosporidium* (Zhu et al. 2000). Finding relic plastids in an early branching heterotrophic lineage would argue strongly against the "plastid-late" scenario, which triggered the search for plastid remnants in many non-photosynthetic lineages with more or less success. Structural features reminiscent of plastids were found after closer inspection (*Perkinsus*: Stelter et al. 2007; Robledo et al. 2011; *Voromonas*: Gile and Slamovits 2014). Besides ultrastructural evidence, "relic genes" from horizontal gene transfer support hypotheses about ancient endosymbioses. Horizontal (endosymbiotic) gene transfer (HGT) has a major role in forming organellar proteomes of the resulting symbiotic organism, and successfully transferred genes may account for high percentage of the total genetic material, while many other genes are lost during evolution

(reviewed by Archibald 2015). Unaccountable phylogenetic signals are sometimes reported, and early branching heterotrophs including oomycetes may hold genes witnessing ancient endosymbioses (Tyler et al. 2006).

Further, some plastid features are too complex to be established several times independently, putting "plastid-late" scenario in apparent doubt. For instance, plastid import machinery (SELMA) seems to have a common origin in cryptophytes, alveolates, stramenopiles, and haptophytes (Felsner et al. 2011). This led to arbitrary grouping of these taxa to the CASH group (Petersen et al. 2014), explaining the discrepancy between plastid and nuclear evolution by higher-order endosymbioses leading to horizontal transfer of an original rhodophyte-derived plastid between the unrelated lineages of the CASH group. Regardless of the order of these endosymbioses, latest phylogenomic studies undoubtedly show that the plastid ancestor of apicoplast could have been acquired in only a handful of different evolutionary points (Burki et al. 2016; He et al. 2016): (1) by the ancestor of SAR group and haptophytes, (2) by the ancestor of SAR group, or (3) by the ancestor of Myzozoa (including dinoflagellates, perkinsids, apicomplexans, chromerids, and colpodellids) (as in Fig. 1). Evidence accumulates that the most probable time point of plastid acquisition was at the root of myzozoans (Sanchez-Puerta and Delwiche 2008; Janouškovec et al. 2013a; Petersen et al. 2014; Ševčíková et al. 2015; Waller et al. 2016). This possibly triggered a radiation of dinoflagellate and apicomplexan species from their rather uniform colponemid-like ancestors (Waller et al. 2016).

Although a great deal of time has passed since this evolutionary milestone, placing the myzozoan endosymbiosis more robustly on the tree of eukaryotes allows some speculations concerning the morphology and biology of the pre-endosymbiotic ancestor. It might have been similar to *Acavomonas* (Tikhonenkov et al. 2014) and *Colponema* (Janouškovec et al. 2013a), heterotrophic flagellates equipped with two heterodynamic flagella at the anterior apex of the cell, a posterior digestive vacuole, extrusive organelles for active prey capture, and alveolar vesicles. Most myzozoans possess an apical complex or its modification, and they are capable of myzocytosis, a unique way of feeding by penetrating the surface of the prey and sucking its content into a feeding vacuole. Development of myzocytosis opened new strategies of predatory behavior and possibly also parasitism (Leander and Keeling 2003; Cavalier-Smith and Chao 2004; Janouškovec et al. 2013). Secondary heterotrophic colpodellids feed on their prey quite similarly to how early branching Apicomplexa parasitize; gregarines are extracellular parasites, feeding on the content of gut epithelial cells by penetrating their cytoplasmic membrane using a membrane-attachment ring. Feeding structures develop inside the host cell (Valigurová et al. 2007). Similarly, myzocytosis is employed by parasitic and heterotrophic dinoflagellates (*Perkinsus*, *Paulsenella*, *Pfiesteria*). Supposedly, the major difference between predation on single-celled organisms by colpodellids and early stages of parasitism in Apicomplexa is, therefore, the multicellularity of the prey in the latter. At the same time, many alveolates (including dinoflagellates *Noctiluca*, *Oxyrrhis*, and *Gyrodinium*) use phagocytosis to ingest prey as a whole. Therefore, it is unclear whether the pre-endosymbiotic ancestor fed using myzocytosis or phagocytosis (or both) and how it eventually captured the endosymbiont.

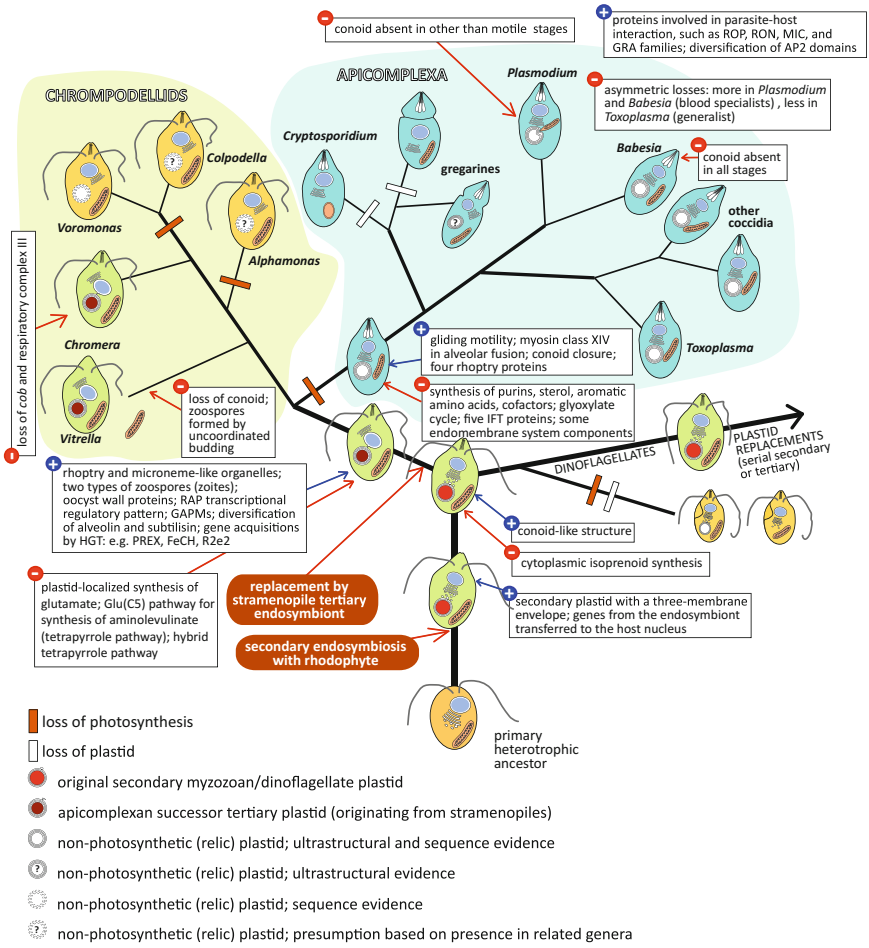


Fig. 1 Evolution of parasitism-related traits in the history of Apicomplexa. The scheme depicts major changes appearing along the diversification of trophic strategies in chrompodellids and apicomplexans. The main events include the plastid endosymbiosis and the appearance of apical-like structures. Many gene acquisitions occurred before the divergence of chrompodellids, followed by further innovations of parasite–host interaction proteins in apicomplexans. However, a switch to parasitism was accompanied by a great reduction in genetic and metabolic complexity (Woo et al. 2015) and repurposing of ancestral features such as the apical conoid, a complex life cycle, and adaptations to anaerobic environments. Abbreviations: AP2, apicomplexan Apetala2-like domain transcription factors; FeCH, ferrochelatase; GAPMs, glideosome-associated proteins; GRA, parasitophorous dense granule protein family; HGT, horizontal gene transfer; IFT, intraflagellar transport; MIC, micronemes proteins; PREX, plastidic replication/repair enzyme complex; R2e2, ribonucleotide reductase; RAP, rhoptry-associated protein; RON, rhoptry neck proteins; ROP, rhoptry proteins

First apicomplexans were probably metabolically complex. *Chromera velia* as the closest known phototrophic relatives of Apicomplexa (Woo et al. 2015) exhibits a peculiar mitochondrial metabolism (Flegontov et al. 2015; Oborník and Lukeš 2015), well-adjusted to aerobic and anaerobic conditions. We can speculate that the heterotrophic ancestors of Myzozoa were highly metabolically flexible as well. Colponemids, close relatives to both Myzozoa and Ciliata, are frequently found in freshwater, saline, and soil environments (Janouškovec et al. 2013a); especially soil species might show adaptations to low oxygen. Such genomic and metabolic complexity and ecological plasticity may reflect complex and unstable environmental conditions the pre-endosymbiotic ancestor of Apicomplexa inhabited.

3 How to Capture a Plastid

Acquisitions of plastid occurred rarely in the evolutionary history of eukaryotes. It is believed that primary (prokaryote-to-eukaryote) endosymbiosis between a eukaryote and a cyanobacterium involved only one major group (the Archaeplastida), while another recent (app. 60 Mya) primary endosymbiosis was established in the cercozoan ameba *Paulinella* (e.g., Delwiche 1999; Keeling 2013). Higher-order (eukaryote-to-eukaryote, or complex) endosymbioses are relatively more frequent than primary endosymbiotic events; the number of higher-order events in previously plastid-less eukaryotic lineages is currently estimated between four and seven (Sanchez-Puerta and Delwiche 2008; Petersen et al. 2014; Stiller et al. 2014; Ševčíková et al. 2015; Waller et al. 2016; Burki et al. 2016). Two endosymbioses involved green algae (yielding euglenophytes and chlorarachniophytes), and two-to-five endosymbioses involved red-algal or red-derived plastids (yielding the CASH lineages). Further, plastids were numerous times serially replaced in dinotoms *Durinskia* and *Glenodinium*/*Kryptoperidinium*, *Karenia*, and the "green" dinoflagellate *Lepidodinium*. We cannot rule out the possibility that cryptophytes were photoautotrophic before the acquisition of their current plastid (Baurain et al. 2010; Burki et al. 2016).

Convincing recognition of a plastid in the multimembranous vesicle of apicomplexans (Köhler et al. 1997) brought about questions concerning its evolutionary history. Apicoplast genomic DNA was isolated from apicomplexan parasites by Kilejian (1975), but it was only later shown that this circular DNA molecule displays similarity to plastid genomes (Gardner et al. 1991, 1994; Howe 1992). Evolutionary affiliation of the apicoplast with the red lineage was proposed already a decade before identification of *C. velia* (McFadden and Waller 1997; Stoebe and Kowallik 1999; Zhang et al. 2000), based on the similar structure of the apicoplast ribosomal superoperon with orthologous regions in rhodophyte and rhodophyte-derived plastid genomes. Affiliation to eukaryotic algae rather than cyanobacteria was also consistent with the presence of four membranes surrounding

the compartment (primary plastids have two). The origin of apicoplast in the red lineage was later supported by other phylogenetic analyses (e.g., Blanchard and Hicks 1999; Coppin et al. 2005) and, most convincingly, the discovery of closely related phototrophic chromerids (Moore et al. 2008; Oborník et al. 2012). These reports disprove data demonstrating a weak association with green algae based on the phylogeny of *tufA* gene (Köhler et al. 1997) and the split of the nuclear-encoded *cox2* gene coding for mitochondrial cytochrome oxidase, a genetic trait found in both Apicomplexa and leguminous plants (Funes et al. 2002; 2004). In contrast, Waller et al. (2003) pointed out that *cox2* is split into two in the mitochondrial genome of ciliates, and the split appeared convergently in unrelated lineages. However, the hypothesis on the green origin of the apicoplast still sometimes emerges from oblivion (Lau et al. 2009).

Despite a supposed plastid acquisition at the base of Myzozoa, accumulating evidence from molecular phylogeny (Moore et al. 2008; Janouškovec et al. 2010, 2015; Woo et al. 2015) and the structure of the ribosomal operon (Janouškovec et al. 2010) indicates that plastids of chromerids, apicomplexans, and stramenopiles share a common ancestor. Some further characters such as pigmentation (Moore et al. 2008; Kotabová et al. 2011; Oborník et al. 2012; Bína et al. 2014), particularly the presence of isofucoxanthin (Oborník et al. 2012), linear-mapping plastid genome of *Chromera* (Janouškovec et al. 2013), and phylogeny of plastid-encoded genes (Ševčíková et al. 2015) point to a tertiary (stramenopile) rather than a secondary (rhodophyte) endosymbiotic origin of chromerid and apicomplexan plastids. Plastid-encoded genes display some affiliation with eustigmatophytes (Ševčíková et al. 2015). Moreover, both eustigmatophytes and chromerids lack chlorophyll *c*, a hallmark pigment of algae with rhodophyte-derived plastids (Moore et al. 2008; Oborník et al. 2012). It remains to be determined whether the original myzozoan plastid, acquired before the divergence of dinoflagellates, was replaced in the common ancestor of chromerids, colpodellids, and apicomplexans by a eustigmatophyte (or a related stramenopile) endosymbiont.

Whatever the source of the plastid, it is important to captivate, along with the organelle, factors to support its functionality. Indeed, a heterotrophic host (exosymbiont) nucleus does not initially contain any genes associated with plastid homeostasis and photosynthesis. Thus, plastid survival must be ensured by factors encoded by the endosymbiont nuclear and plastid genomes. Only later these genes are functionally transferred to host nucleus (Imanian and Keeling 2014). Kleptoplastidic lineages such as the dinoflagellate *Dinophysis* and the foraminifer *Elphidium* might represent an interim stage in the process of endosymbiosis, when plastids are stolen, while the remainder of the prey cell is being digested (Pillet 2013). Remarkably, these stolen organelles have high longevity up to several months after ingestion (Correia and Lee 2002). Since kleptoplastids are not genetically independent (they are supported by factors from the original alga), horizontal gene transfer could provide means how plastid-related factors are supplied from the new host to the organelle to increase its longevity (Wisecaver and Hackett 2010; Pillet 2013).

Serial replacement of plastids could, therefore, occur more easily than acquisitions by primary heterotrophs. Factors for plastid maintenance have been transferred to the host nucleus during the first endosymbiosis, and these factors are available for reuse in the newly acquired plastid. The case of *Lepidodinium*, dinoflagellate with a successor chlorophyte-derived plastid, demonstrates that a new endosymbiont can be acquired even when the previous plastid is still metabolically active. Many proteins of the plastid metabolism show phylogenetic origin in the former peridinin plastid suggesting their continued use all along the new plastid integration (Minge et al. 2010; Cihlář et al. 2016). In dinotoms, two functionally redundant plastid compartments are present. The original peridinin plastid is non-photosynthetic but provides its host with tetrapyrroles and isoprenoids; the diatom plastid is photosynthetic and to a certain extent independent of the host (Hehenberger et al. 2014; Imanian and Keeling 2014). Similarly, if we consider the independent origin of *C. velia* plastid, type II RuBisCO was inherited from the previous plastid according to the "shopping bag" hypothesis of collecting and repurposing genes all along an organisms' evolutionary history (Larkum et al. 2007). Therefore, the newcomer plastid in serial endosymbiosis does not necessarily undergo endosymbiotic gene transfer to the extent seen in the original plastid. In other words, the newcomer plastid needs not to be accompanied by the as many genes to be functional because it can use the molecular machinery of the previous plastid (Minge et al. 2010; Cihlář et al. 2016). Consistently, plastid-bearing lineages possess genes of unknown origin, which suggests previous endosymbioses masked by more recent ones or heavy impact of horizontal gene transfer from prey (Doolittle 1998; Curtis et al. 2012; Cihlář et al. 2016).

In consent with transport mechanisms to ensure plastid viability and function being established, the number of plastid membranes needs to be stabilized quite soon after the endosymbiotic interaction. Indeed, transport of proteins and metabolites across the plastid envelope is achieved by factors specific to individual membranes (e.g., Felsner et al. 2011; Lim et al. 2010; Moog et al. 2015). Later gains and losses of plastid membranes would seriously interfere with the passage of solutes and, more importantly, plastid proteins because of impaired localization. The evolution of the number of plastid membranes is therefore intimately linked with the functionality of protein transport. This can be illustrated again by a relative ease of serial plastid replacement in *Lepidodinium* with a four-membrane successor plastid, compared to dinotoms that contain an almost fully functional diatom endosymbiont separated from the host by a single-unit membrane (Eschbach et al. 1990). An apparent transporter incompatibility in dinotoms resulted in parallel metabolic pathways (Hehenberger et al. 2014; Imanian and Keeling 2014; Cihlář et al. 2016). Protein transport mechanism to the plastid of *C. velia* moderately supports the hypothesis of separate plastid origins in apicomplexans and dinoflagellates, as plastid-targeted proteins show different physicochemical properties than dinoflagellate plastid proteins (our unpublished data).

4 Reductive Evolution of Organelles

All organisms are subject to reductive evolution after short periods of increasing biological complexity (Wolf and Koonin 2013). Genome reductions accompanied by gene transfer into the host nucleus are fundamental processes also in organellar endosymbioses. A typical example of such reductive process is the evolution of mitochondria in alveolates. In the ancestor of alveolates, the mitochondrial genome was large and circular, about 50 Kb in length, and linearization took place before the divergence of ciliates and early branching colponemids such as *Acavomonas* (Janouškovec et al. 2013; Tikhonenkov et al. 2014). In the anaerobic ciliate *Nyctotherus ovalis,* the mitochondrion evolved into a hydrogenosome with a reduced 14-Kb linear genome (de Graaf et al. 2011; Oborník and Lukeš 2015). Drastic reduction of the mitochondrial genome occurred independently also at the root of myzozoans; this was possibly allowed by aminoacyl-tRNA import from the cytosol and the use of an alternative NADH dehydrogenase (Janouškovec et al. 2013).

Mitochondrial genomes of all extant myzozoan lineages are miniature and contain only three structural genes, coding for cytochrome oxidase 1 (*cox1*), cytochrome oxidase 3 (*cox3*), and cytochrome b (*cob*), and a set of genes encoding rRNA fragments (Waller and Jackson 2009; Nash et al. 2008; Jackson et al. 2012; Flegontov et al. 2015; Oborník and Lukeš et al. 2015). The mitochondrial genome of *C. velia* is further reduced by missing the *cob* gene (Flegontov et al. 2015). In apicomplexans, mitochondrial genomes are organized as linear homogeneous molecules from 6 to 11 Kb; nothing is known about the mitochondrial genomes of early branching gregarines, obviously due to their little economic importance. High level of mitochondrial reduction was observed in the genus *Cryptosporidium*, close relatives of gregarines and facultative parasites of immunocompromised patients. *Cryptosporidium* possesses a mitosome-like organelle lacking a genome. In comparison, mitochondrial genomes of dinoflagellates, chromerids, and colpodellids are usually fragmented to heterogeneous DNA molecules, containing single gene, gene fragments, or fused genes (Slamovits et al. 2007; Slamovits and Keeling 2008; Nash et al. 2008; Waller and Jackson 2009; Jackson et al. 2012; Slamovits 2014; Oborník and Lukeš 2015). Despite similar genome topology, the dinoflagellate mitochondrial mRNA is processed by RNA editing, a mechanism not found in chromerids (Flegontov et al. 2015).

The lack of *cob* in the mitochondrial genome of *C. velia* led to an inspection of the presence of nuclear-encoded subunits of respiratory chains in both chromerids (Flegontov et al. 2015). The absence of respective nuclear-encoded subunits suggests that the respiratory chain in *C. velia* lacks complexes I (NADH dehydrogenase; this complex is absent also from Apicomplexa) and III (cytochrome *c* reductase). The remaining complexes are arranged into two disconnected electron transport subchains. L- and D-lactate cytochrome c oxidoreductases were proposed to provide electrons for complex IV (cytochrome *c* oxidase), thus bypassing the missing complex III. The respiratory chain of *V. brassicaformis* has a very similar

organization but, notably, complex III is retained; alternative NADH dehydrogenases, dihydroorotate dehydrogenase, electron transfer flavoprotein:ubiquinone oxidoreductase, glycerol 3-phosphate dehydrogenase, sulfide:ubiquinone oxidoreductase, alternative oxidase, and L-/D- lactate:cytochrome c oxidoreductases are all present in *V. brassicaformis*, only the bidirectional D-lactate dehydrogenase for reversible lactate/pyruvate conversion is missing. Similar mitochondrial biochemistry in *Chromera* and *Vitrella* suggests that this arrangement is ancestral for both species, rather than being a low-oxygen adaptation in *C. velia*.

Strikingly, the plastid genomes of known chromerids differ to such extent that it is rather difficult to find any general characteristics. While *Vitrella* possesses circular, highly compacted, and GC-rich genome (about 80 Kb), *Chromera* hosts a larger (120 Kb) and highly divergent, linear-mapping genome, with inverted repeats of 3 genes at both ends of the DNA molecule (Janouškovec et al. 2010). Moreover, *psaA* and *atpB* are split into two fragments in *Chromera* that are independently transcribed, translated, processed, and incorporated into the functional photosystem I and ATP synthase complexes (Janouškovec et al. 2013b). Further, some proteins encoded in the plastid of *Chromera* use a non-canonical genetic code with UGA encoding tryptophan, while *Vitrella* uses the canonical code (Moore et al. 2008; Janouškovec et al. 2010; Woo et al. 2015). Although the plastid genome of *C. velia* is considerably larger than that of *V. brassicaformis*, it is missing a couple of genes (Janouškovec et al. 2010). *C. velia* possesses a highly divergent and eroded plastid genome when compared not only to *V. brassicaformis* but also to the circular genome of the apicoplast. In fact, with photosynthesis-related genes removed, the plastid genome of *Vitrella* resembles the apicoplast genome much more than the plastid genome of *Chromera*. Therefore, organellar reduction in *C. velia* goes far beyond the reduction found in non-parasitic organisms. This is in good agreement with the basal phylogenetic position of *Vitrella* among chrompodellids (chromerids + colpodellids; Janouškovec et al. 2015), with *Chromera* branching as the more advanced phototroph in the clade.

5 Nuclear Genome Reduction in the Evolution of Apicomplexa

The nuclear genomes of chromerids show similar evolutionary tendencies as the plastid genomes. The nuclear genome in *Chromera* is considerably expanded (193.6 Mb) compared to *Vitrella* (72.7 Mb), but a roughly similar number of predicted genes (*C. velia* 26,112 vs. *V. brassicaformis* 22,817) demonstrate how gene density is substantially higher in *Vitrella*. These differences are mainly caused by the apparently higher occurrence of transposable elements (TEs) in the nuclear genome of *Chromera* and not by reduced metabolic capacity of *V. brassicaformis* (Woo et al 2015; Flegontov et al. 2015). Still, Class I elements prevail over Class II elements in the nuclei of both chromerids. Interestingly, TEs from the

apicomplexan *Eimeria tenella* and chromerids are not related, and variation in the RT domain demonstrates the independent evolution of TEs in these lineages (Woo et al. 2015). TEs are absent from other apicomplexans, pointing out different evolutionary forces forming their nuclear genomes (DeBarry and Kissinger 2011). Evolutionary and metabolic analyses showed that during the trophic transition to parasitism, as many as 3862 genes present in the free-living phototrophic ancestor were lost during the evolution of parasitism in apicomplexans. At the same time, only few novel genes were gained (81 genes), besides lineage- and species-specific gene losses and gains (Woo et al. 2015). This tendency to gene loss suggests that the photoautotrophic ancestor of Apicomplexa possessed most of the factors for parasitism-related processes employed by extant parasitic descendants. Many of these factors, however, gained novel or modified functions to suit parasite–host interactions, including the components of the motility apparatus, DNA- and RNA-binding proteins, and extracellular proteins (Woo et al. 2015).

Similarly, morphological features of the cells were changing with the transition from photoautotrophy to parasitism. The flagellar apparatus, for instance, contributed to the evolution of apical complex to enable effective host cell penetration (Portman and Šlapeta 2014). Great rearrangements also followed from the metabolic transition from organic nutrient self-sustainability to a complete dependence on supplies from the host.

6 Mosaic Pathways in Apicomplexans and Chromerids

Photoautotrophic organisms use their plastids as molecular factories. Many biosynthetic pathways, such as biosynthesis of tetrapyrroles, isoprenoids, fatty acids, vitamins, and iron–sulfur cluster assembly, are directly or indirectly linked to reactions of the photosynthetic apparatus. Upon endosymbiotic interaction, these plastid pathways become redundant with the canonical (cytosolic and mitochondrial) pathways of the host. Some or all steps of the canonical or plastid pathways might be lost during the streamlining of cellular biochemistry. If a canonical pathway is lost, the plastid gains an essential biochemical role for the cell and becomes indispensable. How endosymbiotic events shape metabolic pathways can be exemplified on tetrapyrrole biosynthesis, but similar scenarios have taken place in streamlining other host/endosymbiont metabolic redundancies (Waller et al. 2016).

Tetrapyrroles such as heme or chlorophyll are indispensable for life as we know it. Chlorophyll captures the energy of sunlight, while heme is a cofactor of many proteins associated with energetic catabolism. Among eukaryotes, only a single aerobic organism has been shown to be able to live without heme (Kořený et al. 2012). In primary heterotrophic eukaryotes, such as animals and fungi, heme biosynthesis steps alternate between the mitochondrion and the cytosol.

The pathway starts with formation of aminolevulinate (ALA) by condensation of succinyl-CoA and glycine in the mitochondrion, next three or four steps localize in the cytosol, and the pathway terminates again in mitochondria (Kořený et al. 2011), which likely allows feedback regulation (Masuda and Fujita 2008; Czarnecki and Grimm 2012). In contrast, most studied phototrophs convergently localize their tetrapyrrole (both heme and chlorophyll) biosynthesis entirely in the plastid. ALA is then synthesized from glutamate, but the following enzymatic steps are conserved among all phototrophs and heterotrophs. Individual enzymes, however, display different evolutionary origins in eukaryotic phototrophs, mostly eukaryotic, cyanobacterial, or proteobacterial, which is referred to as mosaic pathway arrangement (Oborník and Green 2005; Kořený et al. 2011; Cihlář et al. 2016). It was suggested that biparallel tetrapyrrole biosynthesis by mitochondrial/cytosolic and plastid pathways as seen in photoautotrophic *Euglena* and *Bigelowiella* (Kořený and Oborník 2011; Cihlář et al. 2016) represents a temporary stage during plastid endosymbiosis.

In Apicomplexa, heme synthesis is unique in its localization partially in three compartments (Sato et al. 2004; Ralph et al. 2004). Although Apicomplexa host a non-photosynthetic plastid, ALA is synthesized in the mitochondrion from glycine (and succinyl-CoA), like in heterotrophic eukaryotes. ALA is then exported to the apicoplast, where the following 3–4 steps take place. The pathway then continues in the cytosol and terminates in the mitochondrion by protoporphyrinogen oxidase and ferrochelatase (Sato et al. 2004; Ralph et al. 2004; Seeber and Soldati-Favre 2010; Kořený et al. 2013). This organization is therefore similar to the situation in primary heterotrophic eukaryotes; still, many of apicomplexan heme pathway enzymes were recruited from the endosymbiont and share origins with photoautotrophic orthologs (Kořený et al. 2011).

Chromerids also possess an unusual tetrapyrrole biosynthesis pathway. Similarly to apicomplexans, both *Chromera* and *Vitrella* synthesize ALA in mitochondria by the "heterotrophic" pathway. Chromerids are therefore the only known phototrophs synthesizing chlorophyll from glycine and not from glutamate (Kořený et al. 2011; Woo et al. 2015; Janouškovec et al. 2015). This is despite their dependence on photosynthesis, and hence a supposed higher need for tetrapyrroles to generate chlorophyll. We believe that this organization of tetrapyrrole pathway in the phototrophic ancestor of apicomplexans played a major role in the transition from photoautotrophy to heterotrophy (van Dooren et al. 2012). Notably, photosynthesis was lost at least three times in the apicomplexan–chromerid–colpodellid lineage, which is evident from the phylogenetic position of *Chromera* and *Vitrella* (Janouškovec et al. 2015). Hence, reduction of metabolic complexity is another process that accompanies the trophic transition from photoautotrophy to parasitism in apicomplexans.

7 Similarities in Life Cycles of Chromerids and Apicomplexans

Life cycles of chromerids have been extensively studied (Moore et al. 2008; Sato 2011; Oborník et al. 2011, 2012; Oborník and Lukeš 2013; Oborník et al. 2016; Füssy et al. 2017). *Chromera* exhibits a simple life cycle involving vegetative coccoid cells, autosporangia containing up to 4 spores, and zoosporangia with up to 10 spores. In adverse conditions, thick-walled cysts resilient to various stresses are formed. Zoosporogenesis in *Chromera* resembles schizogony of apicomplexan parasites (Oborník et al. 2016); the zoospores are equipped with two heterodynamic flagella and possess a primitive form of the apical complex. In contrast, *Vitrella* exhibits a fairly complex life cycle. Vegetative cells and large autosporangia with dozens of autospores represent the vegetative cycle. In *Vitrella*, two kinds of zoosporangia develop with different zoospores; one type of zoospores form flagella inside the cytoplasm similarly to microgametes of *Plasmodium*, the other type build their flagella extracellularly using the intraflagellar transport mechanism similarly to microgametes of *Toxoplasma*. The zoospores of *Vitrella* lack any traces of an apical complex-like structure, which was proposed to be associated with the "uncoordinated" (one-by-one) budding of zoospores from the mother cell (Füssy et al. 2017). Life cycles of apicomplexan parasites are generally quite complicated (for a recent review, see Votýpka et al. in press).

It is possible that *C. velia* and colpodellids minimized their life cycles to the core; and while *Vitrella* exhibits sexual behavior, *Chromera* very likely lost sexuality. If the uncoordinated formation of zoospores by budding and an absence of apical complex-like structure in *V. brassicaformis* (Füssy et al. 2017) are ancestral characters (*Vitrella* occupies basal phylogenetic position among chrompodellids; Janouškovec et al. 2015; Oborník and Lukeš 2015), the apical complex-like structures as well as schizogony-like formation of zoospores had to evolve at least twice independently, in chromerids and apicomplexans (Oborník et al. 2016; Füssy et al. 2017). More likely, the apical complex-like structure was lost from *Vitrella*, and the pseudoconoid in *Chromera*, apical complex-like structures in colpodellids, and the apical complexes in apicomplexan parasites are homologous structures; schizogony might still be discovered in *Vitrella*, as we have insufficient data on zoospore formation in the second type of zoosporangia.

We can speculate that the ancestor of apicomplexans, chromerids, and colpodellids had a very complex life cycle, which reflected the complex and dynamically diverse environmental conditions it inhabited. Parts of this life cycle were gradually reduced in different lineages as an adaptation to diverging environmental niches and trophic strategies. In phototrophic chromerids, vegetative cells apparently represent the dominant stage of the cycle, because only the vegetative cells rely on photosynthesis and are almost filled with the plastid. Zoospores possess a temporarily reduced plastid and occur only at a specific time during cultivation, possibly to allow dispersal (Oborník et al. 2011; Oborník et al. 2016; Füssy et al. 2017). In the related colpodellids, however, zoospores (termed the

trophozoites) are the dominating stage because they are capable of predation and are thus responsible for the energy acquisition (Oborník et al. 2016). The ancestral life cycle complexity might have allowed the evolution of a wide range of strictly specific parasitic species by conservation of cycle segments that were suitable for different surroundings. For instance, life cycle adjustments could give rise to apicomplexan dixenic cycles to accommodate propagation in two distinct host environments. Comparison of molecular processes underlying life cycle progression and switches in apicomplexans parasites and their free-living cousins would shed more light on how life cycle modifications contributed to host recognition and parasite dispersal.

8 Future Perspectives

Trait comparison in free-living and parasitic species of Myzozoa allows us to conclude that the common ancestor of these lineages was a very complex organism, and reductive evolution through loss of genetic and metabolic capacity was most probably the driving force that confined apicomplexans to parasitism. At the same time, it is important to understand how existing processes and factors in free-living species are redefined for strictly parasitic purposes. In parasitic rhodophytes (Salomaki and Lane 2014), higher plants (Krause 2008), and *Helicosporidium*, similarly, reductive evolution (de Koning and Keeling 2006) and repurposing of ancestral traits (Pombert et al. 2014) led to a switch to parasitism. Rhodophytes and plants benefit from a close evolutionary relationship between the parasite and its host, which allowed exploitation of the host's resources. How the green alga *Helicosporidium* adapted to parasitism in insects is a more intriguing question. It is to be discovered which features, besides an expanded chitinase family, this non-photosynthetic alga was able to repurpose for its new trophic strategy, and whether there are similar consequences for the future direction of this evolution as in apicomplexans. Hopefully, investigation of these parasite adaptations will lead to better disease control of these and other widespread parasites.

References

Archibald J (2015) Genomic perspectives on the birth and spread of plastids. Proc Natl Acad Sci USA 112:10147–10153

Baurain D, Brinkmann H, Petersen J et al (2010) Phylogenomic evidence for separate acquisition of plastids in cryptophytes, haptophytes, and stramenopiles. Mol Biol Evol 27:1698–1709

Bína D, Gardian Z, Herbstová M et al (2014) Novel type of red-shifted chlorophyll *a* antenna complex from *Chromera velia*: II. Biochemistry and spectroscopy. Biochim Biophys Acta Bioenerget 1837:802–810

Blanchard JL, Hicks JS (1999) The non-photosynthetic plastid in malarial parasite and other apicomplexans is derived from outside the green plastid lineage. J Euk Microbiol 46:367–375

Blouin NA, Lane CE (2015) Red algae provide fertile ground for exploring parasite evolution. Persp Phycol 3:11–19
Bodył A (2005) Do plastid-related characters support the chromalveolate hypothesis? J Phycol 41:712–719
Bodył A, Stiller JW, Mackiewicz P (2009) Chromalveolate plastids: direct descent or multiple endosymbioses? Trend Ecol Evol 24:119–121
Burki F, Shalchian-Tabrizi K, Minge M et al (2007) Phylogenomics reshuffles the eukaryotic supergroups. PLoS ONE 2:e790
Burki F, Kaplan M, Tikhonenkov DV et al (2016) Untangling the early diversification of eukaryotes: a phylogenomic study of the evolutionary origins of Centrohelida, Haptophyta and Cryptista. Proc R Soc B 283:20152802
Brown MW, Sharpe SC, Silberman JD et al (2013) Phylogenomics demonstrates that breviate flagellates are related to opisthokonts and apusomonads. Proc R Soc B 280:20131755
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 Euk Microbiol 46:347–366
Cavalier-Smith T, Chao EE (2004) Protalveolate phylogeny and systematics and the origins of Sporozoa and dinoflagellates (phylum Myzozoa nom. nov.). Eur J Protistol 40:185–212. doi:10.1016/j.ejop.2004.01.002
Cihlář J, Füssy Z, Horák A et al (2016) Evolution of the tetrapyrrole biosynthetic pathway in secondary algae: conservation, redundancy, and replacement. PLoS ONE 11:e0166338
Coppin A, Varre JS, Lienard L et al (2005) Evolution of plant-like crystalline storage polysaccharide in the protozoan *Toxoplasma gondii* argues for a red alga ancestry. Mol Biol Evol 60:257–267
Correia MJ, Lee JJ (2002) How long do the plastids retained by *Elphidium excavatum* (Terquem) last in their host? Symbiosis 32:27–38
Cumbo VR, Baird AH, Moore R et al (2013) *Chromera velia* is endosymbiotic in larvae of the reef corals *Acropora digitifera* and *A. tenuis*. Protist 164:237–244
Curtis BA, Tanifuji G, Burki F et al (2012) Algal genomes reveal evolutionary mosaicism and the fate of nucleomorphs. Nature 492:59–65
Czarnecki O, Grimm B (2012) Post-translational control of tetrapyrrole biosynthesis in plants, algae, and cyanobacteria. J Exp Bot 63:1675–1687. doi:10.1093/jxb/err437
DeBarry JD, Kissinger JC (2011) Jumbled genomes: missing apicomplexan synteny. Mol Bio Evol 28:2855–2871
de Graaf RM, Ricard G, van Alen TA et al (2011) The organellar genome and metabolic potential of the hydrogen-producing mitochondrion of *Nyctotherus ovalis*. Mol Biol Evol 28:2379–2391
de Koning AP, Keeling PJ (2006) The complete plastid genome sequence of the parasitic green alga Helicosporidium sp. is highly reduced and structured. BMC Biol 4:12
Delwiche CF (1999) Tracing the thread of plastid diversity through the tapestry of life. Am Nat 154:S164–S177
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
Eschbach S, Speth V, Hansmann P, Sitte P (1990) Freeze-fracture study of the single membrane between host cell and endocytobiont in the dinoflagellates *Glenodinium foliaceum* and *Peridinium balticum*. J Phycol 26:324–328
Falkowski PG, Katz ME, Knoll AH et al (2004) The evolution of modern eukaryotic phytoplankton. Science 305:354–360
Felsner G, Sommer MS, Gruenheit N et al (2011) ERAD components in organisms with complex red plastids suggest recruitment of a preexisting protein transport pathway for the periplastid membrane. Genome Biol Evol 3:140–150
Flegontov P, Michálek J, Janouškovec J et al (2015) Divergent mitochondrial respiratory chains in phototrophic relatives of apicomplexan parasites. Mol Biol Evol 32:1115–1131
Flegr J, Prandota J, Sovickova M et al (2014) Toxoplasmosis—a global threat. Correlation of latent toxoplasmosis with specific disease burden in a set of 88 countries. PLoS ONE 9:e90203

Flegr J (2013) How and why *Toxoplasma* makes us crazy. Trend Parasitol 29:156–163
Funes S, Davidson E, Reyes-Prieto A, Magallon S, Herion P, King MP, Gonzales-Halphgen D (2002) A green algal apicoplast ancestor. Science 298:2155–2155
Füssy Z, Masařová P, Kručinská J, Esson H, Oborník M (2017) Budding of the alveolate alga Vitrella brassicaformis resembles sexual and asexual processes in apicomplexan parasites. Protist 168:80–91
Gardner MJ, Williamson DH, Wilson RMJ (1991) A circular DNA in malaria parasites encodes an RNA-polymerase like that of prokaryotes and chloroplasts. Mol Biochem Parasitol 44:115–123
Gardner MJ, Goldman N, Barnett P et al (1994) Phylogenetic analysis of the rpoB gene from the plastid-like DNA of *Plasmodium falciparum*. Mol Biochem Parasitol 66:221–231
Gile GH, Slamovits CH (2014) Transcriptomic analysis reveals evidence for a cryptic plastid in the colpodellid *Voromonas pontica*, a close relative of chromerids and apicomplexan parasites. PLoS ONE 9:e96258
Gornik AG, Febrimarsa Cassin AM et al (2015) Endosymbiosis undone by stepwise elimination of the plastid in a parasitic dinoflagellate. Proc Natl Acad Sci USA 112:5767–5772
He D, Sierra R, Pawlowski J, Baldauf SL (2016) Reducing long-branch effects in multi-protein data uncovers a close relationship between Alveolata and Rhizaria. Mol Phylogenet Evol 101:1–7
Hehenberger E, Imanian B, Burki F, Keeling PJ (2014) Evidence for the retention of two evolutionary distinct plastids in dinoflagellates with diatom endosymbionts. Genome Biol Evol 6:2321–2334
Howe CJ (1992) Plastid origin of an extrachromosomal DNA molecule from *Plasmodium*, the causative agent of malaria. J Theor Biol 158:199–205
Imanian B, Keeling PJ (2014) Horizontal gene transfer and redundancy of tryptophan biosynthetic enzymes in dinotoms. Genom Biol Evol 6:333–343
Jackson CJ, Gornik SG, Waller RF (2012) The mitochondrial genome and transcriptome of the basal dinoflagellate *Hematodinium* sp.: character evolution within the highly derived mitochondrial genomes of dinoflagellates. Genom Biol Evol 4:59–72
Janouškovec J, Horák A, Oborník M, Lukeš J, Keeling PJ (2010) A common red algal origin of the apicomplexan, dinoflagellate, and heterokont plastids. Proc Natl Acad Sci USA 107:10949–10954
Janouškovec J, Tikhonenkov DV, Mikhailov KV et al (2013a) Colponemids represent multiple ancient alveolate lineages. Curr Biol 23:2546–2552
Janouškovec J, Sobotka R, Lai D, et al (2013b) Split photosystem protein, linear-mapping topology and growth of structural complexity in the plastid genome of Chromera velia. Mol Biol Evol 30:2447–2462
Janouškovec J, Tikhonenkov DV, Burki F, Howe AT, Kolísko M, Mylnikov AP, Keeling PJ (2015) Factors mediating plastid dependency and the origins of parasitism in apicomplexans and their close relatives. Proc Natl Acad Sci USA 112:10200–10207
Janouškovec J, Keeling PJ (2016) Evolution: Causality and the origin of parasitism. Curr Biol 26: R174–R177
Keeling PJ (2008) Evolutionary biology: bridge over troublesome plastid. Nature 451:896–897
Keeling PJ (2013) The number, speed, and impact of plastid endosymbioses in eukaryotic evolution. Ann Rev Plant Biol 64:583–607
Kilejian A (1975) Circular mitochondrial DNA from avian malarial parasite *Plasmodium lophurae*. Biochim Biophys Acta 390:276–284
Köhler S, Delwiche C, Denny P et al (1997) A plastid of probable green algal origin in apicomplexan parasites. Science 275:1485–1489
Kotabová E, Kaňa R, Jarešová J, Prášil O (2011) Non-photochemical fluorescence quenching in *Chromera velia* is established by fast violaxanthin de-epoxidation. FEBS Lett 585:1941–1945
Kořený L, Oborník M (2011) Sequence evidence for the presence of two tetrapyrrole pathways in *Euglena gracilis*. Genome Biol Evol 3:359–364
Kořený L, Obornik M, Lukeš J (2013) Make it, take it, or leave it: heme metabolism of parasites. PLoS Pathog 9:e1003088. doi:10.1371/journal.ppat.1003088

Kořený L, Sobotka R, Janouškovec J et al (2011) Tetrapyrrole synthesis of photosynthetic chromerids is likely homologous to the unusual pathway of apicomplexan parasites. Plant Cell 23:3454–3462. doi:10.1105/tpc.111.089102

Kořený L, Sobotka R, Kovářová J et al (2012) Aerobic kinetoplastid flagellate *Phytomonas* does not require heme for viability. Proc Natl Acad Sci USA 109:3808–3813. doi:10.1073/pnas.1201089109

Krause K (2008) From chloroplasts to "cryptic" plastids: evolution of plastid genomes in parasitic plants. Curr Biol 54:111–121

Larkum AWD, Lockhart PJ, Howe CJ (2007) Shopping for plastids. Trends Plant Sci 12:189–195

Lau AO, McElwain TF, Brayton KA et al (2009) *Babesia bovis*: a comprehensive phylogenetic analysis of plastid-encoded genes supports green algal origin of apicoplasts. Exp Parasitol 123:236–243

Leander B, Keeling PJ (2003) Morphostasis in alveolate evolution. Trends Ecol Evol 18:395–402

Lesser MP, Stat M, Gates RD (2013) The endosymbiotic flagellates (*Symbiodinium* sp.) of corals are parasites and mutualists. Coral Reefs 32:603–611

Lester D (2012) Toxoplasma gondii and homicide. Psych Rep 111:196–197

Levine ND, Corliss JO, Cox FEG et al (1980) A newly revised classification of the protozoa. J Protozool 27:37–58

Lim L, Linka M, Mullin K et al (2010) The carbon and energy sources of the non-photosynthetic plastid in the malaria parasite. FEBS Lett 584:549–554

Masuda T, Fujita Y (2008) Regulation and evolution of chlorophyll metabolism. Photochem Photobiol Sci 7:1131–1149. doi:10.1039/b807210h

McFadden GI, Waller RF (1997) Plastids in parasites of humans. BioEssays 19:1033–1040

Minge MA, Shalchian-Tabrizi K, Torresen OK et al (2010) A phylogenetic mosaic plastid proteome and unusual plastid-targeting signals in the green-colored dinoflagellate *Lepidodinium chlorophorum*. BMC Evol Biol 10:191

Moog D, Rensing SA, Archibald JM et al (2015) Localization and evolution of putative triose phosphate translocators in the diatom *Phaeodactylum tricornutum*. Genome Biol Evol 7:2955–2969

Moore RB, Oborník M, Janouškovec J et al (2008) A photosynthetic alveolate closely related to apicomplexan parasites. Nature 451:959–963

Nash EA, Nisbet RER, Barbrook AC (2008) Dinoflagellates: a mitochondrial genome all at sea. Trend Genet 24:328–335

Nuismer S, Otto S (2004) Host-parasite interactions and the evolution of ploidy. Proc Nat Acad Sci USA 101:11036–11039

Oborník M, Green BR (2005) Mosaic origin of the heme biosynthetic pathway in phototrophic eukaryotes. Mol Biol Evol 22:2343–2353

Oborník M, Janouškovec J, Chrudimský T (2009) Evolution of the apicoplast and its host: From heterotrophy to autotrophy and back again. Int J Parasitol 39:1–12

Oborník M, Vancová M, Lai D et al (2011) Morphology and ultrastructure of multiple life stages of the photosynthetic relative of Apicomplexa, Chromera velia. Protist 162:115–130

Oborník M, Lukeš J (2013) Cell biology of chromerids, the autotrophic relatives to apicomplexan parasites. Int Rev Cell Mol Biol 306:333-369

Oborník M, Lukeš J (2015) The organellar genomes of *Chromera* and *Vitrella*, the phototrophic relatives of apicomplexan parasites. Ann Rev Microbiol 69:129–144

Oborník M, Kručinská J, Esson HJ (2016) Life cycles of chromerids resemble those of colpodellids and apicomplexan parasites. Persp Phycol 3:21–27

Oborník M, Modrý D, Lukeš M et al (2012) Morphology ultrastructure and life cycle of *Vitrella brassicaformis* n. sp., n. gen., a novel chromerid from the Great Barrier Reef. Protist 163:306–323

Okamoto N, Chantangsi C, Horák A et al (2009) Molecular phylogeny and description of the novel katablepharid *Roombia truncata* gen. et sp. nov., and establishment of the Hacrobia taxon nov. PLoS ONE 4:e7080

Petersen A, Ludewig A, 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
Pillet L (2013) The role of horizontal gene transfer in kleptoplastidy and the establishment of photosynthesis in the eukaryotes. Mob Genet Elements 3:e24773
Pombert JF, Blouin NA, Lane C et al (2014) A lack of parasitic reduction in the obligate parasitic green alga Helicosporidium. PLoS Genet 10:e1004355
Portman N, Šlapeta J (2014) The flagellar contribution to the apical complex: a new tool for the eukaryotic Swiss Army knife? Trends Parasitol 30:58–64
Ralph SA, van Dooren GG, Waller RF et al (2004) Metabolic maps and functions of the *Plasmodium falciparum* apicoplast. Nat Rev Microbiol 2:203–216. doi:10.1038/nrmicro843
Robledo JAF, Caler E, Matsuzaki M et al (2011) The search for the missing link: a relic plastid in *Perkinsus*? Int J Parasitol 41:1217–1229
Rodríguez-Ezpeleta N, Brinkmann H, Burger G et al (2007) Toward resolving the eukaryotic tree: the phylogenetic positions of jakobids and cercozoans. Curr Biol 17:1420–1425
Saldarriaga JF, Taylor FJR, Keeling PJ et al (2001) Dinoflagellate nuclear SSU rRNA phylogeny suggests multiple plastid losses and replacements. J Mol Evol 53:204–213
Saldarriaga JF, McEwan ML, Fast NM et al (2003) Multiple protein phylogenies show that *Oxyrrhis marina* and *Perkinsus marinus* are early branches of the dinoflagellate lineage. Int J System Evol Micro 53:355–365
Salomaki ED, Lane CE (2014) Are all algal parasites cut from the same cloth? Acta Soc Bot Pol 83:369–375
Sanchez-Puerta M, Delwiche C (2008) A hypothesis for plastid evolution in chromalveolates. J Phycol 44:1097–1107
Sato S, Clough B, Coates L et al (2004) Enzymes for heme biosynthesis are found in both the mitochondrion and plastid of the malaria parasite *Plasmodium falciparum*. Protist 155:117–125. doi:10.1078/1434461000169
Sato S (2011) The apicomplexan plastid and its evolution. Cell Mol Life Sci 68:1285–1296
Seeber F, Soldati-Favre D (2010) Metabolic pathways in the apicoplast of Apicomplexa. Int Rev Cell Mol Biol 281:161–228
Ševčíková T, Horák A, Klimeš V et al (2015) Updating algal evolutionary relationships through plastid genome sequencing: did alveolate plastids emerge through endosymbiosis of an ochrophyte? Sci Reports 5:10134
Slamovits CH, Keeling PJ (2008) Plastid-derived genes in the nonphotosynthetic alveolate *Oxyrrhis marina*. Mol Biol Evol 25:1297–1306
Slamovits CH (2014) Mitochondrial genomes of alveolates. In: Gray MC (ed) Mitochondrial genomes. Springer, Berlin, pp 1–6
Slamovits CH, Saldarriaga JF, Larocque A et al (2007) The highly reduced and fragmented mitochondrial genome of the early-branching dinoflagellate *Oxyrrhis marina* shares characteristics with both apicomplexan and dinoflagellate mitochondrial genomes. J Mol Biol 372:256–268
Shields JD (1994) The parasitic dinoflagellates of marine crustaceans. Ann Rev Fish Dis 4:241–271
Schoenborn HW (1949) Growth of *Astasia longa* in relation to hydrogen ion concentration. J Exp Zool 111:437–447
Schoenborn HW (1952) Studies on the nutrition of colorless euglenoids flagellates. 3. Astasia longa, A. klebsii, and Khawkinea quartana. Phys Zool 25:15–19
Stelter K, El-Sayed NM, Seeber F (2007) The expression of a plant-type ferredoxin redox system provides molecular evidence for a plastid in the early dinoflagellate *Perkinsus marinus*. Protist 158:119–130
Stiller JW, Schreiber J, Yue J et al (2014) The evolution of photosynthesis in chromist algae through serial endosymbioses. Nat Commun 5:5764
Stoebe B, Kowallik KV (1999) Gene-cluster analysis in chloroplast genomics. Trend Genet 15:344–347

Tartar A, Boucias DG, Adams BJ et al (2002) Phylogenetic analysis identifies the invertebrate pathogen *Helicosporidium* sp. as a green alga (Chlorophyta). Int J Syst Evol Microbiol 52:273–279

Tikhonenkov DV, Janouškovec J, Mylnikov A et al (2014) Description of *Colponema vietnamica* sp.n. and *Acavomonas peruviana* n. gen. n. sp., two new alveolate phyla (Colponemidia nom. nov. and Acavomonidia nom. nov.) and their contributions to reconstructing the ancestral state of alveolates and eukaryotes. PLoS ONE 9:e95467

Tyler BM, Tripathy S, Zhang X et al (2006) *Phytophthora* genome sequences uncover evolutionary origins and mechanisms of pathogenesis. Science 313:1261–1266

Valigurová A, Hofmanová L, Koudela B, Vávra J (2007) An ultrastructural comparison of the attachment sites between *Gregarina steini* and *Cryptosporidium muris*. J Euk Micro 54:495–510

van Dooren GG, Kennedy AT, McFadden GI (2012) The use and abuse of heme in apicomplexan parasites. Antioxid Redox Signal 17:634–656

Votýpka J, Modrý D, Oborník M et al (in press) Apicomplexa. In: Archibald JM, Simpson A, Slamovits C (eds) Handbook of Protists. Springer. doi:10.1007/978-3-319-32669-6_20-1

Waller RF, Keeling PJ, van Dooren GG, McFadden GI (2003) Comment on "A green algal apicoplast ancestor". Science 301:5629

Waller RF, Jackson CJ (2009) Dinoflagellate mitochondrial genomes: stretching the rules of molecular biology. BioEssays 31:237–245

Waller RF, Gornik S, Kořený L et al (2016) Metabolic pathway redundancy within the apicomplexan-dinoflagellate radiation argues against an ancient chromalveolate plastid. Commun Integr Biol 9:e1116653

Weiser J (1970) *Helicosporidium parasiticum* Keilin infection in caterpillar of hepialid moth in Argentina. J Protozool 17:436

Wisecaver JH, Hackett JD (2010) Transcriptome analysis reveals nuclear-encoded proteins for the maintenance of temporary plastids in the dinoflagellate *Dinophysis acuminata*. BMC Genom 11:366

Wolf YI, Koonin EV (2013) Genome reduction as the dominant mode of evolution. BioEssays 35:829–837

Woo YH, Ansari H, Otto TD et al (2015) Chromerid genomes reveal the evolutionary path from photosynthetic algae to obligate intracellular parasites. eLife 4:e06974

Zhang ZD, Cavalier-Smith T, Green BR (1999) Single gene circles in dinoflagellate chloroplast genomes. Nature 400:155–159

Zhang ZD, Cavalier-Smith T, Green BR (2000) Phylogeny of ultra-rapidly evolving dinoflagellate chloroplast genes: a possible common origin for sporozoan and dinoflagellate plastids. J Mol Evol 51:26–40

Zhu G, Marchewka MJ, Keithly JS (2000) *Cryptosporidium parvum* appears to lack a plastid genome. Microbiol 146:315–321

Evolution of Milk Oligosaccharides and Their Function in Monotremes and Marsupials

Tadasu Urashima and Michael Messer

Abstract The milk produced by eutherian mammals typically contains a predominance of the disaccharide lactose (Gal(β1-4)Glc) plus trace to minor levels of diverse oligosaccharides that contain lactose at their reducing ends. By contrast, monotremes and marsupials—mammalian lineages that diverged from the lineage that led to eutherians about 190 and 160 million years ago, respectively—produce milk in which diverse oligosaccharides predominate and lactose is absent or a minor constituent. In this paper, we review the evolution of milk and milk oligosaccharides in monotremes and marsupials, including new evidence on neutral and acidic milk oligosaccharides of the brushtail possum and the eastern quoll (marsupials) as well as acidic milk oligosaccharides of the echidna and platypus (monotremes). In milk of the brushtail possum, the linear series of galactosyllactoses was found to predominate over the branched oligosaccharides, as in that of the tammar wallaby and red kangaroo, in contrast to milk of the eastern quoll in which the branched oligosaccharides predominated over the linear ones. Furthermore, we found a unique doubly branched saccharide lacto-N-novooctaose, Gal(β1-3)[Gal(β1-4)GlcNAc(β1-6)]Gal(β1-3)[Gal(β1-4)GlcNAc(β1-6)]Gal(β1-4)Glc, in eastern quoll milk. This oligosaccharide has not previously been identified in the milk/colostrum of any eutherian or monotreme. In milk of the echidna and platypus, most of the acidic oligosaccharides were found to contain 4-O-acetyl N-acetylneuraminic acid (Neu4,5Ac_2), a constituent that is apparently unique to monotreme milks. We hypothesize that the presence of Neu4,5Ac_2 in acidic milk saccharides of monotremes confers resistance against hydrolysis of these saccharides by bacterial neuraminidases. This and other antimicrobial constituents in monotreme and marsupial milks appear to reflect a long evolutionary struggle to provide essential nutrients to highly immature offspring while preventing an onslaught by microbial pathogens. The milk secreted by the earliest mammals may have been especially vulnerable to microbial proliferation as it would have been secreted onto skin

T. Urashima (✉)
Obihiro University of Agriculture and Veterinary Medicine, Obihiro, Japan
e-mail: urashima@obihiro.ac.jp

M. Messer
The University of Sydney, Sydney, Australia

P. Pontarotti (ed.), *Evolutionary Biology: Self/Nonself Evolution, Species and Complex Traits Evolution, Methods and Concepts*,
DOI 10.1007/978-3-319-61569-1_13

surfaces rather than via nipples. We discuss two related phenomena that appear to be important in the evolution of milk constituents: probiotic effects favoring beneficial microorganisms, and antimicrobial effects that hinder pathogens.

Abbreviations

Glc	Glucose
Gal	Galactose
GlcNAc	*N*-acetylglucosamine
GalNAc	*N*-acetylgalactosamine
Fucose	Fucose
Neu5Ac	*N*-acetylneuraminic acid
Neu5Gc	*N*-glycolylneuraminic acid
UDP	Uridine 5'-diphosphate

1 Introduction

It is well known that mammalian milk and colostrum contain lactose as well as a numerous variety of milk oligosaccharides. The milk oligosaccharides usually contain lactose at their reducing ends to which monosaccharide residues including those of N-acetylglucosamine (GlcNAc), galactose (Gal), fucose (Fuc), N-acetylgalactosamine (GalNAc) and sialic acid (N-acetylneuraminic acid (Neu5Ac) or N-glycolylneuraminic acid (Neu5Gc)) are attached. Lactose (Gal (β1-4)Glc) is synthesized within the lactating mammary glands from UDP-Gal and glucose (Glc) by the action of lactose synthase, which is a complex of β4galactosyltransferase 1 and α-lactalbumin. Milk oligosaccharides are synthesized from lactose by the actions of several glycosyltransferases. This means that the expression of α-lactalbumin is essential for the biosynthesis of milk oligosaccharides as well as of lactose. Milk/colostrum of eutherians other than some Arctoidea species contains more lactose than oligosaccharides, while in milk of monotremes and marsupials the oligosaccharides predominate over lactose (Messer and Urashima 2002; Urashima et al. 2011, 2014a). We have previously proposed a hypothesis on the evolution of milk oligosaccharides and lactose (Messer and Urashima 2002; Urashima et al. 2011, 2014a) which is briefly described as follows.

Because α-lactalbumin, which is found only in mammary glands and milk, resembles c-type lysozyme in its primary and tertiary structures, it is believed that α-lactalbumin had evolved from lysozyme. Lysozyme is an enzyme which destroys bacteria by hydrolyzing the peptidoglycans of their cell walls. It is contained in body fluids such as tears and saliva, as well as in many non-mammalian sources including avian egg white. There are two kinds of this enzyme, only one of which has calcium bound to it. It is agreed that mutations occurred in the calcium-binding

lysozyme which resulted in the acquisition of α-lactalbumin and the loss of lysozyme activity. The timing of these mutations is still being debated.

It is proposed that in the ancestor of mammals, the protolacteal secretions had contained only lipids and proteins but no carbohydrates. When α-lactalbumin first appeared its expression level was low, therefore, lactose synthesis was slow and most of lactose produced was utilized for the synthesis of oligosaccharides. As a result, milk oligosaccharides predominated over lactose in these secretions. Initially, these oligosaccharides may have acted primarily as anti-infection agents, but during the course of evolution they came to function also as an energy source for the neonates of monotremes and marsupials. In eutherians, because of an increase in the synthesis of α-lactalbumin and therefore of lactose by their mammary glands, lactose usually came to predominate over milk oligosaccharides. Furthermore, the acquisition of small intestinal neutral lactase (lactose hydrolyzing enzyme) meant that lactose could become a significant energy source for eutherian neonates, while milk oligosaccharides continued to act as anti-infection agents by inhibiting the adhesions of pathogens to epithelial cells. For example, it was shown in a recent in vitro study that two fucosylated trisaccharides that are found in human milk inhibit the adhesion of enteropathogenic bacteria such as using *Campylobacter jejuni* and *Pseudomonas aeruginosa* to the intestinal cell line Caco-2 and to the human respiratory cell line A549 (Weichert et al. 2013). It is thought that milk oligosaccharides act as decoys, preventing the colonization of epithelial cell surfaces by mimicking the structures of cell surface receptors.

We have recently studied the milk oligosaccharides of monotremes including the Tasmanian echidna (Oftedal et al. 2014) and the platypus (Urashima et al. 2015a) and of marsupials including the common brushtail possum (Urashima et al. 2014b), the wombat (Hirayama et al. 2016), the eastern quoll (Urashima et al. 2015b), and the tiger quoll (Urashima et al. 2016). In this review, in light of our new data, we update our ideas on the evolution of milk oligosaccharides in monotremes and marsupial in relation to their life strategy.

2 Monotreme Milk Oligosaccharides

As described previously (Urashima et al. 2014a), the study of monotreme milk oligosaccharides began with the observations on milk from echidnas found on Kangaroo Island and in New South Wales and from platypuses caught in New South Wales, Australia (Messer and Kerry 1973). The structures of these oligosaccharides were characterized as shown in Fig. 3 of our previous chapter (Urashima et al. 2014a; Messer 1974; Kamerling et al. 1982; Jenkins et al. 1984; Amano et al. 1985).

In September, 2012, Tadasu Urashima joined the field work of Dr. Stewart Nicol of the University of Tasmania, for the collection of milk from the Tasmanian echidna (*Tachyglossus aculeatus setosus*). Initially, it was of interest to establish whether milk oligosaccharides of the Tasmanian echidna differ from those of the

Fig. 1 Photograph of the field work on Tasmanian echidna in September, 2012. Dr. Stewart Nichol looking for the animal which a GPS transmitter had been attached

Kangaroo Island and NSW echidnas (*Tachyglossus aculeatus*), as they are different subspecies. Figure 1 illustrates the search for the animal, to which a GPS device had been attached, in an area approximately 50 km north of Hobart. We collected the milk from the milk patch located on the skin of the abdomen of the female which had been injected intravenously with oxytocin (Fig. 2). It should be understood that in monotremes the milk is secreted onto the abdomen from two nipple-less mammary glands. In echidnas, the milk is secreted from around 100 small pores, of the milk patch, which is found in two areas of the abdomen in the lactating echidna's pouch (see Fig. 1, Urashima et al. 2014a). The echidna milk samples were collected at around 39 days (early lactation), 90 days (mid-lactation), and 150 days (late lactation) *post*-hatching. Oligosaccharides in pooled milk from late lactation were purified by gel filtration and high-performance liquid chromatography (HPLC) using a porous graphitized carbon column and characterized by ^{1}H nuclear magnetic resonance spectroscopy (^{1}H-NMR); oligosaccharides in smaller samples from early and mid-lactation were separated by ultra-performance liquid chromatography and characterized by negative electrospray ionization mass spectrum (ESI-MS) and tandem collision mass spectroscopy (MS/MS) product ion patterns (Oftedal et al. 2014).

The oligosaccharides in these milks were characterized as shown in Table 1 (Oftedal et al. 2014). The predominant oligosaccharides in early, mid, and late lactation milk was Neu4,5Ac$_2$(α2-3)Gal(β1-4)Glc (4-O-acetyl-3'-sialyllactose),

Table 1 Echidna milk saccharides (reuse Table 1 of Oftedal et al. 2014)

			Formula[b]	Observed	lactation stage[c]		
	Common name	Structure	Mass	$[M-H]^-$	Early	Mid	Late[a]
				m/z	39 d	90 d	~150 d
Neutral saccharides							
1	lactose	Gal(β1-4)Glc[d]	342.1	341.1	tr	+	+
2	2'-fucosyllactose	Fuc(α1-2)Gal(β1-4)Glc[d]	488.2	487.2	tr	+	+
3	difucosyllactose	Fuc(α1-2)Gal(β1-4)[Fuc(α1-3)]Glc[d]	634.2				+
4	B-tetrasaccharide	Gal(α1-3)[Fuc(α1-2)]Gal(β1-4)Glc	650.2	649.2	tr	+	+
5	B-pentasaccharide	Gal(α1-3)[Fuc(α1-2)]Gal(β1-4)[Fuc(α1-3)]Glc	796.3	795.3	tr	+	++
6	lacto-N-fucopentaose III	Gal(β1-4)[Fuc(α1-3)]GlcNAc(β1-3)Gal(β1-4)Glc	853.3				+
Acidic oligosaccharides							
7	3'-sialyllactose	Neu5Ac(α2-3)Gal(β1-4)Glc	633.2	632.2	+	+	
8	4-O-acetyl 3' sialyllactose	Neu4,5Ac2(α2-3)Gal(β1-4)Glc[d]	675.2	674.2	++	++	++
9	di-O-acetyl-3'-sialyllactose	Neu4,5,*U*Ac3(α2-3)Gal(β1-4)Glc; *U* = 7, 8 or 9	717.2	716.2	+	+	
10	4-O-acetyl-3'-sialyllactose sulfate	[as 8 above, plus sulfate of unknown position]	755.3	754.3	+	+	
11	unknown, 4-O-acetyl 3' sialyllactose core	[as 8 above, plus unknown nitrogen-containing group]	*800.3*	799.3		+	
12	monofucosyl-4-O-acetyl-3'-sialyllactose	Neu4,5Ac2(α2-3)Gal(β1-4)[Fuc(α1-3)]Glc	821.3	820.3	+	+	+
Other oligosaccharides of known mass but unknown structure							
A	unknown		*620.3*	619.3		+	
B	unknown sialylated oligosaccharide		*673.3*	672.3	+		
C	unknown sialylated oligosaccharide		*773.3*	772.3	+	+	

[a]Saccharides as identified by UPLC/MS and/or ^{1}H-NMR

[b]Formula mass and observed m/z for most abundant isotope, i.e., ^{1}H, ^{12}C, ^{14}N, ^{16}O, ^{32}S

[c]Estimated age of offspring given in days post-partum; abundance of each oligosaccharide indicated as tr = trace, + = present, ++ = abundant

[d]Publications that have previously identified these peaks in echidna milk: 1. Messer and Kerry 1973; 2. Messer 1974; 3. Jenkins et al. 1984

while that in late lactation was Gal(α1-3)[Fuc(α1-2)]Gal(β1-4)[Fuc(α1-3)]Glc (B-pentasaccharide) in addition to 4-O-acetyl-3'-sialyllactose. B-pentasaccharide and Gal(α1-3)[Fuc(α1-2)]Gal(β1-4)Glc (B-tetrasaccharide) had not been identified in the milk of Kangaroo Island and New South Wales echidnas (Messer and Kerry 1973; Jenkins et al. 1984), while Fuc(α1-2)Gal(β1-4)Glc (2'-fucosyllactose) and Fuc(α1-2)Gal(β1-4)[Fuc(α1-3)]Glc (difucosyllactose) had been found in these milks. 2'-fucosyllactose and difucosyllactose can be assumed to be acceptors for α1-3galactosyltransferase, which synthesizes B-tetra and B-pentasaccharides. It would seem that the activity of this enzyme is present in the lactating mammary glands of Tasmanian echidnas but is absent from those of Kangaroo Island and New South Wales echidnas, presumably because of loss expression of this enzyme during the course of evolution. The milk contained a small amount of Gal(β1-4)[Fuc(α1-3)]GlcNAc(β1-3)Gal(β1-4)Glc (lacto-N-fucopentaose III), whose core structure is Gal(β1-4)GlcNAc(β1-3)Gal(β1-4)Glc (lacto-N-neotetraose). It is noteworthy that the milk contained di-O-acetyl-3'-sialyllactose and 4-O-acetyl-3'-sialyllactose sulfate, but their concentrations were very small. As in the previous study (Messer and Kerry 1973), free lactose was no more than a minor saccharide in milk of the Tasmanian echidna.

As the neutral platypus milk oligosaccharides had been characterized by Amano et al. (1985), our study was focused on the acidic oligosaccharides of platypus milk. The carbohydrate fraction extracted from the milk was subjected to gel filtration on

Fig. 2 Photograph showing the collection of milk from the milk patch of a lactating echidna which was injected intravenously with oxytocin

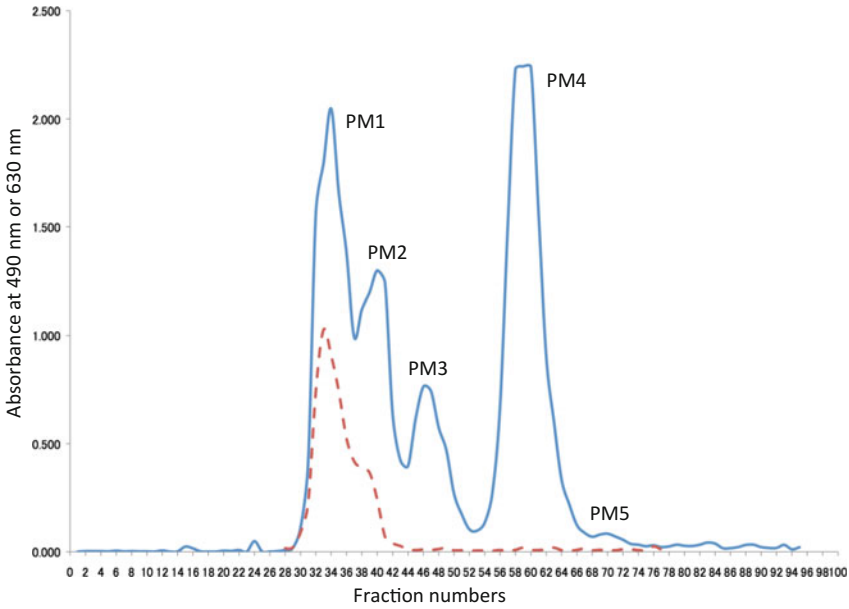


Fig. 3 Gel chromatogram of the carbohydrate fraction from platypus milk on a BioGel P-2 column (2.5 × 100 cm) (reuse Fig. 2 of Urashima et al. 2015a with permission). Elution was done with distilled water at a flow rate of 15 mL/h, and fractions of 5.0 mL were collected. Each fraction was monitored for hexose by the phenol-H_2SO_4 method (*solid line*) and for sialic acid by the periodate–resorcinol (*dotted line*). Acidic oligosaccharides were contained in Peaks PM1 and PM2. Peaks PM3, PM4, and PM5 were found to only contain neutral oligosaccharides

BioGel P-2 column, as shown in Fig. 3, and the fractions containing sialic acid (PM-1 and PM-2) were further fractionated using normal phase HPLC with an Amide-80 column (Urashima et al. 2015a). The fraction designated PM-5 contained lactose, which was shown to be a minor saccharide. The oligosaccharides in each peak were characterized using ^{1}H-NMR and matrix-assisted laser desorption/ionization time-of-flight mass spectrum (MALDI TOFMS). The resulting structures are shown in Fig. 4 (Urashima et al. 2015a). It was concluded that the core units of the oligosaccharides are lactose, lacto-N-neotetraose, or lacto-N-neohexaose (Gal(β1-4)GlcNAc(β1-3)[Gal(β1-4)GlcNAc(β1-6)]Gal(β1-4)Glc) and the structures contained Lewis x (Gal(β1-4)[Fuc(α1-3)]GlcNAc), Lewis y (Fuc(α1-2)Gal(β1-4)[Fuc(α1-3)]GlcNAc), or sialyl Lewis x (in this case, Neu4,5Ac_2(α2-3)Gal(β1-4)[Fuc(α1-3)]GlcNAc). The presence of Lewis x or Lewis y is consistent with the neutral platypus milk oligosaccharides characterized by Amano et al. (1985). The most significant feature is the presence of Neu4,5Ac_2 in most of the acidic platypus milk oligosaccharides. Although an O-acetyl Neu5Ac has been detected in bovine colostrum (Marino et al. 2011), this saccharide was present at very low concentration and the position of its O-acetyl group was not characterized. Thus, the predominant acidic oligosaccharide in the milk of echidnas

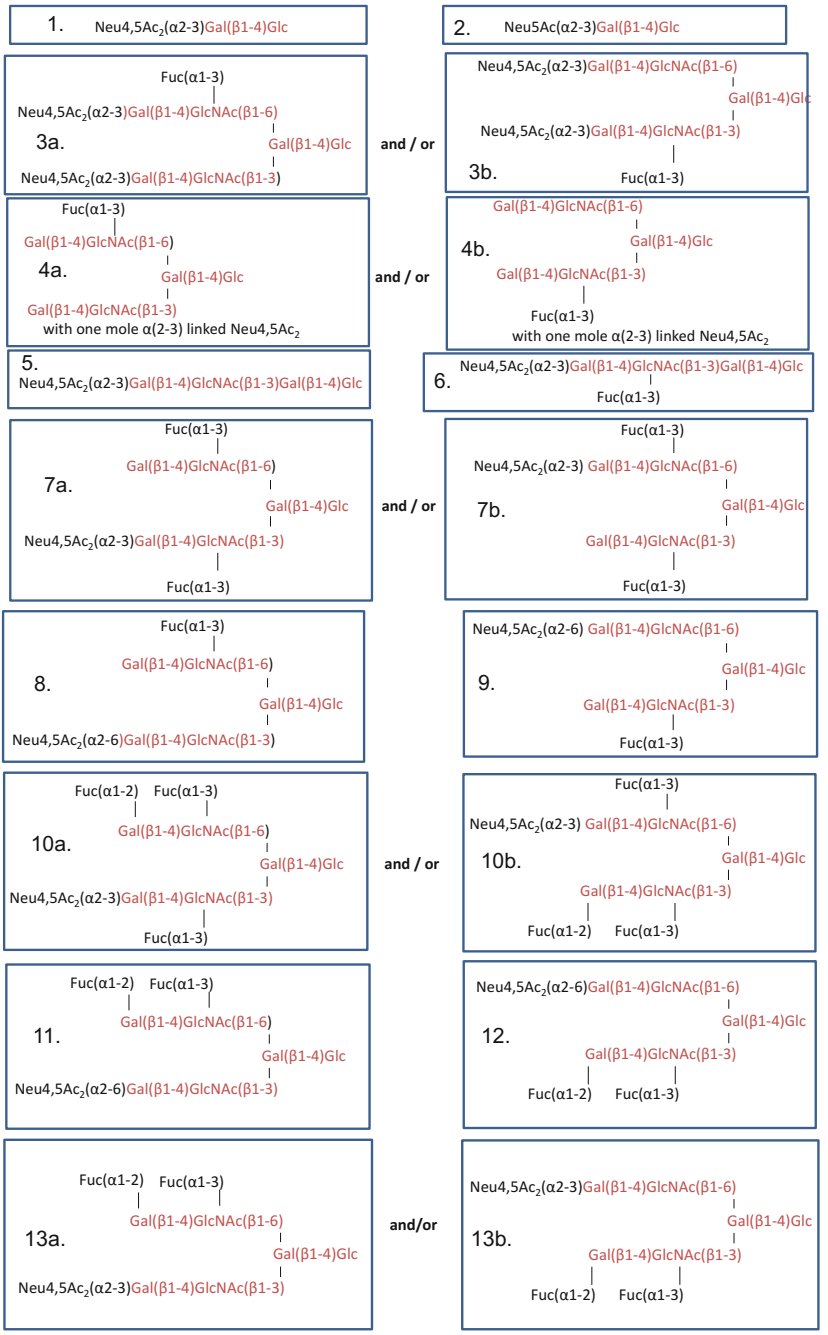


Fig. 4 Structures of acidic oligosaccharides of platypus milk as characterized by ^{1}H-NMR and MALDI TOFMS (reuse Fig. 10 of Urashima et al. 2015a with permission). Oligosaccharides are designated with numbers 1–13 to indicate the minimal number of unique oligosaccharides; where one or more alternative structures are possible these are indicated. Oligosaccharide 4 has additional potential structures depending on placement of Neu4,5Ac$_2$ (not illustrated). Thus, the actual number of acidic oligosaccharides in platypus milk may be ≥ 10

found in Tasmania, Kangaroo Island, and New South Wales (Oftedal et al. 2014; Messer 1974; Kamerling et al. 1982) as well as in platypus milk is Neu4,5Ac_2. Since this unusual sialic acid has not so far been found in the milk of other species, it appears to be specific to and characteristic of the milk of monotremes.

The possible biological significance of Neu4,5Ac_2 in monotreme milk oligosaccharides.

What is the biological significance of Neu4,5Ac_2 in the oligosaccharides in monotreme milk? We suggest that 4-O-acetylation of N-acetylneuraminic acid in acidic oligosaccharides may function to protect monotreme milk from microbial attack. Monotremes secrete milk directly into the infundibula of mammary hair follicle (Griffiths 1978; Oftedal 2002), from which it spreads onto the areolar skin surface and onto mammary hairs within the pouch (in echidna) or incubatorium (platypus). A moist, warm, aerobic surface richly endowed with milk nutrients would be seen an ideal site for the culture of microbial pathogens (see Fig. 5), but if the predominant saccharide is unavailable (due to 4-O-acetylation), microbial growth may be curtailed. Oftedal et al. (2014) regarded the predominance of

Fig. 5 Photograph of the pouch area of a lactating Tasmanian echidna, showing a 6-day-old hatchling adjacent to the areolae or milk patches. These swollen mammary glandular areas which from the borders of the pouch can be clearly seen on each side of the young. Note that milk is visible in the stomach of the young and that moisture and debris particles are evident on the skin, hair, and hatchling, all of which are in contact. Australian 5 cent coin (2 cm diameter) provided for scale. Photograph taken August 20, 2007 by Stewart Nicol (reuse Fig. 1 of Oftedal et al. 2014 with permission)

4-O-acetyl-3'-sialyllactose in echidna milk as evolutionarily significant, given that exposed skin and hair surfaces—as well as the oral cavity and digestive tracts of hatchlings—should be easily colonized by both commensal and pathogenic bacteria. 4-O-Acetylation can be considered to be a lock-up mechanism whereby sialic acid is rendered unavailable to bacterial and mammalian sialidases, apparently due to steric interference with binding to these enzymes (Schauer et al. 2011); catabolism requires a sialate-O-esterase to initially remove the 4-O-acetyl moiety. It is likely that milk secretion onto skin and hair long predated the evolution of nipples (Oftedal and Dhouailly 2013); therefore, there may have been strong selective pressures over a long evolutionary period to deny bacteria access to oligosaccharides in mammary secretions, and this may underlie the evolutionary development of 4-O-acetylation of sialic acid in monotreme oligosaccharides.

4-O-Acetyl 3'-sialyllactose or the oligosaccharides containing 4-O-acetylated Neu5Ac predominate in milk of the echidna or platypus, lactose being present only in small amounts. If they can be digested and absorbed, these oligosaccharides may be critical sources of preformed glucose and sialic acid for the highly altricial young of monotremes (Oftedal et al. 2014). However, 4-O-acetylation of sialic acid blocks mammalian sialidases, unless a sialate-O-esterase initially removes the 4-O-acetyl moiety, as occurs in the liver of the horse (Schauer and Shukla 2008). It may be that 4-O-acetyl sialyloligosaccharides can be taken up into the enterocytes and, after transport to lysosomes, cleaved to their monosaccharide constituents by the sequential actions of esterases, sialidases, and β-galactosidase (Duncan et al. 2009). This has not been studied in the platypus, but it is intriguing that intestinal mucosal homogenates of suckling echidnas can hydrolyze 4-O-acetyl 3'-sialyllactose to lactose, Neu5Ac, Gal, and Glc with intermediate formation of 3'-sialyllactose, even though comparable homogenates from suckling rats achieve no hydrolysis of this substrate (Stewart et al. 1983). This suggests that suckling monotremes may have evolved the ability to utilize 4-O-acetyl-sialo-oligosaccharides, presumably by expression of a 4-O-acetylesterase; if so, this may have been important in the early evolutionary conflict between microbes and milk-fed mammals. In summary, although milk oligosaccharides containing 4-O-acetyl N-acetylneuraminic acid cannot be utilized by bacteria located on the milk patch, it is likely that they can be hydrolyzed by lysosomal enzymes located within monotreme small intestinal cells, which enables monotreme young to utilize them as significant nutrients.

3 Marsupial Milk Oligosaccharides

The milk oligosaccharides of marsupial species that had been studied including the tammar wallaby (Messer et al. 1980, 1982; Collins et al. 1981; Bradbury et al. 1983), red kangaroo (Anraku et al. 2012), and koala (Urashima et al. 2013). The structures of tammar wallaby neutral and red kangaroo acidic oligosaccharides were described in detail in our previous chapter (see Figs. 5 and 11, Urashima et al. 2014a). Since 2013, we have characterized the oligosaccharides in milk of the

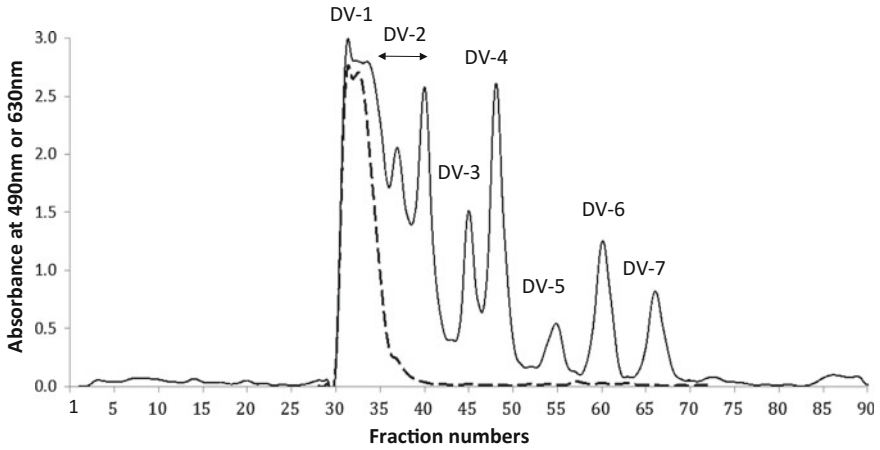
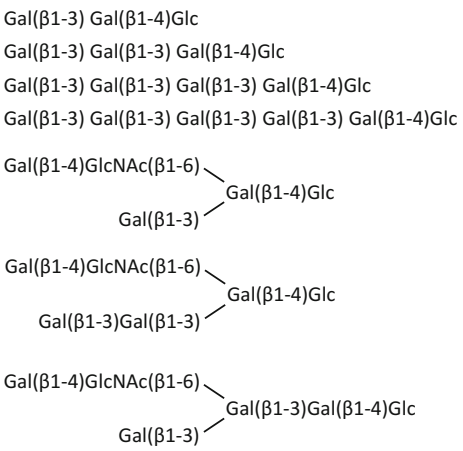


Fig. 6 Gel chromatogram of the carbohydrate fraction from eastern quoll milk on a BioGel P-2 column (2.5 × 100 cm) (reuse Fig. 6a of Urashima et al. 2015b with permission). Each fraction was monitored by the phenol-H_2SO_4 method (*solid line*) and the periodate–resorcinol method (*dotted line*)

common brushtail possum (Urashima et al. 2014b), eastern quoll (Urashima et al. 2015b), wombat (Hirayama et al. 2016), and tiger quoll (Urashima et al. 2016). The carbohydrates extracted from these milks were separated by gel filtration on BioGel P-2. Figure 6 shows the chromatogram obtained with the carbohydrate from eastern quoll milk (lactation: 7–11 weeks post-partum). The acidic oligosaccharides in the early eluted peak fractions, which reacted positively with periodate–resorcinol as well as with phenol–sulfuric acid, were further separated by ion exchange chromatography on DEAE—Sephadex A-50, after which each oligosaccharide was purified by normal phase HPLC using an Amide-80 column. Each of the neutral oligosaccharide, which eluted subsequent to the acidic saccharides on BioGel P-2 and reacted only with phenol–sulfuric acid but not with periodate–resorcinol, was purified from the peak fractions using HPLC with a porous graphitized carbon column. Both the neutral and the acidic milk oligosaccharides were characterized by ^{1}H-NMR and MALDI TOFMS spectroscopies.

The structures of the characterized brushtail possum neutral and acidic milk oligosaccharides are shown in Figs. 7 and 8, respectively (lactation: 60–147 days post-partum) (Urashima et al. 2014b). Lactose was a minor saccharide in the carbohydrate fraction. It was shown that the brushtail possum neutral oligosaccharides were similar to those of the tammar wallaby (see also Fig. 5, Urashima et al. 2014a). Gal(β1-3)[Gal(β1-4)GlcNAc(β1-6)]Gal(β1-3)Gal(β1-4)Glc (galactosyl lacto-N-novopentaose II) was found as a novel oligosaccharide, but it seems likely that this saccharide would also be present in the unidentified fraction of tammar wallaby milk carbohydrate. The core structures of the acidic oligosaccharides were similar to those of the neutral saccharides, with Neu5Ac or sulfate attached to the penultimate or non-reducing Gal residue at OH-3. The Neu5Ac was found to be



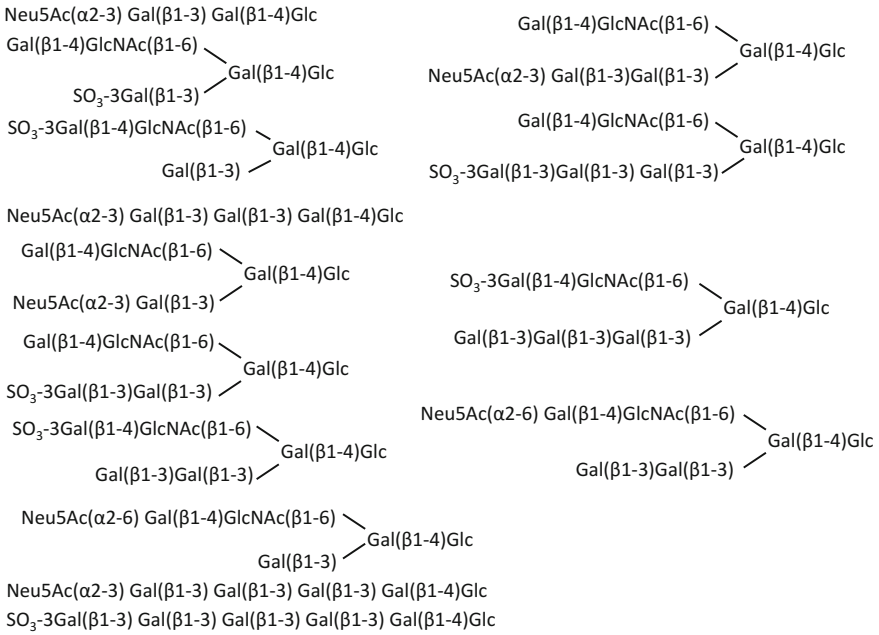


Fig. 7 Structures of the neutral oligosaccharides of brushtail possum milk (reuse Fig. 10 of Urashima et al. 2014b with permission)

Fig. 8 Structures of the acidic oligosaccharides of brushtail possum milk (reuse Fig. 11 of Urashima et al. 2014b with permission)

Fig. 9 Photograph of an eastern quoll taken by Dr. Jim Merchant

attached to this Gal only via an α(2-3) linkage, not via an α(2-6) linkage. It is noteworthy that the brushtail possum milk oligosaccharides contain significant proportions of sulfate, suggesting that the neonates may require sulfate for their development. The compositions and structures of brushtail possum acidic oligosaccharides are similar to those of red kangaroo (see also Fig. 11, Urashima et al. 2014a).

Figure 9 is a photograph of an eastern quoll, while Fig. 10 shows the structures of each of the neutral and acidic eastern quoll milk oligosaccharide (Urashima et al. 2015b). Lactose (fraction DV-7 in Fig. 6) was a minor saccharide in the carbohydrate fraction. It was shown that the branched units of Gal(β1-4)GlcNAc (N-acetyllactosamine) are attached to galactosyllactose (Gal(β1-3)Gal(β1-4)Glc) or digalactosyllactose (Gal(β1-3)Gal(β1-3)Glc(β1-4)Glc). Although Gal(β1-3)[Gal(β1-4)GlcNAc(β1-6)]Gal(β1-3)[Gal(β1-4)GlcNAc(β1-6)]Gal(β1-4)Glc (lacto-N-novooctaose) and its sialyl derivatives, and Gal(β1-3)[Gal(β1-3)Gal(β1-4)GlcNAc(β1-6)]Gal(β1-4)Glc (galactosyl lacto-N-novopentaose III) were identified as novel saccharides, it is possible that these were present in the unidentified fractions of the milk carbohydrates of other marsupial species. In fact, lacto-N-novooctaose was found in milk of the wombat (Hirayama et al. 2016) and tiger quoll (Urashima et al. 2016), too.

Comparison of the structures of the characterized milk oligosaccharides suggests notable differences between those of the brushtail possum and the eastern quoll. Brushtail possum milk contained a major series of linear β(1-3)galactosyllactose series and a minor series of branched saccharides containing β(1-6) linked GlcNAc, a pattern which had previously been observed in the milk of the tammar wallaby and of the red kangaroo. In the carbohydrate fraction of eastern quoll milk, however,

Fig. 10 Structures of the oligosaccharides of eastern quoll milk (Urashima et al. 2015b with permission)

branched saccharides such as lacto-N-novopentaose 1 and lacto-N-novooctaose predominated over the linear series of oligosaccharides. The activities of the following glycolsyltransferases have been observed in lactating tammar wallaby mammary glands: β4galactosyltransferase, which transfer Gal from UDP-Gal to Glc or GlcNAc, β3galactosyltransferase, which transfer Gal to non-reducing Gal of lactose or 3'-galactosyllactose, and β6 N-acetylglucosaminyltransferase, which transfer GlcNAc from UDP-GlcNAc to penultimate Gal of 3'-galactosyllactose or 3',3"-digalactosyllactose, etc. (Messer and Nicholas 1991; Urashima et al. 1992). Assuming that similar transferase activities occur in lactating mammary glands of the eastern quoll, the possible biosynthetic pathway of these milk oligosaccharides is shown in Fig. 11. It can be hypothesized that the differences of milk oligosaccharide patterns between the brushtail possum and the eastern quoll are due to the relative higher activities of the β3galactosyltransferase in lactating mammary glands of the brushtail possum and higher activities of the β6 N-acetylglucosaminyltransferase in those of the eastern quoll.

We have recently found that milk of the tiger quoll surprisingly contains large oligosaccharides with three branches of Gal(β1-4)GlcNAc as shown in Fig. 12 (Urashima et al. 2016). It seems likely that such megasaccharides are contained in milk of some other marsupial species.

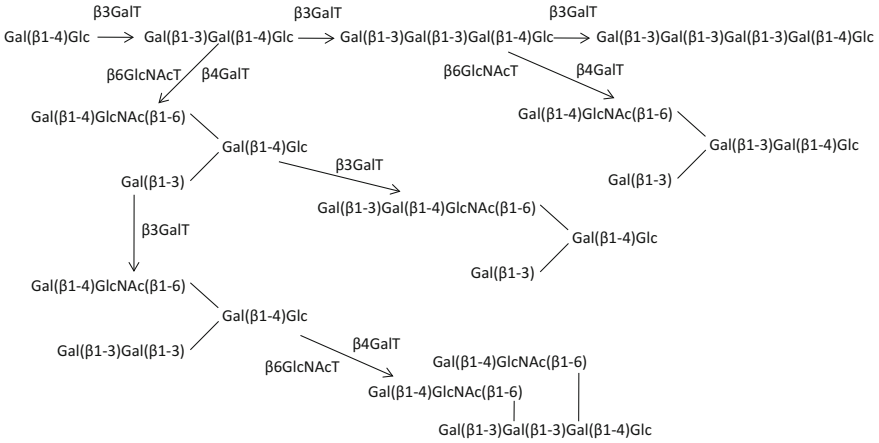


Fig. 11 Possible biosynthetic pathway of eastern quoll milk oligosaccharides (reuse Fig. 6 of Urashima et al. 2015b with permission)

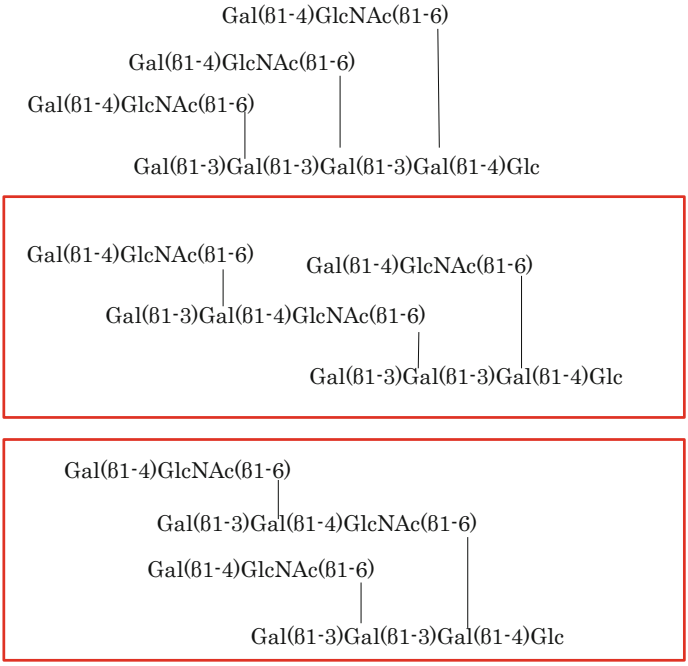


Fig. 12 Three possible structures of the oligosaccharides containing three branches of N-acetyllactosamine, separated from tiger quoll milk. The two structures enclosed in frame are considered to be more plausible than the third one (reuse Fig. 7 of Urashima et al. 2016 with permission)

4 Biological Significance of Marsupial Milk Oligosaccharides

It has been shown that milk of the tammar wallaby contains as much as 14% of oligosaccharides at one lactation stage (Messer and Green 1979). It can be calculated that milk containing 14% of lactose would exert a very large osmotic effect within the small intestine, leading to intense diarrhea, bloating, and dehydration, which are symptoms of lactose intolerance. The high molecular weight of the oligosaccharides as found in the milk of the tammar and other marsupials therefore enables these animals to secrete milk that contains a much higher concentration of carbohydrate than is normally found in the milk of eutherians. In addition, it permits the milk to contain a higher concentration of other osmotically active constituents such as electrolytes. It can reasonably be presumed that higher concentrations of carbohydrate and electrolytes in the milk are nutritionally advantageous to marsupial sucklings.

Most of the marsupial milk oligosaccharides, which range in size from trisaccharide to at least octasaccharides, contain a high proportion of galactose, the nutritional significance of which is still unclear. It can be assumed, however, that the oligosaccharides function mainly as an energy source for the young. The mechanism by which the oligosaccharides are digested and absorbed is still unclear, but it seems likely that they are not hydrolyzed to monosaccharides at the microvillous brush border of the enterocytes, as in eutherians, but instead are taken up into the enterocytes by pinocytosis or endocytosis and are then hydrolyzed to monosaccharides by the action of lysosomal β-galactosidase and other glycosidases, after which they enter the circulation (Messer and Urashima 2002; Urashima et al. 2011).

This lysosomal digestion of milk oligosaccharides, although still hypothetical, is supported by the following observations. Histochemical studies with specific stains for neutral and acid β-galactosidases showed that activity of intestinal neutral lactase (lactose hydrolyzing enzyme) is completely absent from the brush borders of the villi of the small intestine of the suckling tammar wallaby (Wolcott and Messer 1980; Messer et al. 1989). Instead, a very active acid β-galactosidase is present intracellularly, probably located in the lysosomes and supranuclear vacuoles of the enterocytes. This β-galactosidase is able to digest both lactose and the β(1-3)linked galactosides that are found in tammar wallaby milk, in contrast to the neutral brush border lactase of eutherians which splits only lactose. Additional biochemical experiments showed that the neutral lactase is absent also in the red and the gray kangaroos (Messer et al. 1989). On the other hand, the activities of several acid glycosidases including β-galactosidases, N-acetylglucosaminidase, α-L-fucosidase, and neuraminidase could readily be detected in tammar wallaby small intestine (Walcott and Messer 1980). These results indicate, therefore, that the oligosaccharides of tammar milk cannot be hydrolyzed at the microvillous membrane of the small intestine; to be digested, they must first be transported into the enterocytes where they can be hydrolyzed to their constituent monosaccharides by the

lysosomal glycosidases. Thus, the mechanism by which marsupial young digest and absorb milk carbohydrates appears to be very different from that found for lactose in eutherian mammals, and it may be of interest that these observations can explain the susceptibility to lactose intolerance of orphaned marsupials that are bottle-fed with milk that contains lactose instead of oligosaccharides, such as cow milk (Messer et al. 1989).

Hypothetic lysosomal digestion of milk oligosaccharides in monotremes and marsupial neonates by the analogous mechanism as autophagy.

With respect to monotremes, biochemical experiments similar to those done with the small intestine of the tammar wallaby (Walcott and Messer 1980) were done with small intestines of two suckling echidnas. The results showed that neutral lactase activity was entirely absent, but the activities of various acid glycosidases including β-galactosidase, α-L-fucosidase, and neuraminidase could be detected (Stewart et al. 1983). These results were consistent with the concept that the oligosaccharides of echidna milk are digested by lysosomal enzymes contained within the enterocytes of the small intestine.

5 Conclusion

In this paper, we discuss the biological significance of the characteristic ratio of milk oligosaccharides to lactose in the milk of monotreme and marsupial species and also of the presence of a specific 4-O-acetylated N-acetylneuraminic acid in monotreme milk.

We hypothesize that the ratio of milk oligosaccharides to lactose depends on the expression levels of α-lactalbumin and glycosyltransferases in lactating mammary glands (Messer and Urashima 2002; Urashima et al. 2011, 2014a). What regulates these expression levels? It is possible that the uncoded region of the genome but not the genes themselves that may regulate the expression levels of α-lactalbumin and also of glycosyltransferases. The low ratio of lactose to oligosaccharides in monotremes, marsupials, and a few eutherians is hypothesized to be caused by a low level of expression of α-lactalbumin in the lactating mammary glands among these species, but the genomic basis of this is unknown.

Based on the absence of neutral lactase activity in the small intestine of the young of the echidna, a monotreme, and of the tammar wallaby, a macropod marsupial, we hypothesize that the young of these species do not depend on milk lactose as an energy source (Messer and Urashima 2002; Urashima et al. 2011, 2014a, b). However intestinal neutral lactase activity was found late during suckling in young of the brushtail possum, a non-macropod marsupial (Crisp et al. 1989; Messer and Urashima 2002; Urashima et al. 2011, 2014a). This suggests that the neutral lactase gene may have been present in the common ancestor of marsupials and eutherians but was lost in macropods. The genomes of marsupials and monotremes may reveal when the acquisition of this lactase occurred but so far only few of these genomes have been published. It is also possible that the neutral lactase

gene was present in the genome of the common ancestor of monotremes, marsupials, and eutherians but its expression ceased in the young of monotremes and macropod marsupials. It is also possible that regulation of the expression of lactase is regulated within the non-coding region of the genomes of these mammals.

Although milk oligosaccharides containing 4-O-acetylated N-acetylneuraminic acid have been found only in monotremes, Neu4,5Ac_2 has been found in the glycoconjugates of the tissues of some eutherians. This represents a small fraction (<1%) of the Sia in brain tissue in mice (including embryos and suckling young), but a substantial portion of the Sia in glycoproteins (16%) and glycolipids (27%) of the digestive tract (Rinninger et al. 2006). The 4-O-acetylation of N-acetylneuraminic acid is likely to be catalyzed by a 4-O-acetyltransferase, which has been isolated from the Golgi apparatus of guinea pig liver (Iwersen et al. 2003). The question of what regulates the expression of this transferase in the lactating mammary glands of monotremes, and also of the non-expression in the glands of marsupials and eutherians, may perhaps be clarified by mammalian genome information.

References

Amano J, Messer M, Kobata A (1985) Structures of the oligosaccharides isolated from milk of the platypus. Glycoconj J 2:121–135

Anraku T, Fukuda K, Saito T, Messer M, Urashima T (2012) Chemical characterization of acidic oligosaccharides in milk of the red Kangaroo (*Macropus rufus*). Glycoconj J 29:147–156

Bradbury JH, Collins JG, Jenkins GA, Trifonoff E, Messer M (1983) ^{13}C-NMR study of the structures of two branched oligosaccharides from marsupial milk. Carbohydr Res 122:327–331

Collins JG, Bradbury JH, Trifonoff E, Messer M (1981) Structures of four new oligosaccharides from marsupial milk, determined mainly by ^{13}C-NMR spectroscopy. Carbohydr Res 92:136–140

Crisp EA, Messer M, Cowan PE (1989) Intestinal lactase (β-galactosidase) and other disaccharidase activities of suckling and adult common brushtail possums, *Trichosurus vulpecula* (Marsupialia: Phalangeridae). Reprod Fertil Dev 1:315–324

Duncan PI, Raymond F, Fuerholz A, Sprenger N (2009) Sialic acid utilisation and synthesis in the neonatal rat revisited. PLoS ONE 4(12):e8241

Griffiths M (1978) The biology of the monotremes. Academic Press, New York

Hirayama K, Taufik E, Kikuchi M, Nakamura T, Fukuda K, Saito T, Newgrain K, Green B, Messer M, Urashima T (2016) Chemical characterization of milk oligosaccharides of the common wombat (*Vombatus ursinus*). Ani Sci J 87:1167–1177

Iwersen M, Dora H, Kohla G, Gasa S, Schauer R (2003) Solubilisation and properties of the sialate-4-O-acetyltransferase from guinea pig liver. Biol Chem 384:1035–1047

Jenkins GA, Bradbury JH, Messer M, Trifonoff E (1984) Determination of the structures of fucosyl-lactose and difucosyl-lactose from the milk of monotremes, using ^{13}C-n.m.r. spectroscopy. Carbohydr Res 126:157–161

Kamerling JP, Dorland L, van Halbeek H, Vliegenthart JEG, Messer M, Schauer R (1982) Structural studies of 4-O-acetyl-α-N-acetylneuraminyl-(2,3)-lactose, the main oligosaccharide in echidna milk. Carbohydr Res 100:331–340

Marino K, Lane JA, Abrahams JL, Struwe WB, Harvey DJ, Marotta M, Hickey RM, Rudd PM (2011) Method for milk oligosaccharide profiling by 2-aminobenzamide labeling and hydrophilic interaction chromatography. Glycobiology 21:1317–1330

Messer M, Kerry K (1973) Milk carbohydrates of the echidna and the platypus. Science 180:201–203

Messer M (1974) Identification of N-acetyl-4-O-acetylneuraminyl-lactose in echidna milk. Biochem J 139:415–420

Messer M, Green B (1979) Milk carbohydrates of marsupial II. Quantitative and qualitative changes in milk carbohydrates during lactation in the tammar wallaby (*Macropus eugenii*). Aust J Biol Sci 32:415–420

Messer M, Trifonoff E, Stern W, Collins JG, Bradbury JH (1980) Structure of a marsupial milk trisaccharide. Carbohydr Res 83:327–334

Messer M, Trifonoff E, Collins JG, Bradbury JH (1982) Structure of a branched tetrasaccharide from marsupial milk. Carbohydr Res 102:316–320

Messer M, Crisp EA, Crolij R (1989) Lactose digestion in suckling macropodids. In: Grigg G, Jarman P, Hume I (eds) Kangaroos, Wallabies and Rat Kangaroos. Surry Beatty & Sons Pty Ltd, NSW, Australia, pp 217–221

Messer M, Nicholas KR (1991) Biosynthesis of marsupial milk oligosaccharides: characterization and developmental changes of two galactosyltransferases in lactating mammary glands of the tammar wallaby, *Macropus eugenii*. Biochim Biophys Acta 1077:79–85

Messer M, Urashima T (2002) Evolution of milk oligosaccharides and lactose. Trends Glycosci Glycotech 14:153–176

Oftedal OT (2002) The mammary gland and its origin during synapsid evolution. J Mammary Gland Biol Neoplasia 7:225–252

Oftedal OT, Dhouailly D (2013) Evo-Devo of the mammary gland. J Mammary Gland Biol Neoplasia 18:105–120

Oftedal OT, Nicol SC, Davies NW, Sekii N, Taufik E, Fukuda K, Saito T, Urashima T (2014) Can an ancestral condition for milk oligosaccharides be determined? Evidence from the Tasmanian echidna (*Tachyglossus aculeateus setosus*). Glycobiology 24:826–839

Rinninger A, Richet C, Pons A, Kohla G, Schauer R, Bauer HC, Zanetta JP, Vlasak R (2006) Localisation and distribution of O-acetylated N-acetylneuraminic acids, the endogenous substrates of the hemagglutinin-esterases of murine corona-viruses, in mouse tissue. Glycoconj J 23:73–84

Schauer R, Shukla AK (2008) Isolation and propertyes of two sialate-O-acetylesterases from horse liver with 4- and 9-O-acetyl specificities. Glycoconj J 25:625–632

Schauer R, Srinivasan GV, Wipfler D, Kniep B, Schwartz-Albiezet R (2011) O-Acetylated sialic acids and their role in immune defense. Adv Exp Med Biol 705:525–548

Stewart IM, Messer M, Walcott PJ, Gadiel PA, Griffiths M (1983) Intestinal glycosidase activities in one adult and two suckling echidnas: Absence of a neutral lactase (β-D-galactosidase). Aust J Biol Sci 36:139–146

Urashima T, Messer M, Bubb WA (1992) Biosynthesis of marsupial milk oligosaccharides II: characterization of a β^6-N-acetylglucosaminyltransferase in lactating mammary glands of the tammar wallaby, *Macropus eugenii*. Biochim Biophys Acta 1117:223–231

Urashima T, Fukuda K, Messer M (2011) Evolution of milk oligosaccharides and lactose: A hypothesis. Animal 6:369–374

Urashima T, Taufik E, Fukuda R, Nakamura T, Fukuda K, Saito T, Messer M (2013) Chemical characterization of milk oligosaccharides of the koala (*Phascolarctos cinereus)*. Glycoconj J 30:801–811

Urashima T, Messer M, Oftedal OT (2014a) Comparative biochemistry and evolution of milk oligosaccharides of monotremes, marsupials, and eutherians. In: Pontarotti P (ed) Evolutionary biology: genome evolution, speciation, coevolution and origin of life. Springer, Switzerland, pp 3–33

Urashima T, Fujita S, Fukuda K, Nakamura T, Saito T, Cowan P, Messer M (2014b) Chemical characterization of milk oligosaccharides of the common brushtail possum (*Trichosurus vulpecula*). Glycoconj J 31:387–399

Urashima T, Inamori H, Fukuda K, Saito T, Messer M, Oftedal OT (2015a) 4-O-Acetyl-sialic acid (Neu4,5Ac_2) in acidic milk oligosaccharides of the platypus (*Ornithorhynchus anatinus*) and its evolutionary significance. Glycobiology 25:683–697

Urashima T, Sun Y, Fukuda K, Hirayama K, Taufik E, Nakamura T, Saito T, Merchant J, Green B, Messer M (2015b) Chemical characterization of milk oligosaccharides of the eastern quoll (*Dasyurus viverrinus*). Glycoconj J 32:361–370

Urashima T, Yamamoto T, Hirayama K, Fukuda K, Nakamura T, Saito T, Newgrain K, Merchant J, Green B, Messer M (2016) Chemical characterization of milk oligosaccharides of the tiger quoll (*Dasyurus maculates*), a marsupial. Glycoconj J 33:797–807

Walcott PJ, Messer M (1980) Intestinal lactase (β-galactosidase) and other glycosidase activities in suckling and adult tammar wallaby (*Macropus eugenii*). Aust J Biol Sci 33:521–530

Weichert S, Jennewein S, Hufner E, Weiss C, Borkowski J, Putze J, Schroten H (2013) Bioenginnered 2'-fucosyllactose and 3-fucosyllactose inhibit the adhesion of *Pseudomonas aeruginosa* and enteric pathogens to human intestinal and respiratory cell lines. Nutr Res 33:831–838

Modelling the Evolution of Dynamic Regulatory Networks: Some Critical Insights

Anton Crombach

Abstract Regulatory networks are at the centre of cellular decision-making, and understanding their structure and dynamics is a goal in many areas of the life sciences. In this chapter, I present recent studies that, in my opinion, demonstrate how thinking in terms of networks is enriching our understanding of evolution. By studying abstract models of regulatory networks, evolutionary concepts such as robustness, evolvability, modularity, and hierarchy have been clarified. Moreover, models are helping us probe the relationship between network structure, function, and evolution. Models can also be closely linked to experimental systems. I discuss two such data-driven studies, which highlight how combining theory and data allows us to understand how evolution has proceeded on Earth. Finally, I present several open challenges in the field of network evolution, and I suggest how to tackle them.

1 Introduction

Regulatory networks of genes, RNAs, proteins, and other (cellular) components enable the organism to function and adapt, both in the short run—from minutes to hours—and across generations, over evolutionary time scales. They are increasingly accepted as a general property of living systems, which makes it important to understand them. While network thinking in biology dates back to the 1970s (Britten and Davidson 1969; Glass and Kauffman 1973; Kauffman 1969), for a long time it was an exercise in the analysis of abstract computational and mathematical models. Networks only became widely studied with the availability of ubiquitous computational power and genome-scale experimental data around the turn of the century (O'Malley 2012).

A. Crombach (✉)
Centre for Interdisciplinary Research in Biology, College de France, CNRS, INSERM, PSL Research, Paris, France
e-mail: anton.crombach@college-de-france.fr

P. Pontarotti (ed.), *Evolutionary Biology: Self/Nonself Evolution, Species and Complex Traits Evolution, Methods and Concepts*,
DOI 10.1007/978-3-319-61569-1_14

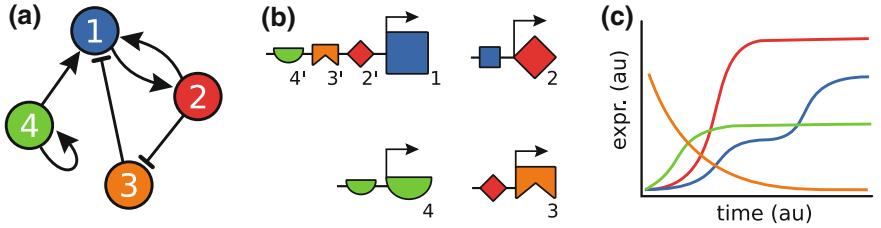


Fig. 1 Dynamic regulatory networks. **a** A simple regulatory network with four components (*1–4*) and their interactions. *Arrows* indicate activating interactions, *T-bars* inhibiting ones. **b** The system of panel (**a**) interpreted as a set of genes with upstream regions. Numbers with a prime (′) refer to regulatory input sites (binding sites) for each of the genes. **c** Hypothetical temporal expression dynamics of the network. (*au*) stands for arbitrary units

The modelling of regulatory networks integrates ideas from biology, physics, computational theory, neuroscience, and early machine learning (Fierst and Phillips 2015). Ever since Glass and Kauffman formulated their Boolean networks (Glass and Kauffman 1973; Kauffman 1969), many variants have been proposed and used over the years, but all are based on a small set of basic assumptions (Gjuvsland et al. 2007). If we restrict ourselves for the moment to gene regulatory networks (GRNs), we can identify the following principles (Fig. 1): (i) gene expression levels are controlled by transcription factors, that combine into an input function; (ii) the effect of the input function on the generation of gene product is modulated by a response function, often a sigmoidal or step function, that generates a threshold behaviour; (iii) transcription factors are themselves the products of gene expression, thus creating a network of genes with feedback loops; and (iv) the input and/or response function capture in a phenomenological manner regulation of transport, splicing, (post)translational modifications, and metabolic processes.

Over the years, experimental evidence has accumulated supporting the general validity of the above-listed principles. For instance, the *cis*-regulatory regions of genes have been characterized as Boolean functions (Abou-Jaoud et al. 2016). Despite such progress, experimentally measuring gene networks and their expression dynamics remains a technically challenging task (Jaeger and Crombach 2012). It is obvious that a divide exists between our theoretical and empirical understanding of the functioning of regulatory networks. And with respect to the evolution of such networks, the discrepancy between theory and experiment is even larger.

The objective of the following sections is to provide an overview of our understanding of network evolution, one that is heavily biased towards a computational and mathematical point of view. Without claiming an exhaustive review of the literature, I critically present work that sparked my interest over the years.[1]

[1]For a thorough review of the literature, see the excellent work on the evolution of innovation by Wagner (2011).

I start by introducing two central concepts in evolutionary thinking, namely that of the genotype–phenotype map and the genotype network. I then move to the relationship between network structure and function. Regarding experimental evidence, I focus on two eukaryotic model systems that have played a central role in my work. I conclude by identifying several key challenges in the field of network evolution.

2 Insights from Abstract Regulatory Networks

2.1 Evolution of the Genotype to Phenotype Mapping

A central question in evolutionary theory is how a genotype maps to a phenotype (the GP map), and how this mapping evolves (Alberch 1991; Catalan et al. 2017; Pigliucci 2010). As with many terms in biology, genotype and phenotype are only loosely defined here. For our purposes, it suffices to take the genotype to stand for the genome, in other words, the heritable material that is transmitted from one generation to the next. Next, the phenotype develops from the genotype—in combination with other factors such as the environment—resulting in an individual organism with a form and function. The mapping evolves as mutations modify the genotype, leading to phenotypic variation on which natural selection may act.

In classic evolutionary theory, however, the process of development from genotype to phenotype is taken to be a linear, one-to-one mapping: one set of genes leads to one phenotype. Under this assumption, the GP map can be eliminated and evolution was indeed simplified to a process of changes in gene (allele) frequencies. Nowadays, there is a lot of evidence that development is not such a simple mapping. Instead, genes cooperate in complex (and complicated) manners to create patterns of expression over space and time. These patterns affect the behaviour of cells and tissues, which in turn feeds back on gene activity. Thus, it is argued that gene regulatory networks are a crucial link between genotype and phenotype and that studying network structure and dynamics will lead to a deeper understanding of the GP map and how it evolves.

In fact, our high-level understanding of the GP mapping is that it is a many-to-many, high-dimensional, nonlinear function (Fig. 2a): (1) Many genotypes result in the same phenotype. Moreover, the same genotype may result in multiple phenotypes, for instance, due to stochastic and environmental factors. (2) Genotypes are high-dimensional entities. One only needs to consider a fruit fly's genome size of 175 Mb (Ellis et al. 2014). (3) Nonlinearity comes about through feedback processes in gene networks, as mentioned above.

The GP mapping is closely related to an important and relatively novel concept in evolutionary theory, namely the genotype network (GN). Where the GP map is focused on the individual, the genotype network generalizes it to the entire space of genotypes: it encompasses all genotypes that generate the same phenotype and links

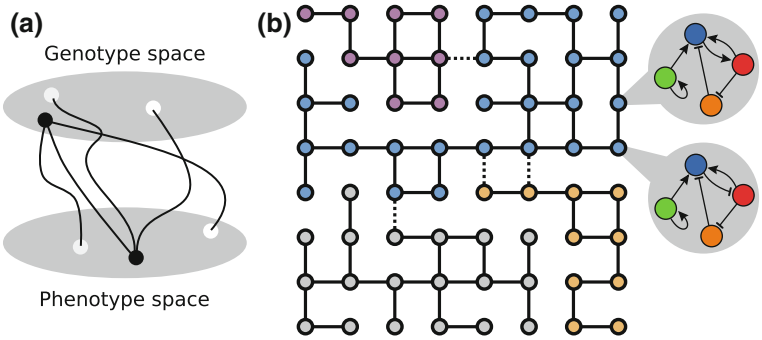


Fig. 2 Genotype–phenotype map and genotype networks. **a** The mapping from genotype to phenotype space is many-to-many, high-dimensional (not depicted), and nonlinear (*curved lines*). **b** Four genotype networks. Each node of a genotype network is a regulatory network and nodes are linked if they differ by a single mutation (compare circular insets for the activation/inhibition of the *blue* gene on the *red* gene)

them through single mutations (Fig. 2b). GNs were originally discovered in models for "simple" genotype–phenotype maps of RNA and protein folding—and later their existence was confirmed experimentally. In case of RNA, it was found that many RNA sequences, the genotypes, fold into the same secondary structure, the phenotype (Schuster et al. 1994). By connecting sequences through single nucleotide substitutions, Schuster et al. realized they could travel across genotype space without changing the phenotype. They called these interconnected mutational paths neutral networks, which were later renamed as genotype networks by Wagner to remove the association with fitness (Wagner 2011).

In our case, the genotype is a gene regulatory network, and the phenotype is defined as the stable expression pattern that the network establishes after updating its internal dynamics for a given amount of time. Since most network models have been studied using a binary set of expression states (genes are "on" or "off", respectively 1 and 0), such a stable expression pattern is a vector of 1s and 0s. The single mutations that link regulatory networks in the GN make qualitative and quantitative changes to network structure. The first type of changes refers to gene duplications, deletions, or the addition and removal of gene–gene interactions (Fig. 2b). The second indicates changes in the logic of the input function of a given gene, the weight of a particular interaction, etc. In effect, we are studying networks of networks, and we want to understand how populations of individual regulatory networks explore this evolutionary space.

In the next section, I will highlight the most important insights that we have gained from studying models of "networks of networks".

2.2 Robustness and Evolvability

Robustness and evolvability are two concepts pervading many areas of biology. They are both defined in various contexts, each slightly different, nevertheless they have a clear intuitive meaning. Robustness indicates the ability to maintain a phenotype as the genotype changes. Evolvability hints at the ability to use genotypic changes to discover novel phenotypes. On first sight, these two concepts seem antagonistic—if you are inert to changes, you cannot change quickly—and in many cases they are. However, we now understand that the two can also reinforce each other. The key is to distinguish between robustness and evolvability on the level of the genotype and the phenotype separately. And this is where the GP map and genotype networks offer insight.

A genotype network is not a homogeneous structure. There are areas with more and with less connections between neighbouring genotypes, since not all neighbours will produce the same phenotype (and thus by definition do not belong to the same GN). A higher density of neighbours means that if you produce mutated offspring, they are more likely to be a neighbour and thus still generate the same phenotype. This means, while the genotype is changing and, therefore, is not robust, the associated phenotype has phenotypic robustness. Since natural selection acts on the phenotype, it means our network has robustness and under a constant (stabilizing) selection regime the population of GRNs will evolve to the most densely connected part of the GN (van Nimwegen et al. 1999). This phenomenon has also been named "survival of the flattest" (Huynen and Hogeweg 1994; Wilke et al. 2001) and is strongly connected to the older ideas of homeostasis, canalization, and cryptic variation (Bergman and Siegal 2003; Gjuvsland et al. 2007; Siegal and Bergman 2002).

As mentioned above, for a long time robustness and evolvability were thought to be each other's complement. If robustness increases, evolvability has to go down. However, by taking into account the GP map and GNs, robustness was found to increase evolvability at the phenotypic level. The explanation is simple and revolves around the idea of "the adjacent possible" (Loreto et al. 2016). Phenotypic robustness allows a population of genotypes to visit different parts of genotype space by travelling (neutrally) over their GN. These areas will have different neighbouring phenotypes, and thus selection has more different phenotypes to choose from. This augmentation in population variation translates into an increase in evolvability (Wagner 2008, 2011). Actually, we have a more refined understanding of genotype space. When exploring a GN, some other GNs tend to be always present in the local neighbourhood of a genotype, nevertheless as one travels across the vast genotype space one always meets novel phenotypes (Fontana and Schuster 1998; Wagner 2011).

This reconciliation of robustness and evolvability is one of the central results of studying genotype networks at the level of GRNs. It has since been reinforced through a range of models varying the research objective, model design, and implementation. The main weakness, though, has been the lack of experimental

evidence for the existence of genotype networks. This is slowly changing, and in Sect. 3 I report on two experimental studies that are providing the first steps in this direction. Moreover, a sequence-based approach that maps the mutational transitions between transcription factor binding sites in mouse is providing an independent line of support (Payne and Wagner 2014).

Previously, a parallel case of neutrality against adaptation was shown to hold for RNA and protein folding (Fontana and Schuster 1998a, b). However, there is an important difference between folding and regulatory networks. While for RNA and protein folding physical–chemical laws define how genotype maps to phenotype, for regulatory networks the GP mapping is an evolved entity itself (Hogeweg 2012). This is called the evolution of evolution, which still is a poorly explored field of study.

Unfortunately, evolution of evolution is also known under the term "evolvability". In fact, evolvability has been (re)defined many times, so this term merits a closer look (Pigliucci 2008). Roughly speaking, evolvability is defined in three ways: firstly, it means that evolution is actually able to find better adapted mutants. If such evolvability is not present, evolution can not take place. Secondly, evolvability may mean improving the ability to generate adaptive mutants. This type of evolvability was accepted as a side effect of other evolutionary processes, but it was long discussed if it could be an evolvable trait itself. Thirdly, evolvability may be understood as innovation, producing truly novel phenotypes and functions, such as the wings of birds and bats.

Evolution of evolution matches best with the second definition. In fact, a strong contribution to the acceptance of this type of evolvability was made through evolutionary simulations of gene regulatory networks (Crombach and Hogeweg 2008). Evolving in dynamic environments led to evolvability "from scratch" through networks with a hub-and-spoke structure. This type of evolvability may be understood as a learning process (Kouvaris et al. 2017; Parter et al. 2008). If certain environmental changes are observed often enough, evolution can encode them in the network structure as closely linked "memories". The re-appearance of an environment then requires only one or few mutations to "recall the memory", quickly leading to well-adapted individuals. In terms of the GP map and GNs, the population has evolved to a specific neighbourhood where many mutational transitions exist between phenotypes that are fit in each of the environments.

2.3 Modularity, Hierarchy, and Sparsity

Robustness and evolvability do not make any specific predictions on the regulatory structure of a network. Two other popular concepts, modularity and hierarchy, do point at specific network topologies. Like robustness and evolvability, modularity has several definitions. I here consider structural modularity, which is defined as a network's structure being decomposable into several nonoverlapping sets of components, named modules, where components within a module interact more closely with

each other, than they do across module boundaries. Ideally, each module has its own function and the full phenotype arises by wiring these functions together. Hierarchy is a closely related concept, which posits that components are themselves decomposable into smaller components. Both concepts are understood to bring benefits to robustness and evolvability. A modular and/or hierarchical network structure localizes the impact of a mutation, and rewiring modules speeds up the adaptation process.

In the context of regulatory networks, the evolution of modularity is observed under a variety of conditions. Kashtan et al. showed that if networks are evolved to solve modularly composed problems, the structure of network can reflect the modularity of the environment (Kashtan and Alon 2005). The introduction of a cost per regulatory interaction may also lead to the evolution of modular and hierarchical networks (Clune et al. 2013; Mengistu et al. 2016). And even the definition of mutational operators impacts on the evolution of modularity (Friedlander et al. 2013). Multiplicative mutations, where regulatory interactions are multiplied with random numbers, tend to reduce the number of connections in a network, while additive mutations, where interactions are added with random numbers, do not.

Closely related to modularity and hierarchy is the observation that many networks only realize a small set of all possible interactions. Such sparsity was found to be more likely if networks are hierarchical (Corominas-Murtra et al. 2013). That, however, does not explain how sparsity evolves. In evolutionary simulations with connection costs (Mengistu et al. 2016), with multiplicative mutation schemes (Friedlander et al. 2013), and with deletions of interactions strongly favoured over insertions (unpublished results) sparsity indeed arises. In all cases, the straightforward explanation is a penalty on establishing a regulatory link, either explicitly (cost) or implicitly (asymmetry of mutations). Whether such costs constitute a sufficiently strong selection pressure in empirical (gene) regulatory networks remains an open question. A case can be made for the energy costs incurred to make regions of the genome accessible through chromatin remodelling and to recruit transcription factors and other RNAs/proteins (e.g. transcriptional and splicing machinery).

Despite the fact that studies on modularity, hierarchy, and sparsity often invoke the argument that such network structures increase robustness and evolvability, to the best of my knowledge they have not been explicitly linked to genotype networks. In addition, a multitude of studies explicitly report networks that lack structural modularity and/or hierarchy (Crombach and Hogeweg 2008; Jimenez et al. 2017; ten Tusscher and Hogeweg 2011), while they do display robustness and evolvability. Clearly, we do not fully understand under which conditions regulatory networks evolve modular, hierarchical, or sparse solutions. The integration of these concepts with GNs is an exciting theme to pursue.

In summary, genotype networks provide a framework to understand a range of evolutionary concepts. While I propose the above-discussed results are valid more generally, one major caveat is that they are based mostly on Boolean network models. The discrete nature of this modelling formalism simplifies definitions, computational aspects, and the interpretation of results. Yet it is largely an open question what artefacts and limitations the discreteness has introduced and how

continuous models will refine or adjust the insight we now have regarding evolution on genotype networks.

2.4 *Network Structure and Function Are Only Loosely Coupled*

At the level of small circuits and network motifs (~3 or 4 genes), structural modularity and hierarchy are not informative concepts any more. The networks are simply too small to decompose into modules. Instead, at this level one of the major questions is how network structure relates to function.

Initial reports of over-representation of specific small network motifs in large transcriptional networks of *E. coli* and *S. cerevisiae* suggested these motifs were special for biological systems (Milo et al. 2004; Milo et al. 2002). Amongst the motifs, the most famous one was the three-gene feed-forward loop (FFL, see Fig. 3). The abundance of these motifs was interpreted as a signature of the constraints under which the networks had evolved, and later studies attributed the motif a signalling function (Mangan and Alon 2003; Mangan et al. 2003). However, follow-up research showed that neutral evolutionary dynamics could already explain the over-representation (Cordero and Hogeweg 2006; Siegal et al. 2007). Moreover, the function of the motif critically depends on the strength of regulatory interactions and the context in which the motif is embedded (Wall 2011). The conclusion is that network structure, the "wiring", is only a partial description of a network. In addition to structure, one needs the precise mathematical formulation of each input function, the interaction weights, and input signal to uniquely determine network function (Ingram et al. 2006; Wall et al. 2005).

Complementary, if we define a function and exhaustively search for networks that robustly perform it, we find only few possible network structures. This was shown to be the case for the "perfect adaptation" function realized by three-enzyme network circuits (Ma et al. 2009) and for stripe-forming networks (Cotterell and Sharpe 2010). Also, evolving small regulatory networks with specific cellular functions, such as bistable switches and oscillators, often showed bias towards typical network structures (Franois and Hakim 2004).

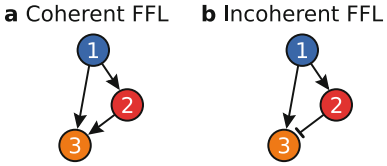


Fig. 3 Feed-forward loop motifs (FFL). **a** The coherent FFL. The motif is known to function as a sign-sensitive delay and a persistence detector. **b** The incoherent FFL may function as a pulse generator and response accelerator. For both types of FFL, three additional variants exist with differences in the sign of interactions

Two sides of the same coin become apparent. On one side, the claim is that for relevant biological functions, only few circuits robustly perform the task (Ma et al. 2009). Logically, the suggestion was made that a central database meticulously describing small functional networks would be extremely useful (Lim et al. 2013). To date I am not aware of the existence of such a repository, though. On the other side, screening large numbers of small networks for many parameter values shows that each network has multiple possible dynamics and hence functions (Cotterell and Sharpe 2010; Jimenez et al. 2015; Jimenez et al. 2017; Wall 2011). At first sight, those two results seem to exclude each other.

The emerging consensus is that on the scale of small three-gene networks there is a many-to-many mapping from structure to function (Payne and Wagner 2014; Payne and Wagner 2015), similarly to the many-to-many property of the more general genotype–phenotype mapping. Each network has more than one function, and each function is performed by more than one network. And this argument of redundancy has been pushed further (Sorrells and Johnson 2015): perhaps the precise topology of a network simply does not matter, as long as its function is correct for the rest of the biological system in which it is embedded. It reminds of RNA and protein folding, where at many positions along the chain the precise nucleotides and amino acids need not be conserved, as long as the overall folding and (enzymatic) function are maintained.

With the exception of Francois and Hakim (2004), the studies mentioned here on small network structure and function rely on large-scale screening of parameter sets and exhaustive enumeration of network structures. Thus, it remains a largely open question how an evolutionary process navigates the genotype space of these small networks. Considering the redundancy between structure and function, I speculate that subtle evolutionary processes can easily leave a signature by favouring one structure (or function) over another.

3 Insight from Data-Driven Regulatory Networks

After having discussed the key theoretical insights of the field, we move to some pioneering experimental efforts that explore the match between what we observe in computational models and evolution as it happened on Earth. I focus on “simple” developmental processes, where the dynamics of gene regulation and signalling are not influenced by tissue growth, rearrangements, and other morphogenetic processes.[2]

A powerful approach to understand the evolution of extant species is a comparative analysis of the same developmental process in different species. This requires a

[2]For examples of the evolution of more complex developmental processes, see studies on mammalian tooth development (Salazar-Ciudad 2012; Salazar-Ciudad and Jernvall 2010; Salazar-Ciudad and Marín-Riera 2013) and compare limb against fin development (Onimaru et al. 2016; Raspopovic et al. 2014).

model organism for which the particular developmental process is well-studied, a complete "parts list" of the process under study, and one or more related species for which any experimental techniques used to gather data work robustly. Below, I discuss two experimental systems are these criteria were met. The first is a comparative study of vulva development in nematode worms, the second an analysis of a body plan patterning network in multiple insect species.

3.1 *Developmental System Drift in the Worm …*

Vulva patterning of the nematode *Caenorhabditis elegans* is one of the best-understood developmental processes from both an experimental and modelling point of view (Sommer 2012). The vulva is the egg-laying and mating organ of *Caenorhabditis*, specified from a row of six precursor cells (Fig. 4a). These adopt one of three alternative fates in a stereotypical pattern (3°3°2°1°2°3°), where 1° and 2° indicate inner and outer vulval fate, respectively, and 3° a nonvulval fate. To correctly commit to one of these cell types, two regulatory pathways are involved. One is based on morphogen induction, the other on lateral cell–cell communication, and they interact through activating and inhibitory crosstalk (Hoyos et al. 2011). In a comparison across *Caenorhabditis* and four closely related species, experimental studies and mathematical modelling were used to characterize how each species uses these two pathways. It was established that network structure remained the same between these species. Yet fitting to experimental results showed that each species used different parameter sets, corresponding to distinct patterning dynamics. Species-specific vulval patterning proceeded through the use of morphogen induction and lateral induction at different ratios.

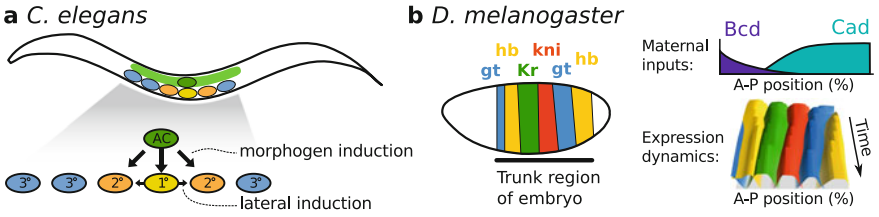


Fig. 4 Two biological systems for which data-driven models exist. **a** Vulva patterning of the nematode *C. elegans*. The *green* area in the worm marks the gonad; AC is the gonadal anchor cell. The six precursor cells are coloured by cell fate. The two regulatory pathways, morphogen and lateral induction, are indicated by *black arrows*. **b** Early body plan patterning in the embryo of the fruit fly *D. melanogaster*. On the *left*, the embryo is shown with anterior (head) to the left and dorsal up. On the *right*, maternal inputs and gap gene dynamics are shown for the trunk region, along the antero–posterior (A–P) axis. Four gap genes are *hunchback* (*hb*) in *yellow*, *Krüppel* (*Kr*) in *green*, *giant* (*gt*) in *blue*, and *knirps* (*kni*) in *red*. Maternal morphogen gradients are Bicoid (Bcd) in *purple* and Caudal (Cad) in *cyan*

The study uncovered a type of cryptic or neutral evolution known as developmental system drift (DSD) (True and Haag 2001; Weiss 2005; Weiss and Fullerton 2000). It is an important concept in current-day evolutionary developmental biology. Originally, DSD was used to identify neutral network-level evolution, where the output of a system remains the same, yet its inner workings display qualitative change, such as changes in the sign of interactions (activation to inhibition and vice versa) or alterations of entire signalling pathways. Here system drift of a quantitative kind was demonstrated as well.

3.2 Developmental System Drift in the Fly

The gap gene system is crucial for early development of the fly. It lays down a pattern of broad expression domains that are later refined into a segmented body plan (reviewed in Jaeger 2011). In effect, the gap genes generate a set of stripes along the main body axis, that is from head to tail (Fig. 4b). It is a well-known regulatory network and an example system for complex patterning. The evolution of the gap gene network was studied by comparing the fruit fly *Drosophila melanogaster* and two other (nonmodel) fly species, the scuttle fly *Megaselia abdita* and moth midge *Clogmia albipunctata*.

Clogmia and *Drosophila* split $\sim$250 MY ago, close to the origin of the dipterans (dipterans are flies, midges, and mosquitoes). A combined experimental and modelling approach to study the gap gene network of *Clogmia* substantially improved our understanding of the evolution of a conserved, essential developmental process. Beforehand little was known about *Clogmia*. Its posterior maternal gradient is Caudal (Cad), but its anterior gradient remains unknown (Jimenez-Guri et al. 2013). With respect to the gap genes, expression data showed that *Clogmia* has fewer expression domains: the posterior domains of *giant* (*gt*) and *hunchback* (*hb*) are missing during the blastoderm stage (Fig. 4b) (Garca-Solache et al. 2010). After fitting models to data, 100 of the best network models were selected for further analysis. All networks shared a core set of interactions. These included the mutual inhibition between *hb* and *kni*, which is also found in *Drosophila*. Using unsupervised learning techniques, we categorized the other genetic interactions into four groups, each resulting in distinct hypotheses regarding network structure. Unfortunately, experimental difficulties with this species have not allowed us so far to test the hypotheses.

The fly *Megaselia* did not have this shortcoming. Its gap gene network was successfully reverse-engineered, and the 20 best gene circuit solutions—fully validated with RNAi knockdown experiments (Wotton et al. 2015)—were selected for a comparison to 20 *Drosophila* gene circuits (Crombach et al. 2016). The external inputs were known to differ: in *Megaselia,* the anterior determinant Bicoid has a broader gradient, and Caudal is not maternal. Instead Cad comes up with the gap genes, which suggests its influence is delayed in comparison to *Drosophila*. Yet, at gastrulation the pattern of stripes is essentially equivalent in both species.

This suggests the *Megaselia* network compensates for upstream changes in Bicoid and Caudal. Indeed, we found species-specific expression dynamics, which were faithfully reproduced by the network models. A detailed analysis of the models helped us understand that while the qualitative network structure is conserved between the flies, genetic interactions have specific quantitative differences. And these differences lead to compensatory mechanisms.

Just like for *C. elegans* and relatives, we carefully documented a case of system drift. The term quantitative system drift (QSD) was coined to describe the subtle, yet significant differences between *Megaselia* and *Drosophila* and we it to the concept of genotype networks (Crombach et al. 2016; Wotton et al. 2015). Indeed, we suggested system drift should be understood as travelling on a genotype network. The consequences of developmental system drift are that network function may not be as conserved as expected (Pavlicev and Wagner 2012; True and Haag 2001; Weiss 2005; Weiss and Fullerton 2000). And that observing a given phenotype, intermediate or "final", does not mean that species employ the same underlying molecular mechanisms.

4 Discussion and Future Directions

Regulatory networks are crucial for understanding how genomic information is interpreted and combined with environmental cues to make cellular decisions, ultimately leading to the implementation of life cycle strategies and/or the successful development of a multicellular body. To understand how such networks are shaped over evolutionary time, it is relatively straightforward to take abstract biological networks and study them through computer simulations. Yet, a much more daunting task is to extract networks and their dynamics from experimental data and to use them to understand how theoretical concepts actually apply to life on Earth. Surely, we have only been scratching the surface in this respect. Currently, our knowledge on eukaryotes is limited to a handful of regulatory networks, for which we have a formal understanding (i.e. a mathematical model) of both their structure and dynamics across species.[3]

I propose two lines of research to deepen our understanding of regulatory networks and their evolution. Obviously, the first one is to perform more comparative studies involving well-known experimental systems and their regulatory networks. There are other eukaryotic systems that lend themselves well for such endeavours. First of all, building upon the knowledge of the gap gene system, the Drosophila pair-rule network is probably one of the best candidate systems. It functions immediately downstream of the gap genes is increasingly well-understood (Clark 2017; Clark and Akam 2016), and comparative knowledge is accumulating (Jiang et al. 2015; Palsson et al. 2014). Another candidate system is the regulatory network

[3]I do not consider bacteria and archaea due to my scientific interests in eukaryotes.

for neuronal fate specification (reviewed in Toma and Hanashima 2015). The comparative analysis of recently evolved cortical structures against older ones, e.g. neocortex versus olfactory cortex, will provide insight into brain evolution (Borrell and Reillo 2012; Klingler 2017). Other cell-centred differentiation systems, like haematopoiesis (Bertolino et al. 2016), will allow similar research strategies.

Second, network models are extremely successful at implementing and explaining a system's understanding at a rather abstract level, where entire signalling cascades are collapsed into a single interaction (see examples in Hoyos et al. 2011; Raspopovic et al. 2014; Salazar-Ciudad 2012; Salazar-Ciudad and Jernvall 2010). Abstraction is part of any modelling process, yet to predict the consequences of manipulating the genotype, we may require more detailed models than we are using at the moment. In my opinion, this need will arise more quickly than perhaps anticipated, especially since genome editing techniques are developing and refining rapidly.[4] I propose to use overlapping modelling perspectives along the axis that takes us from genotype to phenotype. For instance, regulatory processes on a 3D folded genome lead to gene expression at the right moment and in the right amount. Explicitly modelling this process and deriving which variables are crucial will allow us to create well-founded, simplified phenomenological models. Such models may not be too different from our current mathematical equations, though a well-informed choice on whether to use a Hill function or another type of response function is an important advance.

Several other phenomena have been identified that warrant a critical re-evaluation of current network modelling formalisms (Niklas et al. 2015). Amongst them are alternative splicing, histone modifications (epigenetics), and intrinsically disordered protein domains. All create a more complex regulatory system by facilitating spatio-temporal context dependence, without significantly increasing the genotype. Indeed, taking into account an epigenetic time scale has been found to increase evolvability in novel environments (Furusawa and Kaneko 2013) and provided an explanation for the observation of temporal delays in stem cell reprogramming (Miyamoto et al. 2015). At the moment, any consequences for evolution on GNs are unknown, however.

With respect to the question how network structure relates to function, in Sect. 2.4 I suggested the link with GNs should be improved. Interestingly, a connection has been forged with evolutionary genetics. Collections of small networks have been shown to generate some of the signature phenomena of quantitative genetics, such as additivity, dominance, and epistasis (Cotterell and Sharpe 2013; Gjuvsland et al. 2007a, b, 2013; Omholt et al. 2000). These studies provide a basis for a deeper understanding of the connection between the mechanistic approach of regulatory network evolution and the statistical approach of the Modern Synthesis.

In conclusion, exciting times are ahead.

[4]One may expect that advanced genome editing techniques can help us rapidly improve our understanding of the genotype–phenotype map as well.

Acknowledgements I thank Elise Parey for critical reading and comments. And I kindly acknowledge Foundation Bettencourt Schueller.

References

Abou-Jaoud W, Traynard P, Monteiro PT, Saez-Rodriguez J, Helikar T, Thieffry D, Chaouiya C (2016) Logical modeling and dynamical analysis of cellular networks. Front Genet 7:94

Alberch P (1991) From genes to phenotype: dynamical systems and evolvability. Genetica 84(1):5–11

Bergman A, Siegal ML (2003) Evolutionary capacitance as a general feature of complex gene networks. Nature 424(6948):549–552

Bertolino E, Reinitz J, Manu (2016) The analysis of novel distal *Cebpa* enhancers and silencers using a transcriptional model reveals the complex regulatory logic of hematopoietic lineage specification. Dev Biol 413(1):128–144

Borrell V, Reillo I (2012) Emerging roles of neural stem cells in cerebral cortex development and evolution. Dev Neurobiol 72(7):955–971

Britten RJ, Davidson EH (1969) Gene regulation for higher cells: a theory. Science 165 (3891):349–357

Catalan P, Arias CF, Cuesta JA, Manrubia S (2017) Adaptive multiscapes: an up-to-date metaphor to visualize molecular adaptation. Biol Direct 12(1):7

Clark E (2017) Dynamic patterning by the *Drosophila* pair-rule network reconciles long-germ and short-germ segmentation. bioRxiv, p 099671

Clark E, Akam M (2016) Odd-paired controls frequency doubling in *Drosophila* segmentation by altering the pair-rule gene regulatory network. eLife 5

Clune J, Mouret JB, Lipson H (2013) The evolutionary origins of modularity. Proc Biol Sci 280 (1755):20122863

Cordero OX, Hogeweg P (2006) Feed-forward loop circuits as a side effect of genome evolution. Mol Biol Evol 23(10):1931–1936

Corominas-Murtra B, Goi J, Sol RV, Rodrguez-Caso C (2013) On the origins of hierarchy in complex networks. Proc Natl Acad Sci USA 110(33):13316–13321

Cotterell J, Sharpe J (2010) An atlas of gene regulatory networks reveals multiple three-gene mechanisms for interpreting morphogen gradients. Mol Syst Biol 6:425

Cotterell J, Sharpe J (2013) Mechanistic explanations for restricted evolutionary paths that emerge from gene regulatory networks. PLoS ONE 8(4):e61178

Crombach A, Hogeweg P (2008) Evolution of evolvability in gene regulatory networks. PLoS Comput Biol 4(7):e1000112

Crombach A, Wotton KR, Jimenez-Guri E, Jaeger J (2016) Gap gene regulatory dynamics evolve along a genotype network. Mol Biol Evol 33(5):1293–1307

Ellis LL, Huang W, Quinn AM, Ahuja A, Alfrejd B, Gomez FE, Hjelmen CE, Moore KL, Mackay TFC, Johnston JS, Tarone AM (2014) Intrapopulation genome size variation in *D. melanogaster* reflects life history variation and plasticity. PLoS Genet 10(7):e1004522

Fierst JL, Phillips PC (2015) Modeling the evolution of complex genetic systems: the gene network family tree. J Exp Zool B Mol Dev Evol 324(1):1–12

Fontana W, Schuster P (1998a) Continuity in evolution: on the nature of transitions. Science 280(5368):1451–1455

Fontana W, Schuster P (1998b) Shaping space: the possible and the attainable in RNA genotype–phenotype mapping. J Theor Biol 194(4):491–515

Franois P, Hakim V (2004) Design of genetic networks with specified functions by evolution in silico. Proc Natl Acad Sci USA 101(2):580–585

Friedlander T, Mayo AE, Tlusty T, Alon U (2013) Mutation rules and the evolution of sparseness and modularity in biological systems. PLoS ONE 8(8):e70444

Furusawa C, Kaneko K (2013) Epigenetic feedback regulation accelerates adaptation and evolution. PLoS ONE 8(5):e61251

Garca-Solache M, Jaeger J, Akam M (2010) A systematic analysis of the gap gene system in the moth midge *Clogmia albipunctata*. Dev Biol 344(1):306–318

Gjuvsland AB, Hayes BJ, Omholt SW, Carlborg O (2007a) Statistical epistasis is a generic feature of gene regulatory networks. Genetics 175(1):411–420

Gjuvsland AB, Plahte E, Omholt SW (2007b) Threshold-dominated regulation hides genetic variation in gene expression networks. BMC Syst Biol 1:57

Gjuvsland AB, Wang Y, Plahte E, Omholt SW (2013) Monotonicity is a key feature of genotype-phenotype maps. Front Genet 4:216

Glass L, Kauffman SA (1973) The logical analysis of continuous, non-linear biochemical control networks. J Theor Biol 39(1):103–129

Hogeweg P (2012) Toward a theory of multilevel evolution: long-term information integration shapes the mutational landscape and enhances evolvability. Adv Exp Med Biol 751:195–224

Hoyos E, Kim K, Milloz J, Barkoulas M, Pnigault JB, Munro E, Felix MA (2011) Quantitative variation in autocrine signaling and pathway crosstalk in the *Caenorhabditis* vulval network. Curr Biol 21(7):527–538

Huynen MA, Hogeweg P (1994) Pattern generation in molecular evolution: exploitation of the variation in RNA landscapes. J Mol Evol 39(1):71–79

Ingram PJ, Stumpf MPH, Stark J (2006) Network motifs: structure does not determine function. BMC Genom 7:108

Jaeger J (2011) The gap gene network. Cell Mol Life Sci 68(2):243–274

Jaeger J, Crombach A (2012) Life's attractors: understanding developmental systems through reverse engineering and *in silico* evolution. Adv Exp Med Biol 751:93–119

Jiang P, Ludwig MZ, Kreitman M, Reinitz J (2015) Natural variation of the expression pattern of the segmentation gene *even-skipped* in *Drosophila melanogaster*. Dev Biol 405(1):173–181

Jimenez A, Cotterell J, Munteanu A, Sharpe J (2015) Dynamics of gene circuits shapes evolvability. Proc Natl Acad Sci USA 112(7):2103–2108

Jimenez A, Cotterell J, Munteanu A, Sharpe J (2017) A spectrum of modularity in multifunctional gene circuits. Mol Syst Biol 13(4):925

Jimenez-Guri E, Huerta-Cepas J, Cozzuto L, Wotton KR, Kang H, Himmelbauer H, Roma G, Gabaldn T, Jaeger J (2013) Comparative transcriptomics of early dipteran development. BMC Genom 14:123

Kashtan N, Alon U (2005) Spontaneous evolution of modularity and network motifs. Proc Natl Acad Sci USA 102(39):13773–13778

Kauffman SA (1969) Metabolic stability and epigenesis in randomly constructed genetic nets. J Theor Biol 22(3):437–467

Klingler E (2017) Development and organization of the evolutionarily conserved three-layered olfactory cortex. eNeuro 4(1)

Kouvaris K, Clune J, Kounios L, Brede M, Watson RA (2017) How evolution learns to generalise: Using the principles of learning theory to understand the evolution of developmental organisation. PLoS Comput Biol 13(4):e1005358

Lim WA, Lee CM, Tang C (2013) Design principles of regulatory networks: searching for the molecular algorithms of the cell. Mol Cell 49(2):202–212

Loreto V, Servedio VDP, Strogatz SH, Tria F (2016) Dynamics on expanding spaces: modeling the emergence of novelties. In: Esposti MD, Altmann EG, Pachet F (eds) Creativity and universality in language, lecture notes in morphogenesis. Springer International Publishing, pp 59–83

Ma W, Trusina A, El-Samad H, Lim WA, Tang C (2009) Defining network topologies that can achieve biochemical adaptation. Cell 138(4):760–773

Mangan S, Alon U (2003) Structure and function of the feed-forward loop network motif. Proc Natl Acad Sci USA 100(21):11980–11985

Mangan S, Zaslaver A, Alon U (2003) The coherent feedforward loop serves as a sign-sensitive delay element in transcription networks. J Mol Biol 334(2):197–204

Mengistu H, Huizinga J, Mouret JB, Clune J (2016) The evolutionary origins of hierarchy. PLoS Comput Biol 12(6):e1004829
Milo R, Itzkovitz S, Kashtan N, Levitt R, Shen-Orr S, Ayzenshtat I, Sheffer M, Alon U (2004) Superfamilies of evolved and designed networks. Science 303(5663):1538–1542
Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D, Alon U (2002) Network motifs: simple building blocks of complex networks. Science 298(5594):824–827
Miyamoto T, Furusawa C, Kaneko K (2015) Pluripotency, differentiation, and reprogramming: a gene expression dynamics model with epigenetic feedback regulation. PLoS Comput Biol 11(8):e1004476
Niklas KJ, Bondos SE, Dunker AK, Newman SA (2015) Rethinking gene regulatory networks in light of alternative splicing, intrinsically disordered protein domains, and posttranslational modifications. Front Cell Dev Biol 3:8
van Nimwegen E, Crutchfield JP, Huynen M (1999) Neutral evolution of mutational robustness. Proc Natl Acad Sci USA 96(17):9716–9720
O'Malley MA (2012) Evolutionary systems biology: historical and philosophical perspectives on an emerging synthesis. Adv Exp Med Biol 751:1–28
Omholt SW, Plahte E, Oyehaug L, Xiang K (2000) Gene regulatory networks generating the phenomena of additivity, dominance and epistasis. Genetics 155(2):969–980
Onimaru K, Marcon L, Musy M, Tanaka M, Sharpe J (2016) The fin-to-limb transition as the re-organization of a Turing pattern. Nat Commun 7:11582
Palsson A, Wesolowska N, Reynisdttir S, Ludwig MZ, Kreitman M (2014) Naturally occurring deletions of Hunchback binding sites in the *even-skipped* stripe 3 + 7 enhancer. PLoS ONE 9(5):e91924
Parter M, Kashtan N, Alon U (2008) Facilitated variation: how evolution learns from past environments to generalize to new environments. PLoS Comput Biol 4(11):e1000206
Pavlicev M, Wagner GP (2012) A model of developmental evolution: selection, pleiotropy and compensation. Trends Ecol Evol (Amst) 27(6):316–322
Payne JL, Wagner A (2014) The robustness and evolvability of transcription factor binding sites. Science 343(6173):875–877
Payne JL, Wagner A (2015) Function does not follow form in gene regulatory circuits. Sci Rep 5:13015
Pigliucci M (2008) Is evolvability evolvable? Nat Rev Genet 9(1):75–82
Pigliucci M (2010) Genotype-phenotype mapping and the end of the 'genes as blueprint' metaphor. Philos Trans R Soc Lond B Biol Sci 365(1540):557–566
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
Salazar-Ciudad I (2012) Tooth patterning and evolution. Curr Opin Genet Dev 22(6):585–592
Salazar-Ciudad I, Jernvall J (2010) A computational model of teeth and the developmental origins of morphological variation. Nature 464(7288):583–586
Salazar-Ciudad I, Marín-Riera M (2013) Adaptive dynamics under development-based genotype-phenotype maps. Nature 497(7449):361–364
Schuster P, Fontana W, Stadler PF, Hofacker IL (1994) From sequences to shapes and back: a case study in RNA secondary structures. Proc Biol Sci 255(1344):279–284
Siegal ML, Bergman A (2002) Waddington's canalization revisited: developmental stability and evolution. Proc Natl Acad Sci USA 99(16):10528–10532
Siegal ML, Promislow DEL, Bergman A (2007) Functional and evolutionary inference in gene networks: does topology matter? Genetica 129(1):83–103
Sommer RJ (2012) Evolution of regulatory networks: nematode vulva induction as an example of developmental systems drift. Adv Exp Med Biol 751:79–91
Sorrells TR, Johnson AD (2015) Making sense of transcription networks. Cell 161(4):714–723
Toma K, Hanashima C (2015) Switching modes in corticogenesis: mechanisms of neuronal subtype transitions and integration in the cerebral cortex. Front Neurosci 9:274

True JR, Haag ES (2001) Developmental system drift and flexibility in evolutionary trajectories. Evol Dev 3(2):109–119

ten Tusscher KH, Hogeweg P (2011) Evolution of networks for body plan patterning; interplay of modularity, robustness and evolvability. PLoS Comput Biol 7(10):e1002208

Wagner A (2008) Robustness and evolvability: a paradox resolved. Proc Biol Sci 275(1630): 91–100

Wagner A (2011) The origins of evolutionary innovations: a theory of transformative change in living systems. Oxford University Press

Wall ME (2011) Structure-function relations are subtle in genetic regulatory networks. Math Biosci 231(1):61–68

Wall ME, Dunlop MJ, Hlavacek WS (2005) Multiple functions of a feed-forward-loop gene circuit. J Mol Biol 349(3):501–514

Weiss KM (2005) The phenogenetic logic of life. Nat Rev Genet 6(1):36–45

Weiss KM, Fullerton SM (2000) Phenogenetic drift and the evolution of genotype-phenotype relationships. Theor Popul Biol 57(3):187–195

Wilke CO, Wang JL, Ofria C, Lenski RE, Adami C (2001) Evolution of digital organisms at high mutation rates leads to survival of the flattest. Nature 412(6844):331–333

Wotton KR, Jimenez-Guri E, Crombach A, Janssens H, Alcaine-Colet A, Lemke S, Schmidt-Ott U, Jaeger J (2015) Quantitative system drift compensates for altered maternal inputs to the gap gene network of the scuttle fly *Megaselia abdita*. eLife 4

Part III
Methods and Concepts

Mechanistic Models of Protein Evolution

David D. Pollock, Stephen T. Pollard, Jonathan A. Shortt and Richard A. Goldstein

Abstract Models of sequence evolution are used ubiquitously in biology from phylogenetic reconstruction to the analysis of adaptation, coevolution, and convergence. The structure of the model used affects these analyses, and it is therefore preferable to use good models. The field of molecular evolution is currently undergoing an important transformation due to large increases in the ability to collect and analyze massive amounts of data. Here, we briefly review the history of molecular evolution and then discuss how evidence of epistasis and convergent molecular evolution helps overturn traditional models of protein evolution. We conclude by discussing desired features in a simple mechanistic model of protein evolution that is more compatible with patterns observed in real and simulated protein evolution.

1 Introduction

The field of molecular evolution is concerned with how molecules evolve, and the forces that determine their evolutionary path. For functional molecules such as proteins, RNA, and regulatory DNA, the main factors include mutation, how neutral variants spread in a population, and the differential fitness of organisms containing variants. Because most individuals in most populations lived in the inaccessible past, the field is generally focused on using data from currently living (extant) populations to infer past processes, including gene duplications, population divergence (speciation), and phylogenetic relationships. Since we acquired the

D.D. Pollock (✉) · S.T. Pollard · J.A. Shortt
Department of Biochemistry and Molecular Genetics, University of Colorado School of Medicine, Aurora, CO 80045, USA
e-mail: David.Pollock@ucdenver.edu

R.A. Goldstein (✉)
Division of Infection & Immunity, University College London, London WC1E 6BT, UK
e-mail: r.goldstein@ucl.ac.uk

P. Pontarotti (ed.), *Evolutionary Biology: Self/Nonself Evolution, Species and Complex Traits Evolution, Methods and Concepts*,
DOI 10.1007/978-3-319-61569-1_15

ability to sequence proteins and genetic material, the power of utilizing this rich evolutionary record has been demonstrated repeatedly in terms of understanding phylogenetic relationships, predicting the timing and order of species divergence events such as those among human/chimp/gorilla ancestors, as well as looking specifically at evolutionary processes that occur at this molecular level such as when we predict the functional importance of individual positions or regions of molecules (Engelhardt et al. 2005).

These advances were achieved despite our lack of understanding of the mechanistic aspects of functional molecular evolution. We are currently in the midst of an important transformation of our mechanistic description of how functional molecules evolve, primarily through a better understanding of the importance of epistasis and coevolution, and how to incorporate them into evolutionary models. This transformation will strongly impact our ability to resolve conflicts in species and gene phylogenies, predict function and adaptation of function, predict the timing of molecular and systems-level evolutionary events, and predict the functional effect of mutations. To understand the mechanistic view, we first briefly review here the history of how molecular evolution has usually been described, followed by discussion of the molecular evidence that directly contradicts many existing evolutionary models. This will be followed by an overview of recent theoretical advances, which demonstrate the opportunities for simplification and better modeling, despite the potential for epistatic interactions to introduce overwhelming complexity.

2 A Brief History of Molecular Evolution

Since the work of Mendel and Morgan, it has been clear that mutations give rise to variants that can affect higher-level phenotypes such as the wrinkled surface of peas or the white color of a fly's eyes. Variants were reasonably considered as subject to natural selection that would alter the expected evolutionary trajectory of these variants, eliminating variants that give rise to deleterious phenotypes (negative or purifying selection) and increasing the frequency of initially rare variants that give rise to beneficial phenotypes (positive selection). The subsequent development of mathematical descriptions of these processes, especially through the pioneering work of Wright, Fisher, and Haldane, gave rise to the field of population genetics. Although Wright in particular advocated a stochastic approach to population genetics, which considered the role of a finite or even small population sizes, the simpler and more tractable deterministic approach assuming very large/infinite population sizes tended to dominate for decades during the middle of the twentieth century (reviewed in Fenster et al. 1997). Although the deterministic approach allows standing variation in populations due to e.g., mutation/selection balance and overdominance, and transient variation during selective sweeps, this was essentially

an adaptationist description of molecular evolution. The implicit assumption is that most traits are highly adapted if not perfectly optimized, and that positively selected changes, when they do rarely occur, are due to changes in an often vaguely described 'adaptive landscape,' whose hyperdimensionality did not prevent it from being represented in a two-dimensional plot.

As the scientific community learned the nature of transcription, translation, the genetic code, and how to sequence proteins and DNA, it became clear that some nucleotide variants were unlikely to impact phenotypic traits as much as others. This was then largely incorporated into population genetics theory, especially through the work of Kimura and his 'neutral theory' of molecular evolution (Kimura 1968). Once enough sequence data accumulated, it became clear that many variants, including synonymous and non-coding mutations, were also probably essentially neutral and should be included in neutral theory. Because of this, a better appreciation for the importance of stochastic processes began to dominate toward the end of the twentieth century. Although neutral theory was clearly anti-adaptationist in the sense that it was no longer viable to believe that *all* mutations were subject to meaningful levels of natural selection, in retrospect it was still highly adaptationist in the sense that the mutations that *mattered* for selection were all (or nearly all) considered to be deleterious. Most proteins were implicitly considered so optimized that the effect of a mutation, if it had an effect, must be deleterious. Despite the dominance of purely neutral theory during this time, theoreticians such as Gillespie and Ohta developed approaches that incorporated more variable degrees of selection, and laboratory biologists such as Powers and Watt evaluated potentially idiosyncratic systems in which protein variants appeared to be sustained (not fixed) due to varying selection along gradients and overdominant selection (Gillespie 1991; Ohta 1973).

The generality and purity of neutral theory was primarily broken by a combination of events. More examples of variants that were maintained by selection were discovered, and variation at the Major Histocompatibility Complex (MHC) played a big role in convincing many neutral evolution proponents that positive selection mattered (e.g., Mayer and Brunner 2007). MHC was important both because it contains a great deal of long-term standing variation (trans-species polymorphism) that cannot be explained by neutral theory, and because it is a good example of ongoing selection due to constantly varying host–parasite interactions. Other examples of molecular cat–mouse chases involving protein–protein interactions arose in viral proteins, venom–prey, and male–female or mother–offspring conflicts (e.g., Holding et al. 2016; Nourmohammad et al. 2016). All of these produce proteins with evolutionary histories of amino acid substitution rates greater than neutral expectations, what is called diversifying selection. Finally, with the pioneering work of Yang and others, it became possible to detect brief bursts of amino acid substitution greater than neutral expectation along ancestral branches in phylogenetic trees (e.g., Stéphane et al. 2002). These are generally interpreted as 'adaptive bursts,' driven by changing selective requirements (sometimes identified

and sometimes not). There are valid concerns about the statistical certainty with some of these approaches in some cases, but the overwhelming impression is that adaptive bursts are moderately common and identifiable within the vast diversity of life. For example, in our work with snake mitochondrial genomes, we identified perhaps the largest known temporary adaptive burst in multiple proteins at the base of snake diversification (Castoe et al. 2008). This example provides evidence of adaptation, but also a large enough sample of substitutions enriched for adaptive change that we can characterize the differences in evolutionary patterns in adaptive and nearly neutral substitutions (Ohta 1973). With the sequencing of the first two snake genomes, it has also held up as a general systems-level metabolic adaptive phenomenon.

Although Ohta's 'nearly neutral' theory combined with evidence for occasional bursts of adaptive change should give us a healthy respect for the fluctuating nature of molecular evolution, we would argue that nothing discussed so far deviates too much from a modified adaptationist paradigm. Yes, not all variation affects functional adaptation, and yes, sometimes the meaning of 'adapted' changes, but overall these arguments are compatible with mostly constant adaptive pressure at each amino acid position. Missing are explanations for observed epistasis and coevolutionary interactions among amino acid residues, related observations of heterotachy (changes in evolutionary rates over time), and why substitution rates among amino acids differ among positions in proteins. To begin finding explanations, we can move to three-dimensional and experimental considerations, and determine that particular amino acid substitutions have particular effects on stability (changes in the free energy of folding, as measured by $\Delta\Delta G$) or function (e.g., ligand binding or measurable enzymatic parameters). However, experimental results are expected to be of low resolution compared to the sensitivity of evolution and the effects measured in a laboratory may be different than what is important to selection. Experimental results are therefore considered to be generally informative but not definitive, and need to be interpreted with care. Furthermore, there are strong practical limits to the amount of data that can be collected. Computational predictions have questionable utility (Arenas et al. 2015; Bastolla et al. 2017; Thiltgen and Goldstein 2012), with their limited accuracy of $\Delta\Delta G$ prediction that further decreases with multiple substitutions, and binding strength predictions are even more limited. In any case, case-specific measurements do not amount to a general theory of how evolution proceeds (Bastolla et al. 2017). Knowledge progresses on all fronts, but we focus here on our multi-pronged approach, which involves empirical statistical modeling of sequence evolution in the context of phylogenetics, simulation of protein evolution as a thermodynamic system to better understand non-intuitive aspects of how functional molecules evolve, and the continued development and application of theory analogous to statistical mechanics to understand the mechanics of functional molecule evolution.

3 Modeling Principles and Empirical Statistical Models of Molecular Sequence Evolution

3.1 Empirical Statistical Modeling and Phylogenetics

Because we are advocating for more mechanistic models of protein evolution, it is useful at this point to discuss how empirical statistical models of molecular evolution are compared, what we mean by 'mechanistic' models, and how mechanistic models differ from phenomenological models. The statistical models that we discuss are those used to analyze sequence data, and the fundamental calculation in these models is to determine the probability that the data would have been produced if the model had been operating with particular parameter settings. In a frequentist approach, one compares models by finding the parameter combination that is most likely to have produced the data, while a Bayesian approach compares models by integrating the posterior probability over reasonably likely parameter settings, and simultaneously incorporating prior probabilities of models and parameter settings. Good reviews of this topic can be found elsewhere (Goldman and Yang 2008; Thorne 2000).

Empirical statistical models can differ substantially in their theoretical foundations. Here, we emphasize the difference between phenomenology and mechanism-based models. We define pure phenomenology as simple, theory-free measurement, such as might be done to count the number of individuals in a population of organisms, or to measure the height of a person. In the context of molecular evolution, an example of a mostly phenomenological approach might be to measure the frequencies of amino acids at each position in a sequence alignment, or in each protein, or in an entire set of proteins. Other alignment-based measures such as the fraction of sites that are unvarying or correlations between amino acids observed at different sites might also be considered primarily phenomenological. The statistical questions for a purely phenomenological measurement are mostly limited to reproducibility, accuracy, and perhaps how the quantity changes over time. These measurements are a good start, but are of limited utility unless we are able to interpret their meaning, significance, and range of applicability. We usually would prefer to understand what the site-specific amino acid frequencies can tell us about the protein, and its relationship with other proteins, or use the divergence between sequences to estimate evolutionary distances.

An alternative to pure phenomenology is to represent the salient aspects of the process that resulted in the current sequences using a model that embodies some theoretical mechanism. Choosing which aspects to include and how they are represented generally involves a mixture of empirical (phenomenological) observations and (mechanistic) representations of the underlying biology. For example, modern substitution matrices are usually estimated using a phylogenetic tree, which can be considered a mechanistic model for how the species diverged during the course of evolution, part of the process by which the sequences were produced. Another example is the analysis of the DNA sequences that code for proteins using

the genetic code, which constitutes a mechanistic model of how DNA is converted to protein. The most popular rate models for representing protein evolution include the empirical observations that some amino acids are more common than others, some substitutions are more common than others, and some sites change more frequently than others. These are represented phenomenologically with a set of amino acid equilibrium frequencies, a symmetric 'exchangeabilities' matrix, and a (generally Gamma distributed) distribution of rates. Models used for identifying positive selection include the mechanistic consideration that DNA substitutions in protein-coding regions can be synonymous or non-synonymous, but often ignore the observation that some amino acid changes are more likely than others. In both these cases, the probability of substitution from one amino acid or from one nucleotide to another is simply inferred from the number of sequence differences, and thus is a phenomenological component of the model.

There can be borderline cases as well, where the empirical results can be justified from the underlying biology; for instance, the difference between transition rates (between purines A and G or pyrimidines C and T) and transversion rates (between purines and pyrimidines) can be rationalized by considering the chemical structure of DNA. There are also numerous instances where these phenomenological representations are used to gain mechanistic insights, such as in inferences of positive selection or in the analysis of substitution matrices to determine physicochemical protein properties (Koshi and Goldstein 1997; Koshi et al. 1997). There are, conversely, always observations and biological knowledge that are ignored by these models. Many of these simplifications were required due to our lack of knowledge of molecular evolution and the limits of computational resources and sequence data availability at the time in which the models were constructed. In other instances, the phenomenological representations are in conflict with basic molecular biophysics, or are internally inconsistent. For instance, the site-specific rates of amino acid substitutions reflect the degree of selection acting on that site, resulting in a restriction in the amino acids that are appropriate for that site (generally not modeled), which causes the reduced substitution rate (which is modeled). It is, in general, impossible to reconcile the empirical amino acid equilibrium frequencies in these models with the observed overall substitution rate (Goldstein and Pollock 2016a).

Historically, the basis of empirical statistical models used in molecular evolutionary analysis has nearly always been that there are a certain number of states (e.g., nucleotides, amino acids, or codons) with constant substitution rates of exchange among them. These substitution rates might be different for different classes of sites or different genes or genomic regions and occasionally have been allowed to change at discrete points on the phylogenetic tree. Substitution probabilities, $P(t)$, along branches of length t in the phylogenetic tree were then usually (and often still are) calculated first by spectral decomposition to obtain the eigenvalues (Λ) and eigenvectors (S) of the instantaneous rate matrix (Q), and then by calculating $P(t) = Se^{\Lambda t}S^{-1}$. The implicit mathematically necessary but rarely discussed assumption in these approaches is that Q holds over long periods of time.

Thus, even if the average Q is well estimated, it may not be an accurate reflection of the process at any single point in time. This is a problem for a number of reasons, chief among them being that the parameters of the instantaneous rate matrix identified will depend on the particular phylogenetic tree considered, and how long the branches on that tree are, and that this averaged rate matrix may not be accurate for any site at any time during the evolutionary process, obscuring the actual nature of the evolutionary change. This problem was to some extent recognized early on when it was found that PAM matrices determined using many closely related proteins produced very different results than BLOSUM matrices determined with more distantly related proteins (Brenner et al. 1998; Henikoff and Henikoff 1992; Wilbur 1985).

In the last decade or so, more and more Bayesian approaches have incorporated augmented data methods that allow one to avoid time-consuming and computationally expensive spectral decomposition and repeated matrix-vector-matrix multiplication to obtain the substitution probabilities along branches, $P(t)$. Focusing on our own method encoded in the program *PLEX*, we partially sample substitutions to the nearest short branch region to augment the data; in combination with uniformization (de Koning et al. 2010, 2012) of substitution rates, this can be much faster than complete augmentation of fully specified substitution histories. This program was designed to allow greater flexibility in allowing substitution probabilities to differ among positions in a molecule and over time, but this can create an explosion of complexity, or at least an explosion in the number of adjustable model parameters, and the question is how to develop appropriate models that reflect the underlying biology. Models that include rate heterogeneity (e.g., Halpern and Bruno 1998; Koshi and Goldstein 1995; Koshi et al. 1999; Lartillot and Philippe 2004; Tamuri et al. 2012), for example, are limited by the amount of sequence necessary to estimate parameters. They also treat each site in a protein as independent from all others without considering the protein molecule as a whole. Thus, the consequences of selection are modeled but the mechanism of selective action is still treated as an unknown. For reasons that will become clear below, we do not think that continuing to divide sites into more and more small substitution categories is a fruitful or mechanistically justified approach.

3.2 *Epistasis and Coevolution*

The concept of epistasis, that the effect of variants in combination is not always an additive sum of their individual effects, is well known from the early days of genetics and biochemistry. From a biochemical perspective, function and three-dimensional structure arise from interactions among amino acid residues, and if one residue in a protein changes, it is natural to presume that it may alter the effect of a change at another position. Experimentally, epistasis is easily detected by finding mutants in a protein that are deleterious to function, and then selecting for ‘compensatory’ mutants that allow the protein to recover (Stephan 1996).

Early mutagenesis studies in lysozyme also clearly established the prevalence of compensatory relationships among amino acids in protein cores (Baldwin et al. 1996). As with the adaptationist/neutralist arguments of the last century, however, questions arise about how often epistatic changes occur during evolution, and how important the role of positive selection is in preserving these changes. The observation of rampant epistasis among amino acids in proteins (Breen et al. 2012) promotes a more nuanced view on protein evolution and the substitution process, a view in which the probability of substitution at each site is dependent on which amino acids occupy nearby and other interacting sites. A few evolutionary concepts have been extremely important to understanding the role of epistasis in functional molecular evolution: molecular coevolution, deep evolutionary inference, and convergence. The concepts of coevolution and epistasis overlap (Pollock and Pollard 2016), but here we will view coevolution as the long-term evolutionary consequences of epistatic effects and define it (as in Pollock et al. 1999) as what occurs when a substitution at one position alters the propensity to accept substitutions at other positions. It is worth noting, however, that with epistatic changes (such as the biochemist's compensatory changes), it is usually considered that at least one change has a phenotypic or selective effect, whereas coevolution can proceed even if every substitution involved is entirely neutral. Past coevolution between individual residue positions is difficult to prove, especially for small or moderately diverse sequence datasets, but the cumulative evidence across many residues that coevolution is pervasive, and evidence that there is a strong relationship between coevolution and structural proximity, is overwhelming (e.g., Pollock et al. 1999). It has also been noted that residues that are pathogenic in humans are surprisingly often the most frequent residues in related species (Stéphane et al. 2002). Furthermore, even those trying to downplay the role of epistasis and fluctuating amino acid propensities in protein evolution have tended to produce data that confirm it, reducing the argument to questions of the size of the fluctuations under different conditions (Ashenberg et al. 2013; Pollock and Goldstein 2014). Finally, recent papers have demonstrated the ability to filter coevolutionary information in very large and diverse bacterial phylogenies to find sufficiently adjacent amino acid residues that they can be used to predict protein structure (Lunt et al. 2010; Morcos et al. 2011; Weigt et al. 2009).

The problems for simple evolutionary mechanistic theory that arise from deep (ancient) protein evolutionary inferences were recently reviewed (Goldstein and Pollock 2016a), and we refer readers to that paper for details. However, the basic problem is that mutation rates are sufficiently large that neutral substitutions should have saturated individual positions, such that multiple substitutions will have thoroughly obscured the utility of neutral substitutions for inferring deep phylogenetic relationships that we are often interested in (such as mammalian divergence or deeper). Functional molecules such as proteins do not appear to saturate so quickly though (partly reflected in the differences between PAM and BLOSUM matrices, described above), and molecular phylogenetic analyses have long relied on such molecules to resolve ancient phylogenetic questions. Although we agree that this observation does not prove epistasis (McCandlish et al. 2016) as claimed

by Bazykin et al. (2007), it seems likely that the level of coevolution among amino acids that has already been demonstrated is sufficient to cause this effect.

3.3 The Importance of Convergence

Convergence "occurs when two biological traits in two separate lineages independently evolve to similar end points" (Pollock and Pollard 2016). It has long played an important role in evolutionary theory at the organismal level because convergent evolution of similar complex morphologies is seen as a strong sign of adaptation to similar selective forces in the environment (Harvey and Pagel 1991; Mayr 1963). Convergence at the molecular level has been seen as a relatively rare phenomenon, but the huge increase in genomic data and dense taxonomic sampling in recent years has led to an upswing of papers detecting molecular convergence. These efforts have seen a number of false starts, however, beginning with mitochondrial genomes (Rokas and Carroll 2008) and continuing with convergence in echolocating mammals (Parker et al. 2013). An obstacle is that current evolutionary models do a poor job at predicting levels of convergence (Castoe et al. 2009), but further problems arise when indirect methods of detecting convergence are used, and detection of convergence can be conflated with phylogenetic errors (Castoe et al. 2009; Thomas and Hahn 2015; Zou and Zhang 2015a).

Molecular convergence is also beginning to play a big role in understanding epistasis and the mechanics of nearly neutral molecular evolution, and that is because of its relationship to propensity and constraint (Goldstein et al. 2015). To see this, consider evolution at three sites, all with resident amino acid alanine (A), as shown in Fig. 1. Considering only the six amino acids shown (resident alanine and five possible substitutions), at site 1 the amino acids have equal propensity, and nearly equal (1 out of 5) probabilities of convergence (modified only by differences in mutation rates, particularly transition and transversion rates, as shown). At site 2, however, the propensities for S and T (and the resident, A) are much higher than for P, V, and G, meaning that substitutions at the site are almost completely constrained to S and T. If these propensities do not change over time, then the probability of convergence along two different evolutionary lineages given substitutions along both lineages at this site (both with ancestral state A) is nearly 50% (modified by relative mutation rates) because there are only two practical choices of substitutions. Site 3 has the same relative distribution of propensities among the amino acid alternatives to the resident amino acid as site 2 does, but the resident amino acid is far more fit. Thus, site 2 and site 3 will have the same probability of convergence if there are two substitutions at that site at different lineages in a phylogenetic tree, but site 3 is much less likely to substitute, and thus to converge, at all.

Understanding convergence as a biological consequence of constraint allows us to better understand why current evolutionary models do such a poor job at predicting convergence levels. Firstly, when protein positions with different levels of constraint are combined, the combined average model is often less constrained than

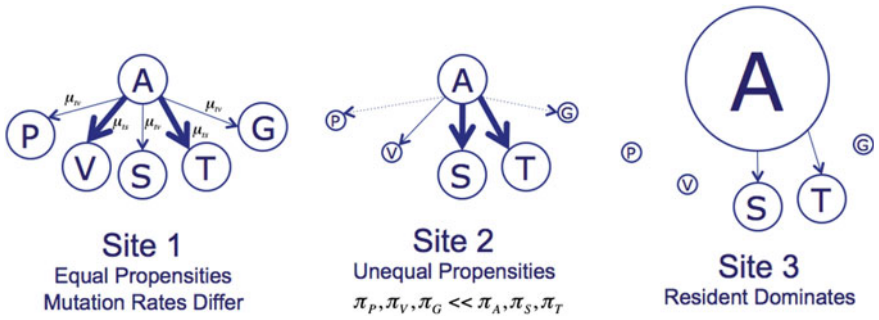


Fig. 1 Convergence depends on constraint. In the examples, at site 1 the amino acids shown (A, alanine; P, proline; V, valine; S, serine; T, tyrosine; and G, glycine) have equal propensities (illustrated as size of *circles*), so evolution is unconstrained, and substitution (indicated by thickness of *arrows*) is determined by the mutation rate for each type of mutation (transition, μ_{ts}; or transversion μ_{tv}) required to change the codon from alanine. At site 2, the greater propensities of serine and tyrosine mean most substitutions will be to one of these two amino acids, and the remaining substitution rates are reduced to thin or *dashed lines*. At site 3, there are few substitutions due to the overwhelmingly large propensity of the resident amino acid, alanine; substitutions to serine and tyrosine are reduced, and other substitutions are so rare as to be essentially absent

any of the individual models. Consider one position that can substitute from alanine to serine or tyrosine, and another that can substitute from the same starting point, alanine, to proline and valine. If substitution probabilities between these sites are combined, one might expect that they both could substitute to any of the four amino acids, reducing the expected probability of convergence by half (25% expected probability rather than the actual 50%). A similar logic applies if the process of substitution changes at a single position at two very distant time points. Applying the previous example, the actual probability of convergence is near 50%. The position can substitute from alanine to serine or tyrosine over a *short-time separation*, but if the position has switched to only accepting proline and valine at some *distant point* on the phylogenetic tree, then the probability of convergence would have fallen to zero. Thus, one can understand that epistasis and coevolution, which by definition alter substitution probabilities, have the necessary effect of reducing the probability of convergence over time (Goldstein et al. 2015). If the evolutionary process differs among positions and over time, which appears to be the case (Goldstein et al. 2015), static evolutionary models would appear to have almost no hope of predicting levels of convergence, although just as a stopped clock may correctly predict the time of day, they may occasionally do so by chance. It is also worth noting that because neutral convergence is equivalent to homoplasy, and inferring levels of homoplasy is one of the main points of evolutionary models in phylogenetics (Castoe et al. 2009), this result has implications for the reliability of phylogenetic inference using functional molecules, although the extent of the problem is currently uncertain.

As an aside, it should be noted that because evolutionary models do such a poor job of predicting levels of convergence, they probably cannot be used for this purpose with any degree of reliability. The problem is particularly insidious because commonly used standard models of amino acid substitution (such as JTT Jones et al. 1992 or WAG Whelan and Goldman 2001); see (Goldstein and Pollock 2016a review) are so broad and unconstrained that they actually do not change predicted convergence levels much over time, even with different ancestral amino acids and the inclusion of the genetic code (Goldstein et al. 2015). Thus, a user would be highly confident of their results even when they should not be. In contrast, moderately constrained but time-invariant models such as CAT models (Quang et al. 2008) or Halpern-Bruno models (Halpern and Bruno 1998) interact strongly with the genetic code and predict that global convergence levels will decrease over time even though the process at each site is not changing. For this reason, great care is needed to distinguish the effect of lowered convergence levels due to previous divergence and the structure of the genetic code; the trivial convergence caused by prior substitution is probably the strongest signal in any protein dataset (Goldstein et al. 2015), and does not demonstrate epistasis and fluctuating constraint. We therefore do not think that convergence predictions based on incorrect models and branch lengths, as in Zou and Zhang (2015b, 2017), are reliable. Instead, we recommend that branch pairwise convergence levels should be compared to branch pairwise double divergence levels, and both only for cases of a common ancestral amino acid (Castoe et al. 2009; Goldstein et al. 2015; Mendes et al. 2016; Zou and Zhang 2017). Such convergence/divergence measures are also not subject to error due to fluctuation of average branch lengths among genes or gene regions. Because convergent molecular evolution may occur in response to both adaptive and non-adaptive causes, it is critical that we obtain a better understanding and use good means to predict non-adaptive convergence, the better to detect adaptive convergence when it does appear. Understanding the difference between adaptive convergence and non-adaptive convergence requires a better understanding of the evolutionary forces that govern the substitution process and the variability in site-specific constraints over time.

4 The Evolutionary Stokes Shift and the Role of Thermodynamic Models

Sensitive readers may at this point be slightly concerned because if, as seems to be the case, amino acid propensities and therefore substitution rates and convergence levels fluctuate due to pervasive epistasis, then it would seem that there is little predictability to molecular evolution, especially if these fluctuations are random. However, the fluctuations are not actually random in the sense that they are undirected or unconstrained. A clear sign of this is the evolutionary Stokes shift (Pollock and Goldstein 2014; Pollock et al. 2012). The basic idea of the

evolutionary Stokes shift is that when an amino acid is substituted at a site, proteins tend to equilibrate to the newly resident amino acid through epistatic substitutions at other sites (Pollock et al. 2012). At the same time, other non-resident amino acids, including the previously resident amino acid, are not necessarily stabilized and may wander away from or into stability states comparable to the resident amino acid (Pollock et al. 2012), thus affecting the probability of substitution. The two components of the evolutionary Stokes shift can be termed 'contingency' (the necessary wandering of an amino acid in stability space to a similar stability level as the resident amino acid) and 'entrenchment' (the tendency of epistatic changes to stabilize the newly resident amino acid), and it has been shown that mutations that fix are contingent on previous substitutions (Shah et al. 2015).

The evolutionary Stokes shift was originally discovered as a consequence of modeling the evolution of functional proteins as thermodynamic, folded entities (Pollock et al. 2012). Surprisingly, even a quite simple energy function, in conjunction with the need to be stable in a particular fold and not spend much time in other folds, can produce patterns of contingency and entrenchment (Pollock et al. 2012). Indeed, modifications of the model and inclusion of functional effects directly (for example, through ligand binding) do not seem to strongly affect the general result (Goldstein unpublished data, Shah et al. 2015), and the expected decrease in reversion rates after substitution has been shown to generalize to arbitrary fitness landscapes (McCandlish et al. 2016). The direction of predicted stabilities between pairwise differences in diverged real proteins that had been crystallized showed remarkable agreement with our thermodynamic model proteins (Pollock et al. 2012), and even skeptics have tended to produce measured stability data for substitutions in divergent proteins that are in rough agreement with theoretical predictions (Ashenberg et al. 2013; Doud et al. 2015; Pollock and Goldstein 2014; Pollock et al. 2012), although the number of protein measurements is necessarily small.

Because modeling the evolution of functional proteins as thermodynamic, folded entities appears to reproduce many important features of protein evolution that are not explained by static models (Goldstein et al. 2015; Pollock et al. 2012), it is worthwhile to consider further what these models are doing and how we are using them. In Fig. 2, it can be seen that while the frequencies in the WAG model are distributed relatively evenly and are constant, the thermodynamic models (sometimes call Stokes-Fisher models) produce highly variable frequencies at a single position, over time changing the relative magnitudes and often the order or amino acid propensities. Major differences also occur among positions (Pollock et al. 2012). These differences occur despite the fact that the underlying amino acid interaction model that drives stabilities, a 20×20 interaction matrix, is no more complicated than the 20×20 WAG substitution matrix. The difference lies in the mechanism, which includes the requirement that the propensities and substitution rates are caused by the effect of substitutions or potential substitutions on the stability of the entire protein sequence.

The use of a simplified thermodynamic models to simulate evolution can reproduce some of the most perplexing features of protein evolution (epistasis,

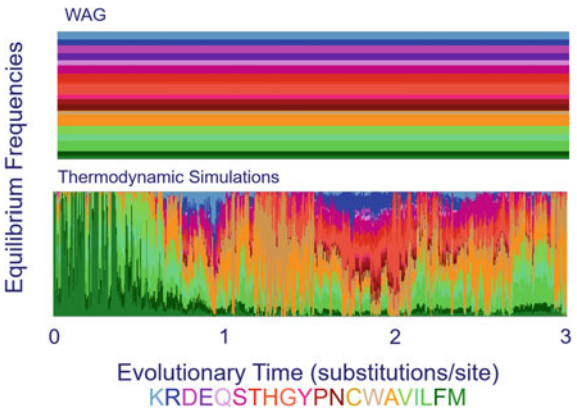


Fig. 2 Changes in equilibrium amino acid frequencies (propensities) over time. Results for the WAG model (*top*) and in thermodynamic simulations (*bottom*). WAG frequencies stay constant over time, while in the thermodynamic model constraints and equilibrium frequencies vary greatly

coevolution, the evolutionary Stokes shift, and changing nearly neutral convergence over time), and may indicate that the thermodynamic three-dimensional folded nature of functional molecules has an impact on their evolution. It also helps to illustrate the utility of developing a more detailed mechanistic approach. Simple mechanisms or processes can produce complicated-seeming results that are nearly impossible to sort out from a purely empirical perspective, but simplicity can be revealed and predictive power greatly improved by focusing on understanding the mechanism that produced these results. Such a scenario seems to be the case with molecular evolution; without a mechanism for how substitution rates are generated, we are faced with trying to find rate matrices for each site, and then for shorter and shorter periods of time, until the point where there is no more data to collect and resolution is still lacking. For example, Fig. 3 shows an example where the substitution rate between threonine and alanine appears to change across the vertebrate mitochondrial tree. The threonine to alanine rate (and the rate of reversion) seems much higher in birds than it is in mammals, and very different amino acid propensities would be predicted if the range of taxa were birds, versus birds plus crocodiles, turtles and mammals, or among all the (tetrapod) vertebrates.

With a mechanism, however, it is possible that data collection can be focused on understanding the simpler question of how amino acids interact, which may be informative across all sites. To be clear, we are not saying that our model demonstrates that there is a single distribution of interactions at all sites; our model runs on a single distribution and reproduces many salient features of real protein evolution, indicating that careful work will need to be done to see if different context-dependent interaction models are truly needed. We view the thermodynamic models as more of a null hypothesis indicating the complexity that can be produced through thermodynamic evolution alone; to demonstrate a strong effect of context-dependence, a truly site-specific effect on the mechanism, one now needs more than just to show that their average rates, observed over a finite time, are significantly different.

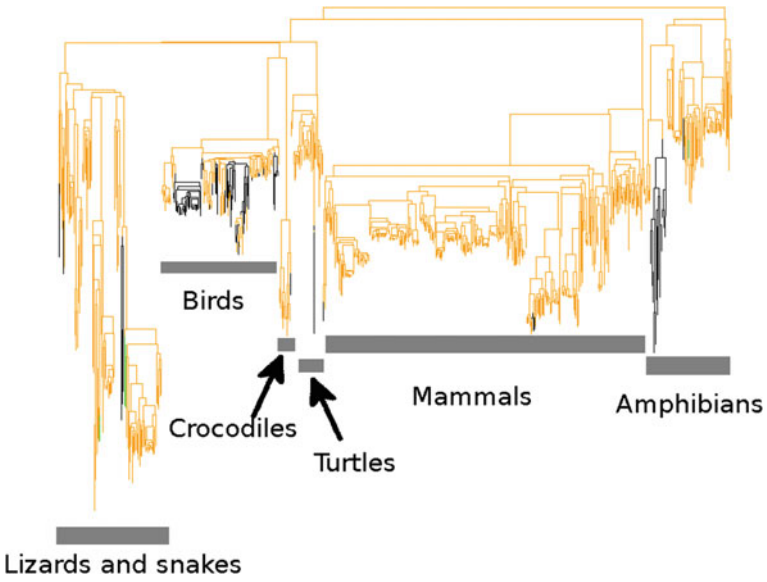


Fig. 3 Substitutions along a site in cytochrome c oxidase from tetrapod mitochondria. Branches are colored by *orange* = threonine, *black* = alanine, and *green* = asparagine. Alanine is produced from threonine by a first codon position transition mutation, while asparagine is produced by a second codon position transversion mutation

Another key feature of our thermodynamic model is that at its base it is a Hamiltonian-Potts model, i.e., an energy model. However, because it is a selected energy model, the amino acid propensities and interactions are not directly inferable from the energy function, and vice versa, as would be the case, for example, in inferring molecular conformation distributions. In the next section, we focus on explaining recent work toward understanding the theoretical dynamics of this situation, and how such theory can be used to inform on relative substitution rates and the strength of the evolutionary Stokes shift (Goldstein and Pollock 2016b).

5 Toward a Statistical Mechanics Theory of Molecular Sequence Evolution

5.1 Introduction to the Statistical Mechanics of Evolution

Up to this point, we have discussed mostly the evidence from statistical empirical models and from thermodynamic simulation models that jointly point to the idea that there is something about thermodynamics that may explain important features of molecular evolution that are incompatible with current models. In this section, we consider the utility of a statistical mechanics framework for this explanation,

following our recent work developing this theory (Goldstein and Pollock 2016b). The topic is difficult conceptually because we need to simultaneously include terms from classical statistical mechanics, thermodynamics, and transition state theory to discuss the folding of molecules and their ability to act as catalysts, but also develop terminology to discuss the application of statistical mechanics and transition state theory to the evolution of the sequences that code for these same proteins. For example, for this reason, we discuss the stability of sequence $\mathbf{X}$, as $\Phi(\mathbf{X})$, which is defined to be in the same direction as fitness (increases in stability correspond to increases in fitness) and is simply the negative of $\Delta G_{folding}(\mathrm{X})$, the free energy change of sequence $\mathbf{X}$ upon folding to a structure that carries out a function. This allows a smooth transition in discussion as to how the results may extend to other fitness functions, including ligand binding, catalysis, and signal propagation.

To separate out the structural component that underlies nearly all protein function, we consider that the probability that the protein is folded at thermodynamic equilibrium is equivalent to fitness (Goldstein 2011; Pollock et al. 2012; Williams et al. 2006). This is partly based on our experience that when fitness is incorporated into thermodynamic models, proteins will crystallize into a single-folded structure that tends to be marginally stable, as do real proteins (Taverna and Goldstein 2002). Because we want to understand how substitution rates at a site come about, we can focus theoretical attention on a single representative site, k, and consider how substitutions at this site alter overall protein stability, and on this basis whether they will be accepted during the course of evolution. This is an entirely reasonable proposition because we have defined fitness to be determined by protein stability. Indeed, we unsurprisingly find in our simulations that substitutions between two amino acids can be extremely well predicted by the distributions of relative stability contributions made by each amino acid, and using Kimura's formula to predict substitution probability from effective population size (N_e) and relative fitness.

It is useful to pause here a moment and consider what this may indicate about real proteins. The distributions of contributions to stability for amino acids in real proteins are unlikely to be exactly what we get in our simulations because it may depend on the target structure and the true interaction energies between these amino acids and may be modified by other functional constraints. We see wide variation in the distributions depending on which amino acids are involved and the average rate of substitution at a site, however, and it seems reasonable that these are factors in real proteins as well. Although we do not know the magnitude of these fluctuations or the rate at which individual sites in real proteins move in stability space, the fluctuation of stability contributions observed in simulations matches the observation of fluctuations in stability seen in real proteins and explains the observed decrease in convergence probabilities with time of divergence (Goldstein et al. 2015).

5.2 *Can We Obtain a Mechanistic Entropic Explanation for the Magnitude of the Evolutionary Stokes Shift and How It Explains Substitution Rates?*

The substitution rate between two amino acids depends on the amount of variation in the stability of the resident amino acid and on how much covariation there is between the stability contributions of possible replacements. Ongoing work (Goldstein and Pollock 2016b) centers around the idea that we can convert sequence space into a statistical mechanics framework by considering, in addition to the stability contribution of an individual site, the stability contribution of the remaining interactions not involving the site, the latter corresponding to the 'Bath' in analogy to classical statistical mechanics. Both of these sum up to the total stability: $\phi = \phi_{k,\alpha} + \phi_{k,Bath}$ (Fig. 4).

The mechanistic process can then be visualized first by considering the forces of sequence entropy (the number of sequences, Ω) and selection (depending on N_e and other factors), which conspire to tightly constrain total stability (ϕ, Fig. 5). There are no selective constraints on the relative proportion of $\phi_{k,\alpha}$ and $\phi_{k,Bath}$ that sum up to ϕ, however, and so for that proportion entropy alone dominates. Because there are so many more interactions involving the bath contribution to stability, it tends to move toward lower stability values that have larger number of sequences (Fig. 5). It is only able to do this, however, if the individual site contribution to stability increases to compensate and keep the total approximately constant.

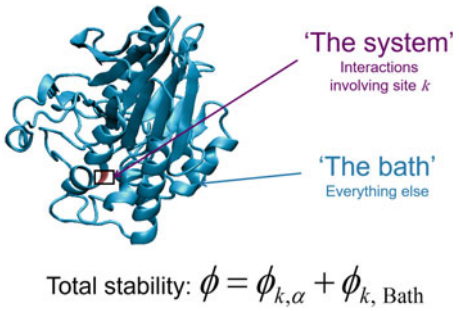


Fig. 4 Total stability. The total stability is divided into the contribution from the amino acid at a site k and stability contributions due to interactions among amino acids not including site k

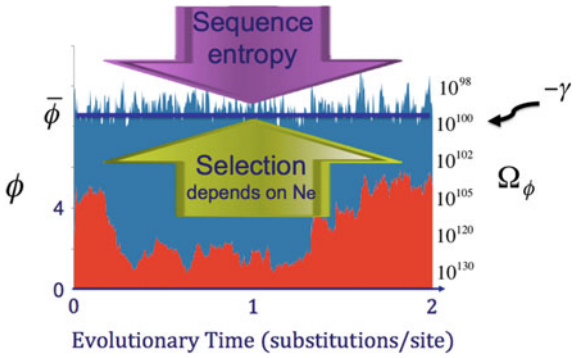


Fig. 5 Constraints on stability. The portion of the total stability occupied by the site-specific interactions and the remaining bath interaction. The constraints on stability due to entropy and selection are indicated. Bath and site-specific interactions are shown in *blue* and *red*, respectively. Plausible example numbers of sequences at each stability value (to the *left*, in kcal/mol) are shown to the *right*

6 Conclusion

We have described here the role of mechanisms and phenomenological descriptions as components of statistical empirical models, and described recent developments in mechanistic descriptions of the evolution of functional molecules, such as proteins. The role of fast thermodynamic evolutionary simulations is pivotal in discerning how proteins, as thermodynamic entities, should evolve, and what sorts of effects thermodynamics have on evolutionary outcomes. These thermodynamic models provide a potential explanation for patterns of epistasis, coevolution, average substitution rate differences over long periods of time, molecular convergence changes over time, and the evolutionary Stokes shift, which are fundamental problems for current statistical empirical models. We believe that a statistical mechanic-like treatment of protein sequence evolution points to a mechanistic explanation for many, if not all, of these phenomena, with the added benefit that it may greatly reduce the number of phenomenological parameters needed for future statistical empirical models of evolution.

Acknowledgements We acknowledge the support of the Medical Research Council (UK) (MC_U117573805) to RAG and the National Institutes of Health (NIH; GM083127 and GM097251) to DDP.

References

Arenas M, Agustin S-C, Bastolla U (2015) Maximum-likelihood phylogenetic inference with selection on protein folding stability. Mol Biol Evol 32:2195–2207

Ashenberg O, Gong L, Bloom J (2013) Mutational effects on stability are largely conserved during protein evolution. Proc Natl Acad Sci 110:21071–21076

Baldwin E, Xu J, Hajiseyedjavadi O, Baase W, Matthews B (1996) Thermodynamic and structural compensation in "size-switch" core repacking variants of bacteriophage T4 lysozyme. J Mol Biol 259:542–559

Bastolla U, Dehouck Y, Echave J (2017) What evolution tells us about protein physics, and protein physics tells us about evolution. Curr Opin Struct Biol 42:59–66

Bazykin G, Kondrashov F, Brudno M, Poliakov A, Dubchak I, Kondrashov A (2007) Extensive parallelism in protein evolution. Biol Direct 2:1–13

Breen M, Kemena C, Vlasov P, Notredame C, Kondrashov F (2012) Epistasis as the primary factor in molecular evolution. Nature 490:535–538

Brenner S, Chothia C, Hubbard T (1998) Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. Proc Natl Acad Sci USA 95:6073–6078

Castoe TA, Jiang ZJ, Gu W, Wang ZO, Pollock DD (2008) Adaptive evolution and functional redesign of core metabolic proteins in snakes. PLoS ONE 3:e2201

Castoe TA, de Koning A, Kim H-MM, 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 106:8986–8991

Doud M, Ashenberg O, Bloom J (2015) Site-specific amino acid preferences are mostly conserved in two closely related protein homologs. Mol Biol Evol 32:2944–2960

Engelhardt BE, Jordan MI, Muratore KE, Brenner SE (2005) Protein molecular function prediction by Bayesian phylogenomics. PLoS Comput Biol 1:e45

Fenster C, Galloway L, Chao L (1997) Epistasis and its consequences for the evolution of natural populations. Trends Ecol Evol Amst 12:282–286

Gillespie J (1991) The causes of molecular evolution. Oxford University Press, New York

Goldman N, Yang Z (2008) Introduction. Statistical and computational challenges in molecular phylogenetics and evolution. Philos Trans R Soc B Biol Sci 363:3889–3892

Goldstein R (2011) The evolution and evolutionary consequences of marginal thermostability in proteins. Proteins Struct Funct Bioinform 79:1396–1407

Goldstein R, Pollock D (2016a) The tangled bank of amino acids. Protein Sci 25:1354–1362

Goldstein R, Pollock D (2016b) Sequence entropy and the absolute rate of amino acid substitutions. bioRxiv

Goldstein R, Pollard S, Shah S, Pollock D (2015) Nonadaptive amino acid convergence rates decrease over time. Mol Biol Evol 32:1373–1381

Halpern A, Bruno W (1998) Evolutionary distances for protein-coding sequences: modeling site-specific residue frequencies. Mol Biol Evol 15:910–917

Harvey P, Pagel M (1991) The comparative method in evolutionary biology. Oxford University Press, New York

Henikoff S, Henikoff J (1992) Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci USA 89:10915–10919

Holding ML, Biardi JE, Gibbs LH (2016) Coevolution of venom function and venom resistance in a rattlesnake predator and its squirrel prey. Proc R Soc B 283:20152841

Jones DT, Taylor WR, Thornton JM (1992) The rapid generation of mutation data matrices from protein sequences 8:275–282

Kimura M (1968) Evolutionary rate at the molecular level. Nature 217:624–626

de Koning A, Gu W, Pollock DD (2010) Rapid likelihood analysis on large phylogenies using partial sampling of substitution histories. Mol Biol Evol 27:249–265

de Koning A, Gu W, Castoe TA, Pollock DD (2012) Phylogenetics, likelihood, evolution and complexity. Bioinformatics 28:2989–2990
Koshi J, Goldstein R (1995) Context-dependent optimal substitution matrices. Protein Eng 8:641–645
Koshi J, Goldstein R (1997) Mutation matrices and physical-chemical properties: correlations and implications. Proteins 27:336–344
Koshi J, Mindell D, Goldstein R (1997) Beyond mutation matrices: physical-chemistry based evolutionary models. Genome Inf Ser Workshop Genome Inf 8:80–89
Koshi JM, Mindell DP, Goldstein RA (1999) Using physical-chemistry-based substitution models in phylogenetic analyses of HIV-1 subtypes 16:173–179
Lartillot N, Philippe H (2004) A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. Mol Biol Evol
Lunt B, Szurmant H, Procaccini A, Hoch JA, Hwa T, Weigt M (2010) Chapter two inference of direct residue contacts in two-component signaling (sciencedirect)
Mayer F, Brunner A (2007) Non-neutral evolution of the major histocompatibility complex class II gene DRB1 in the sac-winged bat Saccopteryx bilineata. Heredity 99:257–264
Mayr E (1963) Animal species and evolution. Harvard University Press, Cambridge, MA
McCandlish DM, Shah P, Plotkin JB (2016) Epistasis and the dynamics of reversion in molecular evolution. Genetics 203:1335–1351
Mendes FK, Hahn Y, Hahn MW (2016) Gene tree discordance can generate patterns of diminishing convergence over time 33:3299–3307
Morcos F, Pagnani A, Lunt B (2011) Direct-coupling analysis of residue coevolution captures native contacts across many protein families. Proc Natl Acad Sci
Nourmohammad A, Otwinowski J, Plotkin JB (2016) Host-pathogen coevolution and the emergence of broadly neutralizing antibodies in chronic infections. PLoS Genet 12:e1006171
Ohta T (1973) Slightly deleterious mutant substitutions in evolution. Nature 246:96–98
Parker J, Tsagkogeorga G, Cotton J, Liu Y, Provero P, Stupka E, Rossiter S (2013) Genome-wide signatures of convergent evolution in echolocating mammals. Nature 502:228–231
Pollock D, Goldstein R (2014) Strong evidence for protein epistasis, weak evidence against it. Proc Natl Acad Sci 111:E1450–E1450
Pollock DD, Pollard S (2016) Parallel and convergent evolution. Academic Press, Oxford
Pollock D, Taylor W, Goldman N (1999) Coevolving protein residues: maximum likelihood identification and relationship to structure. J Mol Biol 287:187–198
Pollock D, Thiltgen G, Goldstein R (2012) Amino acid coevolution induces an evolutionary stokes shift. Proc Natl Acad Sci 109:E1352–E1359
Quang SL, Gascuel O, Lartillot N (2008) Empirical profile mixture models for phylogenetic reconstruction 24:2317–2323
Rokas A, Carroll S (2008) Frequent and widespread parallel evolution of protein sequences. Mol Biol Evol 25:1943–1953
Shah P, McCandlish DM, Plotkin JB (2015) Contingency and entrenchment in protein evolution under purifying selection. Proc Natl Acad Sci 112:E3226-35
Stephan W (1996) The rate of compensatory evolution. Genetics 144:419–426
Stéphane A-B, Yang Z, Huelsenbeck J (2002) Effects of models of rate evolution on estimation of divergence dates with special reference to the metazoan 18S ribosomal RNA phylogeny. Syst Biol 51:703–714
Tamuri AU, dos Reis M, Goldstein RA (2012) Estimating the distribution of selection coefficients from phylogenetic data using sitewise mutation-selection models. Genetics 190:1101–1115
Taverna DM, Goldstein RA (2002) Why are proteins marginally stable? Proteins 46:105–109
Thiltgen G, Goldstein RA (2012) Assessing predictors of changes in protein stability upon mutation using self-consistency. PLoS ONE 7:e46084
Thomas G, Hahn M (2015) Determining the null model for detecting adaptive convergence from genomic data: a case study using echolocating mammals. Mol Biol Evol 32:1232–1236
Thorne JL (2000) Models of protein sequence evolution and their applications. Curr Opin Genet Dev 10:602–605

Weigt M, White RA, Szurmant H, Hoch JA, Hwa T (2009) Identification of direct residue contacts in protein–protein interaction by message passing. Proc Natl Acad Sci 106:67–72
Whelan S, Goldman N (2001) A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Mol Biol Evol 18:691–699
Wilbur W (1985) On the PAM matrix model of protein evolution. Mol Biol Evol 2:434–447
Williams P, Pollock D, Blackburne B (2006) Assessing the accuracy of ancestral protein reconstruction methods. PLoS Comput Biol 2:598–605
Zou Z, Zhang J (2015a) No genome-wide protein sequence convergence for echolocation. Mol Biol Evol 32:1237–1241
Zou Z, Zhang J (2015b) Are convergent and parallel amino acid substitutions in protein evolution more prevalent than neutral expectations? Mol Biol Evol 32:2085–2096
Zou Z, Zhang J (2017) Gene tree discordance does not explain away the temporal decline of convergence in mammalian protein sequence evolution. Mol Biol Evol

Genome-Wide Screens for Molecular Convergent Evolution in Mammals

Jun-Hoe Lee and Michael Hiller

Abstract Convergent evolution can occur at both the phenotypic and molecular level. Of particular interest are cases of convergent molecular changes that underlie convergent phenotypic changes, as they highlight the genomic differences that underlie phenotypic adaptations and can inform us on why evolution has repeatedly chosen the same solution in lineages that have evolved independently. Many approaches to identify convergent molecular evolution have focused on candidate genes with known functions as well as lineages with known convergent phenotypes. The growing amount of genomic sequence data makes it now possible to systematically detect molecular convergence genome-wide. Here, we highlight the advantages and drawbacks of using genomic screens to identify molecular convergence. We present our method to detect convergent substitutions between any pair of lineages in a genome-wide manner, ways of enriching for convergence that are more likely to affect protein function, and present novel cases of convergence in echolocating mammals. Our results suggest that genomic screens have the potential to generate new hypotheses of associations between molecular convergence and phenotypic convergence. Together with experimental assays to test for functional convergence, this will contribute to revealing the genomic changes that underlie convergent phenotypic changes.

1 Convergent Molecular Evolution

Convergent evolution, in the most basic sense, refers to the acquisition of similar traits in independent lineages. Some well-known examples include the wings that birds and bats use for powered flight, the highly streamlined body form of dolphins and fish that allows for efficient movement in an aquatic environment, or adapta-

J.-H. Lee · M. Hiller
Max Planck Institute of Molecular Cell Biology and Genetics, Dresden, Germany

J.-H. Lee · M. Hiller (✉)
Max Planck Institute for the Physics of Complex Systems, Dresden, Germany
e-mail: hiller@mpi-cbg.de

P. Pontarotti (ed.), *Evolutionary Biology: Self/Nonself Evolution, Species and Complex Traits Evolution, Methods and Concepts*,
DOI 10.1007/978-3-319-61569-1_16

tions in ant-foraging mammals across different continents that have evolved powerful digging forelimbs and a long, sticky tongue. These examples of convergent phenotypic evolution document the power of natural selection to repeatedly result in highly similar adaptations that are extremely unlikely to arise by neutral evolution.

From a molecular perspective, convergence can also be observed at different levels of hierarchy, such as pathways, structures, and genes. For instance, many different plants including maize and sugarcane have convergently evolved the C_4 photosynthesis pathway (Williams et al. 2013). Similarly, many distantly related fish and insects have independently evolved antifreeze proteins that share similar structural attributes (Chen et al. 1997; Davies et al. 2002). Convergence can also occur in the same gene through different mutations that confer a similar functional change. This is illustrated by the higher oxygen affinity of hemoglobin in independent bird species that have adapted to high-altitude environments. A recent study compared 56 pairs of high- and low-altitude birds and found that amino acid substitutions at multiple sites can increase oxygen affinity, suggesting that there can be multiple solutions for the same problem (Natarajan et al. 2016). However, this study also revealed several independent high-altitude lineages, where higher oxygen affinity can be traced to identical amino acid substitutions. In the following, we focus on these particular cases of molecular convergence. Such cases where the same nucleotide or amino acid substitution occurs in independent lineages have sometimes been divided into parallel and convergent substitutions, depending on whether the inferred ancestral residues in both lineages are the same (parallel) or different (convergent) (Zhang and Kumar 1997). For simplicity, we refer here to both cases as convergent substitutions.

2 Convergent Molecular Evolution Can Underlie Convergent Phenotypic Evolution

A well-studied example of convergent molecular evolution that contributes to phenotypic convergence is prestin, a motor protein critical for high-frequency hearing in mammals (Dallos and Fakler 2002). High-frequency hearing is essential for echolocating bats and toothed whales as it allows them to detect small, moving prey in conditions with poor visibility. Given the convergence in high-frequency hearing between independent echolocating mammalian lineages, *Slc26a5* encoding prestin was a promising candidate gene to examine whether molecular convergence has occurred. Phylogenetic analysis of the prestin protein sequence clustered the echolocating lineages (dolphin and two independent bat lineages) together (Li et al. 2008, 2010). Thus, the topology of the phylogenetic tree inferred from the prestin protein differs from the generally accepted mammalian phylogeny and suggested molecular convergence between these lineages. Furthermore, a tree that clustered the echolocating mammals was also obtained using the first and second codon

positions that mainly determine the encoded protein sequence. In contrast, a tree computed from only third codon positions that are mostly synonymous has a topology consistent with the mammalian phylogeny (Li et al. 2010). A closer examination of the prestin sequence alignment identified several parallel substitutions in the echolocating mammals (Fig. 1). Subsequent in vitro experiments demonstrated that some of these parallel substitutions (N7T and I384T) alter the voltage-dependent properties of prestin, consistent with a functional change that contributes to hearing higher frequencies (Liu et al. 2014). Thus, the function of prestin in echolocating mammals exemplifies that convergent molecular evolution can be involved in convergent phenotypic changes.

Apart from high-frequency hearing, several other convergent phenotypes have been linked to convergent amino acid substitutions. For example, spectral tuning has been reported in the opsin of different cichlid populations, which enables these species to detect light from different wavelengths at different water depths (Nagai et al. 2011). Adaptations to a herbivorous diet in ruminants and leaf-eating monkeys have been linked to convergent mutations in lysozyme and RNAse1 (Stewart et al. 1987; Zhang 2006). Toxin resistance against cardenolide in insects, reptiles, amphibians, and mammals is associated with various convergent mutations in a sodium–potassium pump (Zhen et al. 2012; Dobler et al. 2012; Ujvari et al. 2015), while resistance against tetrodotoxin in various snakes is mediated by convergent mutations in a voltage-gated sodium channel (Geffeney et al. 2005; Feldman et al. 2012). Recently, it was suggested that convergent evolution in limb development genes is involved in the development of the pseudothumb in the giant and the red panda, two independent bamboo-eating lineages (Hu et al. 2017).

The various examples of convergent molecular evolution involving identical mutations raise the question of why evolution has repeatedly chosen the same

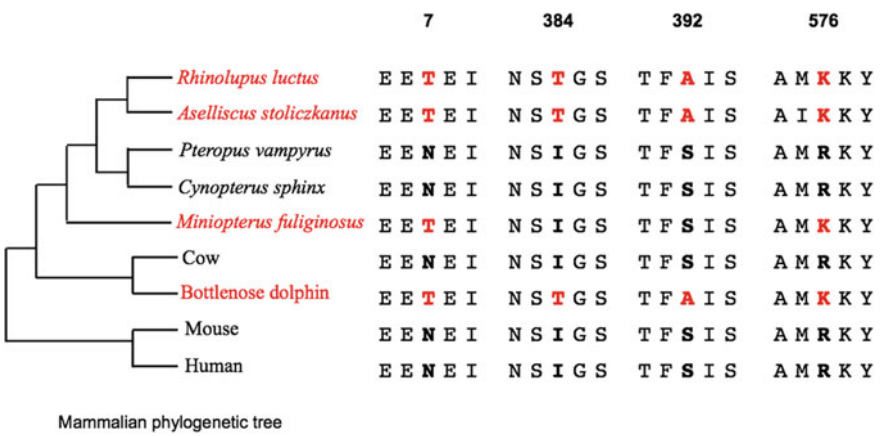


Fig. 1 Simplified mammalian phylogeny and selected regions of the prestin sequence alignment. The echolocating mammals as well as several of the sites showing convergence (N7T, I384T, S392A, and R576 K) are highlighted in *red* (Li et al. 2010). The N7T and I384T sites have been experimentally tested to affect prestin function (Liu et al. 2014)

solution. One possible explanation is that there could be very few amino acid changes that shift the functional activity of a protein in a specific direction (Christin et al. 2010). Consequently, the number of solutions can be very limited. For example, it appears that there are only three critical substitutions (S180A, Y277F, and T285A) in the primate M/LWS opsin that lead to specific color vision changes by causing additive (and reversible) spectral shifts of the wavelength of maximal absorbance (λ_{max}) by −7, −8, −15 nm, respectively (reviewed in Kawamura et al. 2012). In addition, the order of mutations can further constrain the evolutionary trajectory, as shown for the β-lactamase protein (Weinreich 2006). There are five amino acid mutations in β-lactamase leading to increased bacterial resistance. These five mutations could be derived from 120 possible evolutionary trajectories. However, many of these trajectories do not follow a continuous increase in bacterial resistance and are therefore less likely to be favored in evolution. This shows that the effect of mutations is not independent of other mutations. Such nonadditive effects of mutations on protein structure and function are called epistasis (Starr and Thornton 2016). Thus, even if several solutions exist, epistasis within a gene can restrict the possible evolutionary trajectories.

Convergent sequence evolution is particularly interesting as it has the potential to highlight the genomic changes that underlie phenotypic adaptations. Proteins with convergent changes represent candidates that can be experimentally tested for convergence in protein function. Furthermore, since one often lacks a good understanding how changes in sequence affect protein structure and function, it is difficult to narrow down the list of individual amino acid changes to be experimentally tested for their effect on function. The identification of convergent substitutions can provide a starting point for experiments to determine the specific sequence changes that contribute toward convergence in protein function.

3 Do the Convergent Substitutions Always Have the Same Effect?

Although functional experiments on selected convergent sites in prestin demonstrated a similar functional change in echolocating bats and toothed whales, these results cannot be generalized (Liu et al. 2014). In the case of RNAse1, the specific substitutions that modify the catalytic activity of the enzyme in ruminants can have a different effect if these substitutions are introduced into the orthologous RNAse1 of leaf-eating langurs (Zhang 2003). Specifically, even though the Q28L substitution increases the enzymatic activity in cows, it led to decreased enzymatic activity in the langur protein. This suggested that the same substitutions can have different and possibly even opposing effects in homologous proteins because of the different evolutionary history of individual proteins. Due to epistasis, a recent substitution could modify the rate of mutations at other sites to preferentially accommodate substitutions that yield a new stable conformation, a phenomenon that has been termed as Stokes shift (Pollock et al. 2012). Thus, the longer the divergence time

between two lineages, the less likely it is that the same mutation in a homologous protein has the same effect, which decreases the rate of convergence (Goldstein et al. 2015; Zou and Zhang 2015a). Notably, closely-related lineages can exhibit a high level of non-adaptive convergence, i.e., sequence similarity by chance (also known as background convergence), which decreases as divergence time increases. Therefore, caution must be taken when trying to infer adaptive convergence from identical substitutions observed in independent lineages.

To demonstrate that convergent substitutions in a protein are adaptive, Zhang (2006) has proposed 4 criteria: i. The convergent substitutions are observed in lineages that have evolved independently, ii. the convergent substitutions were driven by a common selective pressure, iii. there is convergence in protein function, and iv. the change in protein function can be clearly linked to the convergent substitution. With few exceptions, such as prestin in echolocating mammals, most of the studies that reported on associations between molecular convergence and potentially adaptive convergent phenotypes do not fulfill all criteria, in particular criteria iii and iv that require functional experiments. The fourth criterion is more difficult to satisfy as it requires introducing amino acid mutations in the orthologous protein in other species, followed by functional assays. Even if the convergent substitutions are important for convergence in protein function in the particular lineages, introducing the same mutations in the orthologous protein of another species that has a different evolutionary background might not result in a similar functional change, due to epistasis and/or Stokes shift. Alternatively, the convergent mutations could be introduced in the reconstructed ancestral version of the protein, using the ancestor that predates the adaptive phenotype. However, if several mutations have occurred on the branch descending from this ancestor, the convergent mutation might not be the first mutation that had occurred and experimental results might again depend on the background of other mutations. Thus, demonstrating that the change in protein function is caused directly by the convergent substitutions can be difficult if the effect of these mutations is influenced by the evolutionary background of the protein.

4 Genome-Wide Screens for Convergent Molecular Evolution

The identification of molecular convergence in prestin led to a search for molecular convergence between echolocating mammals in additional candidate genes with known hearing-related functions, based on knockout studies in mice or associations with deafness and hearing disorders in humans. These candidate gene approaches detected molecular convergence in several hearing-related proteins in echolocating mammals: Kcnq4, Tmc1, Pjvk, Cdh23, Pcdh15, and Otof (Liu et al. 2011; Davies et al. 2012; Shen et al. 2012). However, functional convergence of these six proteins or the effect of the convergent substitutions (criteria iii and iv according to Zhang (2006)) has not been experimentally explored. Consequently, it is currently

unknown whether there is convergence in protein function and whether the convergent mutations are involved.

The main limitation of the candidate gene approach to discover molecular convergence is the requirement for functionally well-characterized genes that could be associated with a well-characterized convergent phenotype in selected species. Advancements in sequencing technologies have led to the sequencing of hundreds of genomes, and the number continues to grow rapidly. This wealth of genomic data has made it possible to carry out genomic screens to detect genes with molecular convergence in the selected species. For example, Parker et al. (2013) sequenced the genomes of four bats and screened 2326 orthologous proteins in 22 mammals for signatures of molecular convergence in the echolocating bats and dolphin. For each gene, Parker et al. (2013) calculated the difference in the site-specific likelihood between a null tree (the accepted mammalian phylogeny) and hypothetical trees in which the echolocating mammals are artificially clustered together. A higher value for the hypothetical trees was taken as support for convergence, based on comparisons with a simulated null distribution of the site under the same parameters. In this way, Parker et al. (2013) suggested that 117 proteins exhibit signatures of convergence, particularly those linked to hearing and vision. Besides echolocating mammals, a genomic screen in three marine mammalian lineages (cetaceans, pinnipeds, sirenia) reported positive selection and molecular convergence in a small subset of proteins that could be linked to marine adaptations (Foote et al. 2015).

However, the use of site-specific support values and simulations by Parker et al. (2013) was subsequently criticized by several studies, mostly because many of the reported proteins do not exhibit convergent amino acid substitutions between echolocators. Out of the 117 candidate genes identified by Parker et al. (2013), only 19 were found to exhibit convergent substitutions between the echolocating lineages (Thomas and Hahn 2015). Although convergent mutations can result in a higher likelihood for the hypothetical tree, there are other factors unrelated to convergence that affect the likelihood of the null tree and hypothetical tree. As succinctly stated by Zou and Zhang (2015b), "convergence does not necessarily result in a wrong phylogeny and a wrong phylogeny is not necessarily caused by convergence." Furthermore, a re-examination of the 22 novel candidate genes related to hearing (Parker et al. 2013) found that 45% (10 of 22) did not pass statistical tests to demonstrate a higher than expected amount of convergence (Zou and Zhang 2015b). Using a set of 6400 orthologous proteins from 9 mammals, Thomas and Hahn (2015) also found 1951 genes that show convergence between microbat and cow, which is higher than the 1372 genes that show convergence between microbat and dolphin. In addition, the majority of the candidate genes exhibit convergent substitutions in other non-echolocating lineages; thus, the observed convergence cannot be wholly attributed to the adaptive phenotype (echolocation). Similarly, Foote et al. (2015) noted in their study of marine mammals that the terrestrial sister taxa of the marine mammals, which were used as control, exhibited a higher amount of convergent substitutions. Nonetheless, both follow-up studies agreed that the existence of background convergence in many

species does not exclude the possibility of adaptive molecular convergence in a small number of proteins in specific lineages, such as prestin and possibly other hearing-related genes in echolocating mammals (Thomas and Hahn 2015; Zou and Zhang 2015b).

The main challenge, as discussed by the studies above, is the lack of an accurate probabilistic model of sequence evolution that applies to all sites in a protein. Most sequence evolution models tend to underestimate or do not take into account background convergence (Castoe et al. 2009). This is problematic as several studies have reported that background convergence is more prevalent than previously expected (Stayton 2008; Rokas and Carroll 2008; Thomas and Hahn 2015; Zou and Zhang 2015a). Several solutions have been proposed to address these issues, such as tools that employ different statistical methods to determine whether convergent substitutions in a pair of lineages are higher than expected under certain models or tests that compare the number of divergent to the number of convergent amino acid changes between all lineage pairs (Castoe et al. 2009; Qian et al. 2015). Nevertheless, determining whether adaptive molecular convergence has occurred remains an active area of research.

5 Development of a Pipeline to Search for Convergent Molecular Evolution in an Unbiased Fashion

Despite the lack of a proper statistical framework to assess the significance of convergent mutations, the rapidly increasing number of sequenced genomes provides an unprecedented opportunity to carry out genomic screens on a multi-species scale to develop new hypotheses about associations between molecular convergence and phenotypic convergence. This approach is analogous to the use of genomic screens for positively selected proteins or proteins with lineage-specific amino acid mutations, which uncovered compelling examples that could be related to phenotypic changes. For example, proteins related to DNA replication and repair such as APEX1 and RFC1 exhibit amino acid changes that are unique to the naked mole rat and did not occur in other mammals (Kim et al. 2012). These naked mole rat-specific changes might be linked to the extraordinarily long life span and cancer resistance in that species. Another study identified a number of aging and cancer-related genes that were under positive selection such as SOCS2, APTX, NOG, and LEP in the long-lived bowhead whale (Keane et al. 2015).

There are two advantages in using genomic screens. First, they have the potential to uncover sets of functionally related genes that exhibit a higher number of convergent substitutions in particular lineages. The effect of molecular convergence on phenotypic convergence might be easier to interpret within such a gene set, even though some of those genes could exhibit convergence due to random chance. Second, the genomes of many species provide an opportunity to not only search for convergence in species known to exhibit convergent phenotypes, but to detect molecular convergence between any pair of independent lineages. This is

advantageous as it allows for a thorough test on whether a set of functionally-related genes (such as hearing-related genes) also exhibits a similar number of convergent mutations in other lineages (such as non-echolocating mammals).

To this end, we have developed a pipeline to look for genes exhibiting convergent evolution in a systematic, genome-wide manner (Lee et al. 2017). In our approach, the first step involved obtaining one-to-one orthologous protein sequences of 31 mammalian species from Ensembl (Hubbard et al. 2009; Kinsella et al. 2011), followed by filtering of ambiguous and poor-quality sequences. Next, we carried out a multiple sequence alignment for all ortholog sets, followed by ancestral reconstruction using a maximum likelihood approach. Subsequently, we iterated over every position in the alignment and systematically detected all convergent amino acid substitutions in all independent pairs of lineages (Fig. 2). Applying this pipeline to 14,407 sets of orthologous proteins across 31 mammals required approximately 180 CPU hours, though this can be greatly reduced through parallelization on a computer cluster.

We found 13,330 proteins that have at least one convergent substitution in one or more independent pairs of lineages. In total, we found over a million entries that consist of a protein with one or more convergent substitutions between a lineage pair. This finding is consistent with studies that reported on widespread background convergence in various genomes (Rokas and Carroll 2008; Thomas and Hahn 2015; Zou and Zhang 2015a). We further observed that most of the convergent substitutions are either conservative (substitution to another amino acid with similar physicochemical properties) or occur at alignment positions that are poorly conserved (Fig. 3). Both factors indicate that the majority of convergent amino acid substitutions are less likely to affect protein function. Indeed, the convergent changes in prestin that affect protein function (N7T and I384T) occur at highly conserved positions and are radical substitutions i.e., a replacement of an amino acid with another residue that has different physicochemical properties (Li et al. 2010). To enrich for molecular convergence that is more likely to affect function, we specifically filtered for convergent substitutions that are both radical and occur

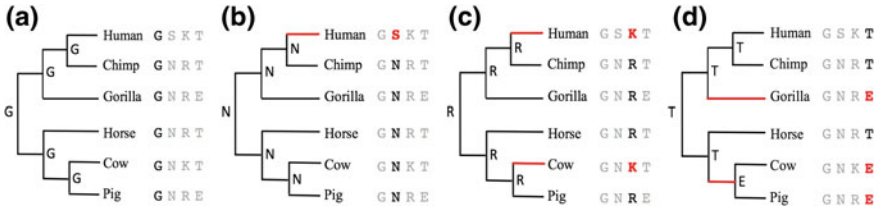


Fig. 2 Illustration of the systematic search for convergent substitutions in all pairs of lineages. First, ancestral reconstruction was carried out for every internal node in the species phylogeny. Next, we screen for convergent substitutions in independent lineages in each column of the alignment. **a** There are no substitutions observed in any lineages. **b** A substitution of N → S was observed only in human. **c** A convergent change of R → K is observed in human and cow, and this is recorded as an entry. **d** Another convergent change of T → E is observed in gorilla and the ancestral lineage of cow-pig, which is also recorded as an entry

at a highly conserved position, which greatly reduced the candidate list to 145,465 lineage pairs with one or more convergent substitutions in a protein (Lee et al. 2017).

Complex phenotypic changes likely require multiple changes in functionally related genes, exemplified by several hearing-related genes that exhibit convergence in echolocating mammals. Therefore, we proceeded to perform functional enrichment tests of all proteins that exhibit convergence between a pair of lineages. As shown in Table 1, for the echolocating microbat and dolphin, we found enrichments for hearing-related terms such as "abnormal ear morphology" that refer to genes that affect ear morphology in a mouse knockout. These genes include those that have been reported previously (e.g., prestin and Pjvk) (Li et al. 2010; Davies et al. 2012). In addition, we found several genes that have not been reported previously. For example, we detected the Bardet–Biedl syndrome 2 protein (Bbs2) and tyrosine-related protein 1 (Tyrp1), two proteins that are expressed in the cochlea (Fig. 4). Bbs2 is implicated in Bardet–Biedl syndrome, an autosomal recessive disorder that includes speech impairment and sensorineural hearing loss (May-Simera et al. 2009). Tyrp1 is a melanosomal enzyme that has been implicated in the decline of the endocochlear potential, which is one of the factors that contributes toward age-related hearing loss (Ohlemiller 2009).

The design of our genome-wide screen in detecting convergence between any pair of lineages makes it possible to test whether the observed enrichment in hearing-related genes is higher for microbat and dolphin compared to all other pairs of lineages. To this end, we iterated over all pairs of independent lineages, defined as two branches that do not share a direct common ancestor and where one branch is not a descendant of the other. For each independent pair of branches, we counted the number of convergent amino acid changes in all proteins that are associated

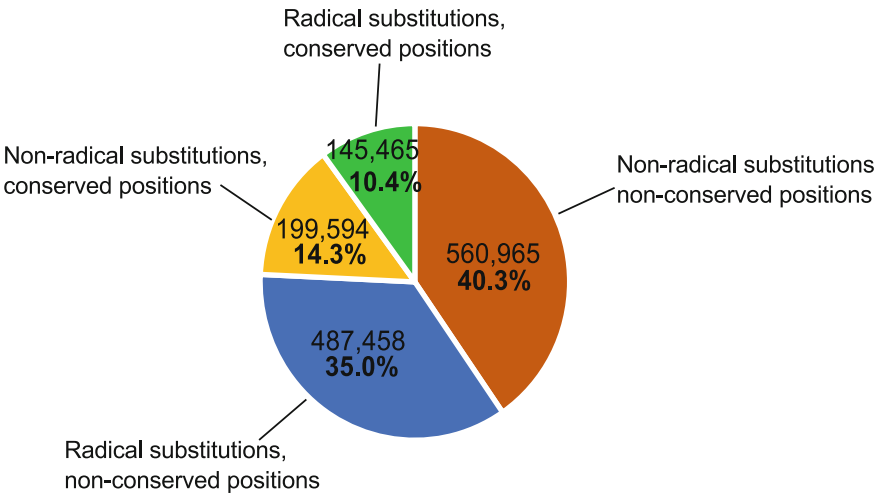


Fig. 3 Classification of all detected convergent amino acid substitutions

Table 1 Functional enrichments of the set of genes with radical convergent amino acid substitutions at conserved positions using Enrichr (Kuleshov et al. 2016) for mammalian phenotypes and GeneTrail2 (Stöckel et al. 2016) for gene ontology, ranked by P-values adjusted for multiple testing. The terms highlighted in *red* represent terms that are related to the physiology/morphology of ear and cellular components of contractile muscle fibers. The "Hits" column indicates the number of genes annotated with the phenotype or gene ontology term that has convergent amino acid substitutions

MGI Mammalian Phenotype Level 3	Hits	Adjusted P-value
MP0001919 abnormal reproductive system physiology	47	1.86E-003
MP0010769 abnormal survival	95	1.86E-003
MP0001672 abnormal embryo development	46	3.00E-003
MP0002102 abnormal ear morphology	19	4.09E-003
MP0001968 abnormal touch/nociception	12	1.76E-002
MP0003632 abnormal nervous system morphology	57	1.76E-002
MP0003633 abnormal nervous system physiology	44	1.83E-002
MP0004924 abnormal behavior	67	2.44E-002
MP0004196 abnormal prenatal growth/weight/body size	28	2.59E-002
MP0009389 abnormal extracutaneous pigmentation	8	3.61E-002
MP0003878 abnormal ear physiology	16	4.98E-002
MP0005621 abnormal cell physiology	36	4.98E-002
MP0002163 abnormal gland morphology	35	5.22E-002
MP0000358 abnormal cell morphology	12	5.29E-002
MP0002106 abnormal muscle physiology	23	5.29E-002

GO – Cellular Component	Hits	Adjusted P-value
contractile fiber part	15	1.29E-007
contractile fiber	15	1.40E-007
sarcomere	14	1.40E-007
ciliary part	17	1.69E-007
myofibril	14	3.83E-007
I band	10	2.72E-005
organelle subcompartment	13	9.84E-005
clathrin coated vesicle	9	9.85E-004
primary cilium	10	9.85E-004
sarcolemma	8	9.85E-004
Golgi subcompartment	11	1.54E-003
nuclear membrane	11	1.54E-003
apical plasma membrane	11	1.55E-003
axoneme part	5	1.55E-003
axoneme & ciliary cytoplasm	7	1.55E-003

with abnormal ear morphology. This shows that out of all 1984 pairs of lineages, microbat and dolphin rank 14th with a total of 22 observed convergent changes (Fig. 5). In other words, 99.3% of all other lineage pairs exhibit fewer convergent changes, suggesting that the convergence in at least some of these proteins could play a role in the evolution of echolocation.

Fig. 4 Alignment of Bbs2 and Tyrp1, two hearing-related proteins identified to show convergent substitutions in the echolocating microbat and dolphin

6 The Evolution of Superfast Muscles in Echolocating Mammals

Apart from ear-related enrichments, the proteins with convergence between dolphin and microbat also show statistical enrichments for genes related to muscle function (Table 1). Interestingly, some of these genes are specifically related to the physiology of fast-twitch muscles. As explained below in detail and in Lee et al. (2017), this suggests that there could be a connection between molecular convergence in fast muscle proteins and the less studied convergent aspect of echolocation, which is vocalization.

The ability to echolocate is a complex phenotype that requires the ability to produce and detect high frequency calls and subsequently convert this information into an acoustic representation of their environment. Most studies on convergence have focused on the high-frequency hearing aspect of echolocation, while little attention has been directed to vocalization. This is partially due to the fact that there is no obvious convergence in vocalization between echolocating bats and toothed whales as bats produce sounds in their larynx and toothed whales in their nasal complex (Clement et al. 2006; Berta et al. 2014). Despite the different anatomical

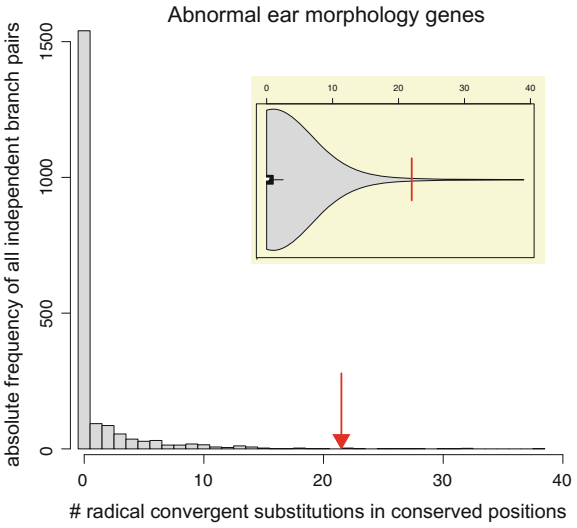


Fig. 5 Histogram of the number of convergent substitutions for each independent pair of lineages observed in 222 proteins that affect ear morphology in a mouse knockout (MGI phenotype identifier MP0002102). Only radical convergent amino acid substitutions at conserved positions are considered. The *red arrow* shows the microbat–dolphin pair. The *inset* shows the same data as a violin plot

structures used for sound production, both lineages have converged in the way clicks and calls are produced in the final moments of capturing prey. During foraging, there are two phases of call production in bats which are termed as the search phase (flying without a specific direction) and approach phase (flying toward a target). Once prey is detected, the bat will transition into the approach phase, which is accompanied by drastically increasing the repetition rate of its calls to precisely track its prey (Fenton et al. 2012). The final moments of capturing prey end with a terminal buzz, which is a period of extremely high repetition call rate up to 200 calls per second (Moss et al. 2011). The terminal buzz provides the bat with near-instantaneous feedback as it homes in on the prey's position, in case of any sudden escape maneuvers. Strikingly, the use of terminal buzz has also been observed in toothed whales, indicating that both lineages converged not only in the hearing-related aspect of echolocation but also in a similar strategy of producing calls to maximize the prey capture success rate (Johnson et al. 2006).

The extremely rapid call rates in bats during the terminal buzz are powered by equally rapid superfast muscles found in the larynx (Elemans et al. 2011). Isolated fibers from these highly unique muscles have been experimentally demonstrated to produce power cycles up to 180 Hz, which is close to the call rates produced during the terminal buzz. Although the bat laryngeal muscle is the first superfast muscle that was experimentally investigated in mammals, other experiments have demonstrated the existence of superfast muscles in other vertebrates including the swim bladder muscles that produce the "boatwhistle" mating call in male toadfish (Rome et al. 1996), vocal muscles in songbirds (Elemans et al. 2004), shaker muscles in rattlesnakes (Schaeffer et al. 1996), and most recently, wing muscles in manakin (Fuxjager et al. 2016).

Based on our finding that several fast muscle proteins show convergent substitutions in echolocating mammals, we hypothesized that these proteins could be involved in the physiological adaptations toward building superfast muscles that power the incredibly rapid calls during the terminal buzz. Further detailed work including functional experiments will be necessary to investigate this hypothesis.

7 Summary

The availability of many sequenced genomes has made it possible to extend candidate gene approaches and systematically screen for molecular convergence. This in turn intensifies the problem of assessing whether the observed convergence is higher than expected by chance. Despite several proposed approaches, this problem is an active area of research without a general consensus on the best solution. Nonetheless, genome-wide screens have the potential to detect novel associations between phenotypic change and genomic differences (Hiller et al. 2012; Prudent et al. 2016). As we have shown here, genome-wide screens for convergence in particular lineages can detect not only individual genes, but sets of functionally related genes or pathways by using statistical enrichment tests. Furthermore, unbiased genome-wide screens that simultaneously record convergence between any pair of lineages will allow subsequent testing on whether the convergence observed in the gene set is exceptional compared to all other pairs of lineages. Although this is not a formal proof of adaptive convergence, it can still provide new hints toward potentially unknown phenotypic convergence in these species and motivate further investigation. Ultimately, experiments are necessary to demonstrate convergence in the function of the discovered proteins and where possible, to test the role of the convergent mutations. Similar to genomic screens that searched for other molecular patterns (positive selection or lineage-specific mutations in genes, gene family expansions or gene losses), genomic screens for molecular convergence have the potential to generate new hypotheses of associations between molecular convergence and phenotypic convergence, which will help to illuminate the genomic changes that underlie nature's fascinating phenotypic diversity.

References

Berta A, Ekdale EG, Cranford TW (2014) Review of the cetacean nose: form, function, and evolution. Anat Rec 297:2205–2215

Castoe TA, de Koning APJ, Kim H-M et al (2009) Evidence for an ancient adaptive episode of convergent molecular evolution. Proc Natl Acad Sci USA 106:8986–8991

Chen L, DeVries AL, Cheng CH (1997) Convergent evolution of antifreeze glycoproteins in Antarctic notothenioid fish and Arctic cod. Proc Natl Acad Sci USA 94:3817–3822

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

Clement M, Dietz N, Gupta P (2006) Audiovocal communication and social behavior in mustached bats. In: Kanwal J, Ehret G (eds) Behavior and neurodynamics for auditory communication. Cambridge University Press, Cambridge, pp 57–84

Dallos P, Fakler B (2002) Prestin, a new type of motor protein. Nat Rev Mol Cell Biol 3:104–111

Davies KTJ, Cotton JA, Kirwan JD et al (2012) Parallel signatures of sequence evolution among hearing genes in echolocating mammals: an emerging model of genetic convergence. Heredity 108:480–489

Davies PL, Baardsnes J, Kuiper MJ, Walker VK (2002) Structure and function of antifreeze proteins. Philos Trans R Soc Lond B Biol Sci 357:927–935

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

Elemans CPH, Mead AF, Jakobsen L, Ratcliffe JM (2011) Superfast muscles set maximum call rate in echolocating bats. Science 333:1885–1888

Elemans CPH, Spierts ILY, Müller UK et al (2004) Bird song: superfast muscles control dove's trill. Nature 431:146

Feldman CR, Brodie ED, Pfrender ME (2012) Constraint shapes convergence in tetrodotoxin-resistant sodium channels of snakes. Proc Natl Acad Sci USA 109:4556–4561

Fenton M, Faure P, Ratcliffe J (2012) Evolution of high duty cycle echolocation in bats. J Exp Biol 215:2935–2944

Foote AD, Liu Y, Thomas GWC et al (2015) Convergent evolution of the genomes of marine mammals. Nat Genet 47:272–275

Fuxjager MJ, Goller F, Dirkse A et al (2016) Select forelimb muscles have evolved superfast contractile speed to support acrobatic social displays. Elife 5:3523–3528

Geffeney SL, Fujimoto E, Brodie ED III et al (2005) Evolutionary diversification of TTX-resistant sodium channels in a predator-prey interaction. Nature 434:759–763

Goldstein RA, Pollard ST, Shah SD, Pollock DD (2015) Nonadaptive amino acid convergence rates decrease over time. Mol Biol Evol 32:1373–1381

Hiller M, Schaar BT, Indjeian VB, et al (2012) A "Forward Genomics" approach links genotype to phenotype using independent phenotypic losses among related species. Cell Rep 1–7

Hu Y, Wu Q, Ma S et al (2017) Comparative genomics reveals convergent evolution between the bamboo-eating giant and red pandas. Proc Natl Acad Sci USA 114:201613870

Hubbard TJP, Aken BL, Ayling S et al (2009) Ensembl 2009. Nucleic Acids Res 37:D690–D697

Johnson M, Madsen PT, Zimmer WMX et al (2006) Foraging Blainville's beaked whales (*Mesoplodon densirostris*) produce distinct click types matched to different phases of echolocation. J Exp Biol 209:5038–5050

Kawamura S, Hiramatsu C, Melin AD, et al (2012) Polymorphic color vision in primates: evolutionary considerations. In: Post-genome biology of primates. Springer, Tokyo, pp 269–280

Keane M, Semeiks J, Webb AE et al (2015) Insights into the evolution of longevity from the bowhead whale genome. Cell Rep 10:112–122

Kim EB, Fang X, Fushan AA et al (2012) Genome sequencing reveals insights into physiology and longevity of the naked mole rat. Nature 479:223–227

Kinsella RJ, Kähäri A, Haider S, et al (2011) Ensembl BioMarts: A hub for data retrieval across taxonomic space. Database 2011:bar030

Kuleshov MV, Jones MR, Rouillard AD et al (2016) Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic Acids Res 44:W90–W97

Lee J-H, Lewis KM, Moural TW, et al (2017) Building superfast muscles: insights from molecular parallelism in fast-twitch muscle proteins in echolocating mammals. Submitted

Li G, Wang J, Rossiter SJ et al (2008) The hearing gene *Prestin* reunites echolocating bats. Proc Natl Acad Sci USA 105:13959–13964

Li Y, Liu Z, Shi P, Zhang J (2010) The hearing gene *Prestin* unites echolocating bats and whales. Curr Biol 20:R55–R56

Liu Z, Li S, Wang W et al (2011) Parallel evolution of *KCNQ4* in echolocating bats. PLoS ONE 6: e26618
Liu Z, Qi F-Y, Zhou X et al (2014) Parallel sites implicate functional convergence of the hearing gene *prestin* among echolocating mammals. Mol Biol Evol 31:2415–2424
May-Simera HL, Ross A, Rix S et al (2009) Patterns of expression of Bardet-Biedl syndrome proteins in the mammalian cochlea suggest noncentrosomal functions. J Comp Neurol 514:174–188
Moss CF, Chiu C, Surlykke A (2011) Adaptive vocal behavior drives perception by echolocation in bats. Curr Opin Neurobiol 21:645–652
Nagai H, Terai Y, Sugawara T et al (2011) Reverse evolution in *RH1* for adaptation of cichlids to water depth in Lake Tanganyika. Mol Biol Evol 28:1769–1776
Natarajan C, Hoffmann FG, Weber RE et al (2016) Predictable convergence in hemoglobin function has unpredictable molecular underpinnings. Science 354:336–339
Ohlemiller KK (2009) Mechanisms and genes in human strial presbycusis from animal models. Brain Res 1277:70–83
Parker J, Tsagkogeorga G, Cotton JA et al (2013) Genome-wide signatures of convergent evolution in echolocating mammals. Nature 502:1–9
Pollock DD, Thiltgen G, Goldstein RA (2012) Amino acid coevolution induces an evolutionary Stokes shift. Proc Natl Acad Sci USA 109:E1352–E1359
Prudent X, Parra G, Schwede P et al (2016) Controlling for hylogenetic relatedness and evolutionary rates improves the discovery of associations between species' phenotypic and genomic differences. Mol Biol Evol 33:2135–2150
Qian C, Bryans N, Kruykov I, de Koning APJ (2015) Visualization and analysis of statistical signatures of convergent molecular evolution. University of Calgary. http://lab.jasondk.io
Rokas A, Carroll SB (2008) Frequent and widespread parallel evolution of protein sequences. Mol Biol Evol 25:1943–1953
Rome LC, Syme DA, Hollingworth S et al (1996) The whistle and the rattle: the design of sound producing muscles. Proc Natl Acad Sci USA 93:8095–8100
Schaeffer PJ, Conley KE, Lindstedt STL (1996) Structural correlates of speed and endurance in skeletal muscle: the rattlesnake tailshaker muscle. J Exp Biol 199:351–358
Shen YY, Liang L, Li GS et al (2012) Parallel evolution of auditory genes for echolocation in bats and toothed whales. PLoS Genet 8:e1002788
Starr TN, Thornton JW (2016) Epistasis in protein evolution. Protein Sci 25:1204–1218
Stayton CT (2008) Is convergence surprising? An examination of the frequency of convergence in simulated datasets. J Theor Biol 252:1–14
Stewart CB, Schilling JW, Wilson AC (1987) Adaptive evolution in the stomach lysozymes of foregut fermenters. Nature 330:401–404
Stöckel D, Kehl T, Trampert P et al (2016) Multi-omics enrichment analysis using the GeneTrail2 web service. Bioinformatics 32:1502–1508
Thomas GWC, Hahn MW (2015) Determining the null model for detecting adaptive convergence from genomic data: A case study using echolocating mammals. Mol Biol Evol 32:1232–1236
Ujvari B, Casewell NR, Sunagar K et al (2015) Widespread convergence in toxin resistance by predictable molecular evolution. Proc Natl Acad Sci USA 112:201511706
Weinreich DM (2006) Darwinian evolution can follow only very few mutational paths to fitter proteins. Science 312:111–114
Williams BP, Johnston IG, Covshoff S, Hibberd JM (2013) Phenotypic landscape inference reveals multiple evolutionary paths to C4 photosynthesis. Elife 2:e00961
Zhang J (2006) Parallel adaptive origins of digestive RNases in Asian and African leaf monkeys. Nat Genet 38:819–823
Zhang J (2003) Parallel functional changes in the digestive RNases of ruminants and colobines by divergent amino acid substitutions. Mol Biol Evol 20:1310–1317
Zhang J, Kumar S (1997) Detection of convergent and parallel evolution at the amino acid sequence level. Mol Biol Evol 14:527–536

Zhen Y, Aardema ML, Medina EM et al (2012) Parallel molecular evolution in an herbivore community. Science 337:1634–1637

Zou Z, Zhang J (2015a) Are convergent and parallel amino acid substitutions in protein evolution more prevalent than neutral expectations? Mol Biol Evol 32:2085–2096

Zou Z, Zhang J (2015b) No genome-wide protein sequence convergence for echolocation. Mol Biol Evol 32:1237–1241

Assessing Evolutionary Potential in Tree Species Through Ecology-Informed Genome Screening

Hanne De Kort and Olivier Honnay

Abstract Gaining insight into the ability of tree populations to evolve pest, disease, and drought resistance is of major importance with regard to the conservation of adaptive tree genetic resource variation under global change. However, the longevity of tree species challenges accurate assessment of additive genetic variation driving trait evolution through common garden experiments and pedigree-informed quantitative genetic analyses. Here, we argue that recent developments in landscape and population genomics pave the way for molecular marker-based analysis of evolutionary potential as a more time and cost-efficient alternative. Focusing on phenotype- or ecology-informed molecular markers increases the per-marker contribution to the genetic variation underpinning phenotypic trait evolution. Considering that most tree species lack a reference genome, landscape genomic analysis of anonymous markers can be used in concert with in situ and ex situ phenotypic monitoring of tree populations to quantify tree adaptive potential. Global forest restoration efforts and strategies to conserve tree genetic resources are likely to benefit considerably from marker-based insights in the spatial distribution of tree *adaptive* genetic *diversity*.

1 Introduction

1.1 Increasing Global Threats to Forest Genetic Resources

Rising demands for forest resources are highly challenged by the global increase in disease and pest pressure, temperature, and drought (Pautasso et al. 2013; Trumbore

H. De Kort (✉) · O. Honnay
Plant Conservation and Population Biology, Biology Department, University of Leuven, Kasteelpark Arenberg 31, 3001 Heverlee, Belgium
e-mail: hanne.dekort@kuleuven.be

H. De Kort
Station d'Ecologie Théorique et Expérimentale (SETE), National Center for Scientific Research (CNRS), 2 Route du CNRS, 09200 Moulis, France

P. Pontarotti (ed.), *Evolutionary Biology: Self/Nonself Evolution, Species and Complex Traits Evolution, Methods and Concepts*,
DOI 10.1007/978-3-319-61569-1_17

et al. 2015). The compromised ability of tree populations from both natural and managed forest stands to keep pace with these global environmental threats is fueled by ongoing habitat destruction and fragmentation (Hansen et al. 2013; Haddad et al. 2015; Trumbore et al. 2015), which dramatically affect forest genetic resources through decreasing population sizes and connectivity between populations. On one hand, smaller populations harbor less genetic variation, which may decrease their ability to resist biotic and abiotic stress, both in the short and the long term (Willi et al. 2006; Ellegren and Galtier 2016). On the other hand, poorly connected populations are less effective in dispersing genetic material through seeds and pollen, exacerbating the genetic effects of small population sizes. Whereas fragmented tree populations have long time been thought to be less susceptible to loss of genetic variation due to their longevity and ability for long distance pollen dispersal, a recent meta-analysis, involving 97 shrub and tree species, demonstrated significant negative effects of habitat fragmentation on population genetic diversity (Vranckx et al. 2012). The global extent of habitat fragmentation, therefore, urges for a better understanding of the evolutionary processes mediating the ability of natural and managed tree stands to cope with global environmental changes.

1.2 Assessing Tree Evolutionary Potential: Molecular Markers Versus Quantitative Traits

Evolutionary potential, i.e., the ability of a population to evolve under changing environmental conditions, is mainly shaped by standing adaptive genetic variation (Hoffmann and Willi 2008). Heritability estimates, which represent the proportion of phenotypic variation available for evolution (i.e., additive genetic variation, Falconer and MacKay 1996; Hoffmann and Merilä 1999), have proven extremely informative for assessing the evolutionary potential of populations. Yet, estimates of the additive genetic component of heritability in tree populations are scarce (Alberto et al. 2013a).This lack of accurate heritability measurements in tree species can be attributed to the technical limitations associated with quantitative genetic research in tree species. Obtaining heritability estimates for trees is time and labor intensive, as this requires pedigree information, and should ideally occur under controlled environmental conditions, for example, in common gardens. Moreover, because quantitative genetic studies commonly focus on a handful of traits that are often interdependent due to genetic and environmentally driven phenotypic correlations, it remains challenging to provide realistic estimates of phenome-wide adaptive potential (Granier and Vile 2014; Blows and McGuigan 2015).

Molecular marker analyses of samples collected in natural populations offer a cheaper and faster alternative to extensive quantitative trait analysis and could be particularly useful for long-lived species and species that are difficult to manipulate in controlled conditions. Combining the advantages of a molecular marker approach with the accuracy of a quantitative trait analysis would greatly improve our ability

to assess the evolutionary potential of tree species, and consequently also benefit breeding programs and in situ forest genetic resource management and conservation under global change (de Villemereuil et al. 2016; Holliday et al. 2017). However, the relation between molecular marker variation and adaptive quantitative trait variation has been under debate for decades for several reasons (Reed and Frankham 2001; McKay and Latta 2002; Koonin and Wolf 2010). First, the complex polygenic and interactive architecture of traits prevents direct translation of genetic variation into phenotypic variation. Nonetheless, dense marker panels may capture phenome-wide variation available for evolution in populations lacking kinship information. A recent study comparing pedigree-informed heritability with single nucleotide polymorphism (SNP)-based heritability in a wild population of Soay sheep found that genomic relatedness based on ca. 18,000 anonymous SNPs accounted for up to 95% of pedigree-informed heritability of several traits related to body size, including weight and foreleg length (Bérénos et al. 2014). It is unclear, however, to what extent the use of anonymous markers as a proxy for the potential of evolutionary change can be extrapolated to trees, which are often featured by high linkage disequilibrium and thus likely require more SNPs to accurately capture quantitative genetic variation (Krutovsky and Neale 2005; Neale and Ingvarsson 2008; Sork et al. 2016, but see Slavov et al. 2012). Second, not all genetic variation is affected by ongoing or future natural selection imposed by environmental change. Most markers are assumed to be selectively neutral and therefore do not contribute to the adaptive potential of populations. Whereas the majority of population genetic studies so far has applied selectively neutral genetic markers, implicitly assuming some relation between neutral genetic variation and evolutionary potential, a crucial and largely unanswered question in evolutionary biology is how adaptive genetic markers can be applied as a proxy for the long-term adaptive potential of tree populations.

2 Genetic Marker Diversity Underlying Tree Evolutionary Potential

2.1 Quantifying Adaptive Genomic Diversity

As opposed to neutral loci, adaptive loci directly interact with the local environment to maximize fitness, for example, through driving disease resistance and synchronized phenology. Therefore, adaptive marker analysis can inform us about the potential of natural populations to adapt to changing environmental conditions. If genetic variation at adaptive markers is lower than neutral genetic diversity, neutral markers may overestimate evolutionary potential when environmental conditions, and thus fitness optima, change (Fig. 1a). On the other hand, if adaptive genetic variation exceeds neutral genetic diversity, neutral markers are indicative of limited short-term fitness, but they may underestimate evolutionary potential and long-term

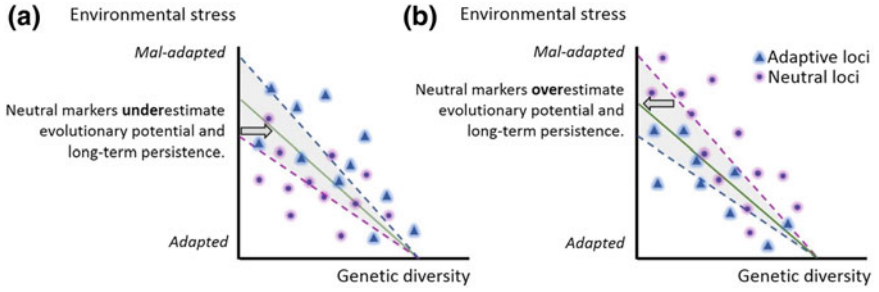


Fig. 1 Simplified illustration of how adaptive molecular variation can affect evolutionary potential of tree populations. Suboptimal environmental conditions decrease population fitness and hence genetic diversity, but the decrease in genetic diversity can differ between adaptive and neutral genetic markers. *Solid lines* represent average genetic diversity, and the direction of *gray arrows* indicates whether neutral marker genetic diversity under- or overestimates potential long-term population persistence when ignoring adaptive marker genetic diversity

population persistence (Fig. 1b). Although the overall genetic variation at adaptive markers is expected to be low due to selective pressures, skewing allele frequencies across environmental gradients (Excoffier et al. 2009), the within-population genetic diversity across adaptive markers can nevertheless exceed neutral genetic diversity (see Sect. 2.2). Assessing adaptive genetic diversity could, therefore, reveal important insights into the potential of populations to evolve new fitness optima in changing environmental conditions, without the need for long-lasting phenotypic monitoring in common gardens.

Although the genomes of most tree species are unexplored, recent developments in population and landscape genetics allow uncovering adaptive genetic variation using large sets of anonymous markers (Manel and Holderegger 2013; Sork et al. 2013; Rellstab et al. 2015). Outlier analysis can be used to identify markers with exceptional levels of genetic differentiation which are indicative of strong selective clines directly at the identified loci, or more likely, at loci linked to the loci under selection (Excoffier et al. 2009). In tree species, proportions of anonymous outliers vary between 2 and 5% for anonymous markers (e.g., Cox et al. 2011; De Kort et al. 2014; Pais et al. 2017), and between 4 and 10% for markers derived from known genes (e.g., Namroud et al. 2008; Alberto et al. 2013; Guichoux et al. 2013; Olson et al. 2013; De Kort et al. 2015), indicating that a rather small part of the genome plays a role in adaptive processes. However, many loci are expected to have small individual phenotypic effects mirrored by subtle allele frequency shifts across environmental clines and are therefore undetectable by outlier approaches. Alternatively, landscape genetic approaches aiming to find correlations between environmental drivers of selection and allele frequency patterns (Frichot et al. 2013; Manel and Holderegger 2013; de Villemereuil et al. 2014) have shown great potential in finding both weak and strong adaptive variation in genetic markers, with up to 20% of genetic markers aligning with environmental clines in woody species (De Kort et al. 2014; Christmas et al. 2016; Pluess et al. 2016; Vangestel

et al. 2016; Izuno et al. 2017). For example, Pais et al. (2017) used population and landscape genomic approaches to uncover molecular signatures of adaptation to environmental factors in dogwood tree (*Cornus florida*) populations inhabiting divergent ecological habitats in North Carolina. The authors relied on genotyping-by-sequencing (GBS) to obtain anonymous SNPs that were subsequently screened for outliers and associations with climate and soil variables as well as disease (e.g., necrosis and branch dieback). Out of ca. 2000 SNPs, 2% was identified as outliers, while 6% was significantly associated with at least one environmental variable (Pais et al. 2017). The most important environmental variables were temperature and growing period (ca. 2.5%) followed by soil properties (1.5%) and disease (0.2%). On overall, such population and landscape genomic studies have demonstrated the power of anonymous markers in generating insights into adaptive processes in the context of global change (De Kort et al. 2014; Pais et al. 2017).

2.2 *Using Ecology-Informed Marker Diversity to Quantify Adaptive Phenotypic Variation*

Although associations between putatively adaptive markers and environmental variation suggest the involvement of these markers in environment-driven evolutionary processes, it is necessary to assess the contribution of ecology-informed marker variation to adaptive phenotypic variation, which is the actual target of natural selection. While whole-genome sequencing and genome-wide association studies (GWAS) are too costly to identify the genetic basis of phenotypic adaptive diversity in the many species of conservation concern, landscape genomics analysis can be complemented with phenotypic trait assessment to reveal the contribution of modest marker panels to adaptive trait variation. Such a complementary approach, involving a common garden, outlier approaches, and multivariate landscape genetic tools, has been successfully used by De Kort et al. (2014) to uncover the role of ecology-informed anonymous SNPs in phenotypic adaptation of *Alnus glutinosa* populations sampled across a latitudinal gradient. Through comparing the allele frequency of anonymous SNPs with quantitative genetic data, it was demonstrated that only 15 temperature-associated outlier SNPs could explain 28% of the variation in adaptive traits, including bud set and leaf size (De Kort et al. 2014). These traits showed strong environmental clines, with delayed bud set (longer growing season) and smaller leaves (less transpiration) in genotypes originating from warmer and drier regions as compared to colder regions. On the contrary, only 5% of this adaptive phenotypic variation could be explained by a set of 1933 neutral SNPs. Although a large proportion of adaptive trait variation remained unexplained, this study showed that landscape genetic approaches can strongly improve the power of anonymous markers to assess phenotypic adaptive genetic variation (Lepais and Bacles 2014).

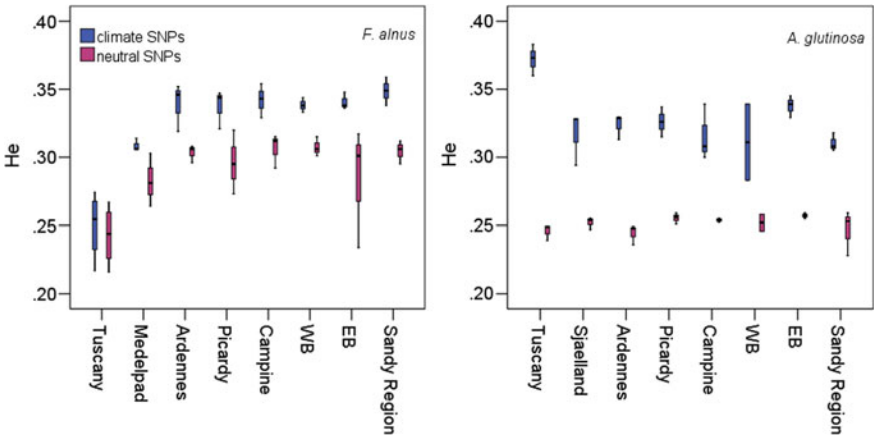


Fig. 2 Genetic diversity (He, expected heterozygosity) across Europe in two tree species for putatively adaptive versus neutral SNPs. Countries involved are Italy (Tuscany), Sweden (Medelpad), Denmark (Sjaelland), France (Picardy), and Belgium (Campine, WB, EB, Sandy Region). *WB* and *EB* are abbreviations for "Western Brabant" and "Eastern Brabant", respectively

Comparing the population genetic diversity across these adaptive markers with neutral genetic diversity (Fig. 1) further demonstrated that adaptive genetic diversity is generally higher than neutral genetic diversity in *A. glutinosa* (Fig. 2, right panel). Although it is unknown whether this is a general pattern in tree species, a similar discrepancy between adaptive and neutral genetic diversity was observed in the Glossy buckthorn (*Frangula alnus* (syn. *Rhamnus frangula*)), (Fig. 2, left panel) (De Kort et al. 2015). The latter study was based on a limited set of genic SNPs (n = 183) aiming to infer population genetic patterns across different spatial scales. The results of both studies suggest varying adaptive genetic variation across populations and thus environment-dependent resilience to future climate change. Why adaptive genetic diversity exceeds neutral genetic diversity in both tree species remains elusive, but several eco-evolutionary processes could explain this pattern. First, long life spans may allow accumulation of selective alleles within populations to a greater extent than of neutral alleles. This is because environmental variation creates populations of older trees bearing alleles adapted to previous conditions, and younger trees bearing alleles beneficial under current conditions (Petit and Hampe 2006; Kremer et al. 2012). An *A. glutinosa* population, for example, can harbor adaptive genetic variation resulting from more than a century of environmental fluctuations. The potential role of life span could be inferred by comparing adaptive with neutral genetic diversity in short-lived species. Second, the complex genetic architecture underlying adaptive traits may support high allelic diversity within populations for the loci underlying these traits. When many small-effect loci affect an adaptive trait, interactions among loci may prevent genetic drift of alleles that are neutral or slightly deleterious in the local

environment, but are beneficial in a different environment (conditional neutrality and antagonistic pleiotropy, respectively) (Latta 1998; Colautti et al. 2012; Alonso-Blanco and Méndez-Vigo 2014).

2.3 Genomic Marker-Based Heritability Estimation

Genetic markers can be used to obtain marker-derived heritability estimates in wild and managed tree populations where pedigree information or information of trait variation under controlled environmental conditions is lacking (Ritland 2000; Bérénos et al. 2014; Castellanos et al. 2015). Dense panels of randomly positioned SNPs (tens of thousands) allow inferring realized genomic relatedness and genetic variance estimates based on the level of identity-by-state (IBS) of SNPs and can yield heritability estimates that approximate pedigree-based estimates (Robinson et al. 2013; Bérénos et al. 2014). Although obtaining such dense SNP panels remains challenging in many organisms, a few recent studies have shown that the use of a limited set of ecology-informed SNPs provides an equivalent accuracy when compared to large amounts of random SNPs when the aim is to assess the heritability of a specific (set of) phenotypic trait(s). For example, Castellanos et al. (2015) used transcriptome-derived SNPs from individual Mediterranean pine trees differing in fire resistance to obtain heritability patterns of serotiny (i.e., seed accumulation in unopened cones until the next fire). They found that a limited set of 251 fire-related SNPs provided the same accuracy as a denser panel of 1480 random SNPs with regard to serotiny inheritance.

A study focusing on a breeding population of Eucalyptus trees contrasted pedigree-informed against genomic relatedness-based heritability estimates (Resende et al. 2017). Relying on regional heritability mapping (RHM) (Uemoto et al. 2013), both rare and common SNP variants in genomic segments underlying quantitative trait variation could be identified. The focus of RHM on narrow genomic regions (a few Mbp), as opposed to genome-wide single SNP analyses (e.g., GWAS), allows detecting a larger proportion of missing genetic variation underlying adaptive traits. Resende et al. (2017) found that 24,806 random genic SNPs explained 64–89% of pedigree-informed heritability estimates in Eucalyptus wood and disease traits, while 2191 RHM SNPs and 13 GWAS SNPs (SNPs found to be associated with one or more phenotypic traits) explained 45–78% and 12–50% of the pedigree-informed heritability estimates, respectively. Thus, although most tree species lack a reference genome and are therefore unsuitable for GWAS or RHM, the Eucalyptus study demonstrated that the per-SNP contribution to trait heritability increases considerably when targeting SNPs associated with quantitative trait variation. Taken together, by focusing on ecology-informed SNPs (e.g., SNPs associated with phenotypic traits or environmental clines), anonymous SNP panels can provide important insights into the ability of natural populations to respond to changes in corresponding environmental factors (De Kort et al. 2014; Castellanos et al. 2015; Resende et al. 2017).

3 Epigenetic Evolutionary Potential

3.1 Epigenetic Variation as a Basis for Evolution

Long-lived species, including trees, experience strong environmental variation and a relatively high probability of experiencing extreme events throughout their life cycles. To cope with this environmental variability, trees require mechanisms that allow them to adjust their phenotypes quickly to changes in their environment. One such a mechanism that has received increasing attention in the last decade is the accumulation of environment-induced epigenetic variation (Verhoeven et al. 2010; Bräutigam et al. 2013; Quadrana and Colot 2016). Epigenetic variation refers to changes in gene function without changes in the underlying DNA sequence, for example, through DNA (de)methylation, histone modifications, and small RNA-based gene expression regulation (Bossdorf et al. 2008; Bonasio et al. 2010). Through their ability to drive within-generation phenotypic change, epigenetic modifications play an important role in physiological development, environmental plasticity, and immunity (Sung and Amasino 2004; Shi 2007; Saeed et al. 2014; Nicotra et al. 2015). Moreover, faithful transgenerational transmission of DNA methylation signatures has been shown to provide short- and long-term adjustments of phenotypes to local environmental change (Pennisi 2013; Verhoeven et al. 2016; Quadrana and Colot 2016). Such epigenetic adaptive potential and heritability could not only increase the probability of populations surviving extreme events associated with climate change, such as unusual droughts and the arrival of novel disease variants, but also provide the time to adapt genetically over contemporary timescales.

3.2 Assessing Adaptive Epigenetic Variation in Trees

Evidence for epigenetic evolution in trees has been accumulating recently, with the earliest study demonstrating epigenetic regulation of adaptive tree responses dating back a decade ago (Skrøppa et al. 2007; Yakovlev et al. 2010). Skrøppa and Johnsen (2000) found that temperature during embryogenesis and seed maturation causes a shift in vital phenological processes in offspring of Norway spruce. This phenotypic change lasted for more than 20 years after germination, suggesting that adaptive traits are regulated by an epigenetic memory (Skrøppa et al. 2007; Yakovlev et al. 2012). In addition, Lira-Medeiros (2010) showed that epigenetic profiles of mangrove trees based on methylation-sensitive AFLP markers were more closely associated with habitat type (riverside versus salt march) than genetic profiles based on AFLP markers. In another epigenetic study, Herrera and Bazaga (2010) focused on heterophyllous plants, i.e., plants producing different leaf types

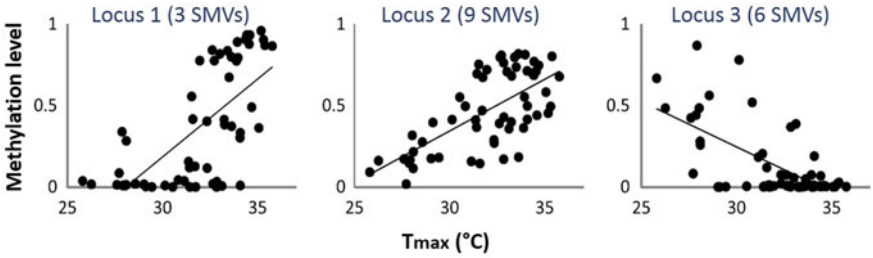


Fig. 3 Average methylation level of single methylation variants (SMVs) at three loci in relation to maximum temperature of the warmest month for 58 valley oak trees sampled across a climatic gradient in California (Gugger et al. 2016)

depending on light or herbivore conditions. These authors showed that *Ilex aquifolium* tree crowns closer to the ground exhibited more browsing damage to the leaves, and therefore showed a higher proportion of prickly leaves. Interestingly, these prickly leave exhibited less DNA methylation at particular zones of the genome compared to non-prickly leaves. In *Castanea sativa*, a decrease in DNA methylation and an increase in acetylation of histone 4 were observed during bud burst when conditions were favorable for active growth (Santamaría et al. 2009). Similarly, significantly more DNA methylation levels and lower acetylation of histone 4 were observed in *Populus* during winter than during summer, illustrating strong epigenetic regulation of phenological events (Conde et al. 2012). In a more recent study, Gugger et al. (2016) found that climate and spatial variables explained patterns in CG methylation to a larger extent than variation in SNP, CHG methylation, and CHH methylation (where H is A, C or T) across 58 Valley oak (*Quercus lobata*) individuals sampled across climatic gradients (Fig. 3). Altogether, these studies show that epigenetic signatures of adaptation may play a prominent role in adaptive evolution and in the ability of tree populations to cope with environmental change.

Although the relative contribution of inheritable versus plastic adaptive epigenetic variation to the adaptive potential of tree populations can only be derived from multi-generational quantitative genetic research (Whipple and Holeski 2016), their combined contribution can be studied analogously to anonymous genetic marker analysis. While landscape epigenetic approaches can be employed to identify ecology-informed epigenetic variation, trait assessment in the field or in common garden conditions allows inferring the potential of ecology-informed epigenetic variation to assess adaptive phenotypic trait variation. Furthermore, under the assumption that the heritable component of adaptive epigenetic variation approximates adaptive genetic variation, comparing adaptive epigenetic with genetic variation may also shed light on the plastic component of adaptive epigenetic diversity (total adaptive epigenetic variation—adaptive genetic variation $\approx$ plastic adaptive epigenetic variation).

4 Conserving and Improving Tree Adaptive Potential

Quick, cost-efficient, and accurate molecular marker-based estimates of tree adaptive potential can increase our understanding of the spatial distribution of tree adaptive potential and would allow identifying natural cold and hot spots of adaptive potential as primary targets for conservation. This in situ conservation of wild gene pools is of utmost importance because natural tree populations harbor the (epi)genetic variants that are required to breed novel tree varieties and to cope with emerging pests and extreme climatic events, among other global environmental threats (Fernie et al. 2006; Vincent et al. 2013; Castañeda-Álvarez et al. 2016). Whereas the conservation of wild gene pools as germplasm accessions in gene banks or in botanical gardens can be complementary, a recent study revealed that over 95% of crop wild relative taxa (the wild relatives of cultivated crops) are underrepresented in gene banks with regard to their full range of geographic and ecological variation (Castañeda-Álvarez et al. 2016). Thus, the vast majority of tree genetic variation underlying phenotypic variation is to be found in natural populations across the native range. Moreover, as opposed to natural populations, gene banks and other ex situ collections do not evolve novel genetic variation in response to environmental change (Schoen and Brown 2001).

Ongoing human-mediated losses of natural adaptive molecular variation underlying disease and abiotic stress resistance compromise our ability to synchronize forest resource production with both human population growth and ongoing climate change and crop pest spread (Fernie et al. 2006; Bebber et al. 2013; Trumbore et al. 2015). To prevent further losses of forest genetic resources, sustainable management and conservation of the standing genetic diversity and evolutionary potential of natural tree populations are crucial. Several strategies have been proposed to increase tree resilience toward environmental change, including assisted gene flow (e.g., climate-adjusted provenancing and composite provenancing) (Broadhurst et al. 2008; Prober et al. 2015; Aitken and Bemmels 2016), and gene flow-promoting landscaping through establishment of ecological corridors and landscape restoration (Gilbert-Norton et al. 2010; Menz et al. 2013; Krosby et al. 2015). The rationale behind assisted gene flow is to introduce genotypes that are pre-adapted to projected climate change, as to match phenotypes to future conditions through artificial fast-forward evolution. Furthermore, using mixed seed sources can reduce the risks associated with the uncertainty of climate projections. These approaches may optimize climate resilience of specific tree species in existing and newly established forests even in highly fragmented landscapes. Alternatively, conservation efforts could aim at restoring landscape connectivity, thereby promoting the flow of adaptive genetic diversity across the landscape through seed and pollen dispersal. Although dispersal is a highly species-specific process, cross-habitat landscape networking will cover many more species simultaneously as compared to species-specific approaches. As opposed to assisted gene flow, however, landscape connectivity restoration does not guarantee sufficient gene dispersal in all species to allow evolution and keep pace with climate change.

Nevertheless, through restoring natural meta-community dynamics and allowing in situ natural selection to unforeseen biotic and abiotic changes, research-informed landscape-wide restoration could greatly increase the evolutionary resilience of ecosystems to ongoing global environmental changes (Sgrò et al. 2011; Gillson et al. 2013; Menz et al. 2013; Webster et al. 2017).

Successful large-scale and cross-species forest conservation thus requires cross-species monitoring of tree adaptive potential in distinct natural settings and spanning various levels of landscape connectivity. Such an approach has been implemented in a recently established pan-European conservation network (http://portal.eufgis.org/) to safeguard tree genetic resources (Koskela et al. 2013). Today, this network counts 3428 conservation units across 34 countries, covering 101 tree species, and is monitored every 5 or 10 years for regeneration success and population size to update management strategies that serve to maintain diversity across the network. Through allowing in situ and ex situ natural selection and evolution, the *eufgis* network offers a dynamic research and conservation framework for long-term evaluation of evolutionary potential. A recent assessment of the vulnerability of the tree species covered by the conservation network predicted unfavorable climatic conditions at 33–65% of conservation units, indicating that additional research and conservation measures are necessary to protect forest genetic resources (Schueler et al. 2014). Such research ideally involves (i) landscape (epi)genetic assessment of multiple species across multiple conservation units (ii) comparing in situ with ex situ adaptive potential based on marker-informed heritability estimates, and (iii) evaluating to what extent effects of landscape connectivity on adaptive potential can be integrated in large-scale conservation strategies.

5 Concluding Remarks

Insight in tree evolutionary potential and in the resilience of forest stands with regard to ongoing global change is limited, and the development of quick, cheap, and accurate methods for inferring evolutionary potential is still in its infancy. Considering the ecological and economic importance of the ability of tree species to evolve, and the difficulty of obtaining pedigree-informed heritability estimates, there is a high demand for tools that facilitate our understanding of tree evolutionary potential. Here, we argued that ecology-informed anonymous (epi)genetic markers should be exploited to assess the adaptive potential of tree populations across environmental gradients. However, more research is needed to fully understand to what extent adaptive (epi)genomic variation represents the ability of a population to adapt its phenome to various environmental stressors such as drought and disease. Questions that need to be answered include (i) why and to what extent does adaptive potential differ among environmental settings and among tree species? (ii) what is the role of adaptive epigenetic variation in providing evolutionary potential? and (iii) how can we maximize evolutionary resilience of entire

communities and ecosystems for sustainable resource provisioning? The European tree conservation network is exemplary as it allows the screening the adaptive genetic diversity across conservation units and species and the assessment of the rate of in situ and ex situ evolution in the context of habitat fragmentation and climate change.

References

Aitken SN, Bemmels JB (2016) Time to get moving: assisted gene flow of forest trees. Evol Appl 9:271–290

Alberto FJ, Aitken SN, Alía R et al (2013a) Potential for evolutionary responses to climate change—evidence from tree populations. Glob Change Biol 19:1645–1661

Alberto FJ, Derory J, Boury C et al (2013b) Imprints of natural selection along environmental gradients in phenology-related genes of quercus petraea. Genetics 195:495–512

Alonso-Blanco C, Méndez-Vigo B (2014) Genetic architecture of naturally occurring quantitative traits in plants: an updated synthesis. Curr Opin Plant Biol 18:37–43

Bebber DP, Ramotowski MAT, Gurr SJ (2013) Crop pests and pathogens move polewards in a warming world. Nature Climate Change 3:985–988

Bérénos C, Ellis PA, Pilkington JG, Pemberton JM (2014) Estimating quantitative genetic parameters in wild populations: a comparison of pedigree and genomic approaches. Mol Ecol 23:3434–3451

Blows MW, McGuigan K (2015) The distribution of genetic variance across phenotypic space and the response to selection. Mol Ecol 24:2056–2072

Bonasio R, Tu S, Reinberg D (2010) Molecular signals of epigenetic states. Science 330:612–616

Bossdorf O, Richards CL, Pigliucci M (2008) Epigenetics for ecologists. Ecol Lett 11:106–115

Bräutigam K, Vining KJ, Lafon-Placette C et al (2013) Epigenetic regulation of adaptive responses of forest tree species to the environment. Ecology and evolution 3:399–415

Broadhurst LM, Lowe A, Coates DJ et al (2008) Seed supply for broadscale restoration: maximizing evolutionary potential. Evol Appl 1–11

Castañeda-Álvarez NP, Khoury CK, Achicanoy HA et al (2016) Global conservation priorities for crop wild relatives. Nat Plants 2:16022

Castellanos MC, González-Martínez SC, Pausas JG (2015) Field heritability of a plant adaptation to fire in heterogeneous landscapes. Mol Ecol 24:5633–5642

Christmas MJ, Biffin E, Breed MF, Lowe AJ (2016) Finding needles in a genomic haystack: targeted capture identifies clear signatures of selection in a nonmodel plant species. Mol Ecol 25:4216–4233

Colautti RI, Lee C-R, Mitchell-Olds T (2012) Origin, fate, and architecture of ecologically relevant genetic variation. Curr Opin Plant Biol 15:199–204

Conde D, González-Melendi P, Allona I (2012) Poplar stems show opposite epigenetic patterns during winter dormancy and vegetative growth. Trees 27:311–320

Cox K, Vanden Broeck A, Van Calster H, Mergeay J (2011) Temperature-related natural selection in a wind-pollinated tree across regional and continental scales. Mol Ecol 20:2724–2738

Ellegren H, Galtier N (2016) Determinants of genetic diversity. Nat Rev Genet 17:422–433

Excoffier L, Hofer T, Foll M (2009) Detecting loci under selection in a hierarchically structured population. Heredity 103:285–298

Falconer DS, Mackay TFC (1996) Introduction to quantitative genetics. Longman, Harlow UK

Fernie AR, Tadmor Y, Zamir D (2006) Natural genetic variation for improving crop quality. Curr Opin Plant Biol 9:196–202

Frichot E, Schoville SD, Bouchard G, François O (2013) Testing for associations between loci and environmental gradients using latent factor mixed models. Mol Biol Evol 30:1687–1699

Gilbert-Norton L, Wilson R, Stevens JR, Beard KH (2010) A meta-analytic review of corridor effectiveness. Conserv Biol 24:660–668
Gillson L, Dawson TP, Jack S, McGeoch MA (2013) Accommodating climate change contingencies in conservation strategy. Trends Ecol Evol 28:135–142
Granier C, Vile D (2014) Phenotyping and beyond: modelling the relationships between traits. Curr Opin Plant Biol 18:96–102
Gugger PF, Fitz-Gibbon S, PellEgrini M, Sork VL (2016) Species-wide patterns of DNA methylation variation in *Quercus lobata* and their association with climate gradients. Mol Ecol 25:1665–1680
Guichoux E, Garnier-Géré P, Lagache L et al (2013) Outlier loci highlight the direction of introgression in oaks. Mol Ecol 22:450–462
Haddad NM, Brudvig LA, Clobert J et al (2015) Habitat fragmentation and its lasting impact on Earth's ecosystems. Sci Adv 1
Hansen MC, Potapov P V., Moore R et al (2013) High-resolution global maps of 21st-century forest cover change. Science 342
Herrera CM, Bazaga P (2010) Epigenetic differentiation and relationship to adaptive genetic divergence in discrete populations of the violet Viola cazorlensis. New Phytol 187:867–876
Hoffmann AA, Merilä J (1999) Heritable variation and evolution under favourable and unfavourable conditions. Trends Ecol Evol 14:96–101
Hoffmann AA, Willi Y (2008) Detecting genetic responses to environmental change. Nat Rev Genet 9:421–432
Holliday JA, Aitken SN, Cooke JEK et al (2017) Advances in ecological genomics in forest trees and applications to genetic resources conservation and breeding. Mol Ecol 26:706–717
Izuno A, Kitayama K, Onoda Y et al (2017) The population genomic signature of environmental association and gene flow in an ecologically divergent tree species Metrosideros polymorpha (Myrtaceae). Mol Ecol 26:1515–1532
Koonin EV, Wolf YI (2010) Constraints and plasticity in genome and molecular-phenome evolution. Nat Rev Genet 11:487–498
De Kort H, Vandepitte K, Bruun HH et al (2014) Landscape genomics and a common garden trial reveal adaptive differentiation to temperature across Europe in the tree species Alnus glutinosa. Mol Ecol 23:4709–4721
De Kort H, Vandepitte K, Mergeay J, Mijnsbrugge KV, Honnay O (2015) The population genomic signature of environmental selection in the widespread insect-pollinated tree species Frangula alnus at different geographical scales. Heredity 115:415–425
Koskela J, Lefèvre F, Schueler S et al (2013) Translating conservation genetics into management: Pan-European minimum requirements for dynamic conservation units of forest tree genetic diversity. Biol Cons 157:39–49
Kremer A, Ronce O, Robledo-Arnuncio JJ et al (2012) Long-distance gene flow and adaptation of forest trees to rapid climate change. Ecol Lett 378–392
Krosby M, Breckheimer I, John Pierce D et al (2015) Focal species and landscape "naturalness" corridor models offer complementary approaches for connectivity conservation planning. Landscape Ecol 30:2121–2132
Krutovsky KV, Neale DB (2005) Nucleotide diversity and linkage disequilibrium in cold-hardiness- and wood quality-related candidate genes in douglas fir. Genetics 171:2029–2041
Latta RG (1998) Differentiation of allelic frequencies at quantitative trait loci affecting locally adaptive traits. Am Nat 151:283–292
Lepais O, Bacles CFE (2014) Two are better than one: combining landscape genomics and common gardens for detecting local adaptation in forest trees. Mol Ecol 23:4671–4673
Lira-Medeiros CF, Parisod C, Fernandes RA et al (2010) Epigenetic variation in mangrove plants occurring in contrasting natural environment. PLoS ONE 5:e10326
Manel S, Holderegger R (2013) Ten years of landscape genetics. Trends Ecol Evol 28:614–621
McKay JK, Latta RG (2002) Adaptive population divergence: markers, QTL and traits. Trends Ecol Evol 17:285–291

Menz M, Dixon K, Hobbs R (2013) Hurdles and opportunities for landscape-scale restoration. Science 339:526–527
Namroud M-C, Beaulieu J, Juge N, Laroche J, Bousquet J (2008) Scanning the genome for gene single nucleotide polymorphisms involved in adaptive population differentiation in white spruce. Mol Ecol 17:3599–3613
Neale DB, Ingvarsson PK (2008) Population, quantitative and comparative genomics of adaptation in forest trees. Curr Opin Plant Biol 11:149–155
Nicotra AB, Segal DL, Hoyle GL et al. (2015) Adaptive plasticity and epigenetic variation in response to warming in an Alpine plant. Ecol Evol 5:634–47
Olson MS, Levsen N, Soolanayakanahally RY et al (2013) The adaptive potential of Populus balsamifera L. to phenology requirements in a warmer global climate. Mol Ecol 22:1214–1230
Pais AL, Whetten RW, Xiang Q-YJ (2017) Ecological genomics of local adaptation in *Cornus florida* L. by genotyping by sequencing. Eco Evol 7:441–465
Pautasso M, Aas G, Queloz V, Holdenrieder O (2013) European ash (Fraxinus excelsior) dieback —a conservation biology challenge. Biol Cons 158:37–49
Pennisi E (2013) Evolution heresy? Epigenetics underlies heritable plant traits. Science 341
Petit RJ, Hampe A (2006) Some evolutionary consequences of being a tree. Annu Rev Ecol Evol Syst 37:187–214
Pluess AR, Frank A, Heiri C et al (2016) Genome-environment association study suggests local adaptation to climate at the regional scale in *Fagus sylvatica*. New Phytol 210:589–601
Prober SM, Byrne M, McLean EH et al (2015) Climate-adjusted provenancing: a strategy for climate-resilient ecological restoration. Front Ecol Evol 3:65
Quadrana L, Colot V (2016) Plant transgenerational epigenetics. Annu Rev Genet 50:467–491
Reed DH, Frankham R (2001) How closely correlated are molecular and quantitative measures of genetic variation? A meta-analysis. Evolution 55:1095–1103
Rellstab C, Gugerli F, Eckert AJ, Hancock AM, Holderegger R (2015) A practical guide to environmental association analysis in landscape genomics. Mol Ecol 24:4348–4370
Resende RT, Resende MDV, Silva FF et al (2017) Regional heritability mapping and genome-wide association identify loci for complex growth, wood and disease resistance traits in *Eucalyptus*. New Phytol 213:1287–1300
Ritland K (2000) Marker-inferred relatedness as a tool for detecting heritability in nature. Mol Ecol 9:1195–1204
Robinson MR, Santure AW, DeCauwer I, Sheldon BC, Slate J (2013) Partitioning of genetic variation across the genome using multimarker methods in a wild bird population. Mol Ecol 22:3963–3980
Saeed S, Quintin J, Kerstens HHD et al (2014) Epigenetic programming of monocyte-to-macrophage differentiation and trained innate immunity. Science 345
Santamaría ME, Hasbún R, Valera MJ et al (2009) Acetylated H4 histone and genomic DNA methylation patterns during bud set and bud burst in Castanea sativa. J Plant Physiol 166:1360–1369
Schoen DJ, Brown AHD (2001) The conservation of wild plant species in seed banks: attention to both taconomic coverage and population biology will improve the role of seed banks as conservation tools. Bioscience 51:960–966
Schueler S, Falk W, Koskela J et al (2014) Vulnerability of dynamic genetic conservation units of forest trees in Europe to climate change. Glob Change Biol 20:1498–1511
Sgrò CM, Lowe AJ, Hoffmann AA (2011) Building evolutionary resilience for conserving biodiversity under climate change. Evol Appl 4:326–337
Shi Y (2007) Histone lysine demethylases: emerging roles in development, physiology and disease. Nat Rev Genet 8:829–833
Skrøppa T, Johnsen Ø (2000) Patterns of adaptive genetic variation in forest tree species; the reproductive environment as an evolutionary force in Picea abies. Springer, Netherlands, pp 49–58

Skrøppa T, Kohmann K, Johnsen Ø, Steffenrem A, Edvardsen ØM (2007) Field performance and early test results of offspring from two Norway spruce seed orchards containing clones transferred to warmer climates. Can J For Res 37:515–522

Slavov GT, DiFazio SP, Martin J et al (2012) Genome resequencing reveals multiscale geographic structure and extensive linkage disequilibrium in the forest tree Populus trichocarpa. New Phytol 196:713–725

Sork VL, Aitken SN, Dyer RJ et al (2013) Putting the landscape into the genomics of trees: approaches for understanding local adaptation and population responses to changing climate. Tree Genetics & Genomes 9:901

Sork VL, Squire K, Gugger PF et al (2016) Landscape genomic analysis of candidate genes for climate adaptation in a California endemic oak, Quercus lobata. Am J Bot 103:33–46

Sung S, Amasino RM (2004) Vernalization and epigenetics: how plants remember winter. Curr Opin Plant Biol 7:4–10

Trumbore S, Brando P, Hartmann H (2015) Forest health and global change. Science 349

Uemoto Y, Pong-Wong R, Navarro P et al (2013) The power of regional heritability analysis for rare and common variant detection: simulations and application to eye biometrical traits. Front Genet 4:232

Vangestel C, Vázquez-Lobo A, Martínez-García PJ et al (2016) Patterns of neutral and adaptive genetic diversity across the natural range of sugar pine (Pinus lambertiana Dougl.). Tree Genet Genomes 12:51

Verhoeven KJF, Jansen JJ, van Dijk PJ, Biere A (2010) Stress-induced DNA methylation changes and their heritability in asexual dandelions. New Phytol 185:1108–1118

Verhoeven KJF, vonHoldt BM, Sork VL (2016) Epigenetics in ecology and evolution: what we know and what we need to know. Mol Ecol 25:1631–1638

de Villemereuil P, Frichot E, Bazin E, François O, Gaggiotti OE (2014) Genome scan methods against more complex models: when and how much should we trust them? Mol Ecol 23:2006–2019

de Villemereuil P, Gaggiotti OE, Mouterde M, Till-Bottraud I (2016) Common garden experiments in the genomic era: new perspectives and opportunities. Heredity 116:249–254

Vincent H, Wiersema J, Kell S et al (2013) A prioritized crop wild relative inventory to help underpin global food security. Biol Cons 167:265–275

Vranckx G, Jacquemyn H, Muys B, Honnay O (2012) Meta-analysis of susceptibility of woody plants to loss of genetic diversity through habitat fragmentation. Conserv Biol 26:228–237

Webster M, Colton M, Darling E et al (2017) Who should pick the winners of climate change? Trends Ecol Evol 32:167–173

Whipple AV, Holeski LM (2016) Epigenetic inheritance across the landscape. Front Genet 7:189

Willi Y, Van Buskirk J, Hoffmann AA (2006) Limits to the adaptive potential of small populations. Annu Rev Ecol Evol Syst 37:433–458

Yakovlev IA, Fossdal CG, Johnsen Ø (2010) MicroRNAs, the epigenetic memory and climatic adaptation in Norway spruce. New Phytol 187:1154–1169

Yakovlev I, Fossdal CG, Skrøppa T et al (2012) An adaptive epigenetic memory in conifers with important implications for seed production. Seed Science Research 22:63–76

Evolutionary Constraints on Coding Sequences at the Nucleotidic Level: A Statistical Physics Approach

Didier Chatenay, Simona Cocco, Benjamin Greenbaum, Rémi Monasson and Pierre Netter

Abstract Selection at the molecular level is generally measured by amino-acid alterations, for instance, through the ratio of non-synonymous and synonymous substitutions. While it is known that codons coding for identical amino acids are not perfectly identical in terms of fitness cost, e.g. due to differences in the kinetics of the associated t-RNAs, mechanisms exist for selection acting at the nucleotide level rather than the amino-acid level. In this work, we consider two such mechanisms. The first is the action of the innate immune system, with pattern recognition receptors capable of recognizing small nucleotidic motifs, such as CpG dinucleotides. Pathogens such as viruses are under this selective pressure while strongly constrained by the fact that their short genomes must code for essential proteins. A second tentative mechanism, referred to as the Ambush Hypothesis, suggests that codons are optimized to favor the presence of off-frame stop codons, which are useful to abort translation of non-functional proteins in case of accidental ribosomal

D. Chatenay
Laboratoire Jean Perrin (LJP), CNRS UMR8237, Sorbonne Universités, UPMC University Paris 06, 4 place Jussieu, Case Courrier 114, 75005 Paris, France

S. Cocco
Laboratoire de Physique Statistique, Ecole Normale Supérieure and CNRS-UMR8550, PSL Research University, Sorbonne Universités UPMC, 24 Rue Lhomond, 75005 Paris, France

B. Greenbaum
Icahn School of Medicine at Mount Sinai, Tisch Cancer Institute, 1190 One Gustave L. Levy Place, 1st Floor Box 1128 Icahn Building, New York, NY 10029, USA

R. Monasson (✉)
Laboratoire de Physique Théorique, Ecole Normale Supérieure and CNRS-UMR8549, PSL Research University, Sorbonne Universités UPMC, 24 Rue Lhomond, 75005 Paris, France
e-mail: monasson@lpt.ens.fr

P. Netter
Sorbonne Universités, UPMC University Paris 06, CNRS UMR7138, Evolution Paris Seine, IBPS, 7 quai Saint-Bernard, 75005 Paris, France

P. Pontarotti (ed.), *Evolutionary Biology: Self/Nonself Evolution, Species and Complex Traits Evolution, Methods and Concepts*,
DOI 10.1007/978-3-319-61569-1_18

frame-shift. We show how the same statistical physics inspired formalism can be applied to both questions to compute selective pressure or make predictions in a null model, called random codon model, in which the coding nature of the genomic sequence and its essential statistical features are retained. Our formalism is based on the notion of transfer matrix, developed in statistical physics to deal with systems of particles with short-range interactions; here, particles are codons and interactions result from the presence of selection mechanism acting at the nucleotidic level, possibly on contiguous codons along the sequence. Our approach is computationally efficient as it requires a computation time growing only linearly with the length of the sequence under study.

1 Introduction

Selection is generally measured in terms of modifications to proteins. A popular approach to estimate the level of evolutionary pressure on a protein is the ratio K_a/K_s for amino acid residues, which estimates the ratio between the number of non-synonymous substitutions at a particular site over the number of synonymous mutations. This approach allows one to estimate how much amino acid evolution at that site is dictated by natural selection, versus how much change an be expected randomly (Li et al. 1985; Nei and Gojobori 1986). However there are other patterns of natural selection that cannot be captured by looking at amino acid changes. In particular, synonymous mutations may not actually be equivalent, but are themselves influenced by natural selection. For instance, codon usage depends on the tissue under consideration and varies across genes. One possible explanation is that the kinetics of corresponding t-RNA varies. This can create a codon usage bias, where more favorable codon usage can offer an organism a replicative fitness advantage (Plotkin and Kudla 2011; Sharp and Li 1987). In the case of, say, an amino acid which is coded for by four codons, synonymous changes at the third position that would be assumed neutral could have a fitness cost.

A clear case where synonymous changes may have a fitness cost is when the genome of a pathogen is targeted by the innate immune system. The innate immune system is a non-specific set of receptors that may target sequence features found in pathogens, but rare or absent in host genomic material found in the receptor's location (Medzhitov and Janeway 2000). Such features may be sequence specific, such as nucleic acid motifs or structural features, and as a result nucleotide changes that alter the presence of such features will have a consequence for pathogen fitness. For instance, the CpG dinucleotide is avoided in the DNA of many genomes, and hence has become a target of the innate immune system which can detect its presence in pathogen genomes (Hemmi et al. 2000). This is just one example of sequence specific patterns which can be sensed (Vabret et al. 2016). In the case of the genomes of RNA viruses, their compact genome is mostly devoted to protein coding. Hence, if one wants to detect the evolution of recognizable patterns, the protein coding aspects of a genome become a constraint (Greenbaum et al. 2014, 2008).

To capture these evolutionary processes in a theoretical framework, we developed a formalism where selective evolutionary forces on motifs and structures are pitted against randomizing forces of constrained nucleotide sequences (Greenbaum et al. 2014). Hence, a viral genome, such as influenza, will avoid a recognizable pattern due to innate immune mediated forces, even when randomizing patterns in codon usage are accounted for in a genome constrained by protein coding and codon usage. To calculate selective and entropic forces we utilized a transfer matrix formalism from statistical physics, which was originally developed to treat systems with short-range interactions in low dimension. Here, the dimension of the "system" is one as a coding sequence can be seen as a linear chain of codons, and the effective interactions between nearest codons along the coding sequence are produced by the selective pressure acting on motifs overlapping contiguous codons. The payoff for the formal development is a reward in terms of computational speed, which allows such forces to be calculated efficiently in large datasets. We showed the forces on CpG dinucleotides in influenza, a motif predicted to be stimulatory in RNA viruses, have the greatest selective forces in influenza and HIV, and created dynamical models based on these principles (Jimenez-Baranda et al. 2011).

Here, after reviewing briefly applications of this framework, we present new results detecting abnormal short nucleotidic motifs. In particular, we present new simultaneous calculations of forces acting on different motifs. This allows us to decide whether the pressures acting on those motifs are independent or not. We also show Monte Carlo (MC) simulations of simple mutational dynamical models that reproduce the equilibrium calculations. We also better characterize the nature of the space of sequences under pressure from the immune system, in particular how similar two randomly picked up sequences are. This information can be useful to understand how constrained are viral sequences by selective pressure, and how the virus can evolve in the constrained space.

The generality of our statistical-physics formalism allows us to adapt it to detect and measure any kind of pressure acting at the nucleotidic level, not necessarily related to the immune system. An example of interest is the so-called Ambush Hypothesis introduced by Seligmann and Pollock (2004). According to the Ambush Hypothesis deleterious effects (production of long and non-functional proteins) due to ribosome frame-shifts during translation can be avoided by increasing the frequency of off-frame STOP codons. This hypothesis is similar, in spirit, to the pressure exerted by the immune system evoked above, as it acts at the nucleotidic level (to produce excess STOP codons in shifted frames by virtue of the genetic code degeneracy) under the constraint of having coding sequences (in the right frame). In the present work, we introduce a new estimator of the presence of off-frame STOP codons, which is not sensitive to the genomic AT content (contrary to most estimators). Our statistical analysis of ~ 1800 bacterial genomes shows no evidence at all in favor of the Ambush Hypothesis. In addition, extending our transfer-matrix formalism to the study of off-frame STOP codons, we compute the distribution of distances between the position at which the frameshift takes place and the first off-frame STOP codon in the same random codon model used to estimate the immune system pressure. We obtain that the average distance is small

(less than 10 codons), giving further statistical evidence for the fact that, even if the Amubush hypothesis does not hold, off-frame translation rapidly aborts.

The plan of the paper is as follows. In Sect. 2 we review previous works on the estimation of selective pressure based on our statistical physics formalism. New results for nucleotidic motifs under immune pressure and the Ambush hypothesis are reported in, respectively, Sects. 3 and 4. A short discussion with perspectives is given in Sect. 5.

2 Statistical Physics Framework for Detecting Aberrant Short Nucleotide Motifs

2.1 Viral Evolution and Pressures on Nucleotide Usage

The particular problem we are studying is what drives the evolution of a virus which changes its host, and, therefore, its environment. In addition to "local pressures" whose fitness effects derive from the consequences of changing residues to protein function, there are "global pressures", such as the codon bias of the new host, or changes in the innate immune system from one host to the next. Separating these two effects can be challenging.

For example, suppose a DNA virus were to change from a non-mammalian host to a human host. That virus, if it contained many CpG dinucleotides, could stimulate the human innate immune system via Toll-like receptor 9. Such feedback could generate a selective pressure to eliminate CpG dinucleotides. At the same time, altering the number of CpGs could effect the codon usage bias of arginine codons, since two thirds of these codons start with CpGs. If such a pressure were strong enough and arginine not particularly essential, one might even imagine cases where the amino acid itself would change, in a way that might be mistaken for positive selection at the protein level if that site were examined in isolation. As shown in Greenbaum et al. (2014), such a pressure may also exist in an RNA virus, where elimination of the CpG dinucleotide was detectable in the sequence history of influenza and where the codon bias of arginine also was altered as a consequence. This non-random evolution was associated with avoiding motifs that may be detectable (Jimenez-Baranda et al. 2011).

Hence there are at least three possible selective effects: a virus may alter replication efficiency by adopting host codon usage, detectability by altering chemical signatures that bind to host immune receptors, and adaptation via mutations that alter amino acids. We have recently developed an approach from statistical physics which is particularly useful in quantifying the first two of these effects, while offering a general program for analyzing sequences evolving under these global pressures and, therefore, broadly separating the contributions from all three types of effects. The goal is to quantify how much information one can superimpose the nucleotide sequence, at fixed amino acid sequence, thanks to the degeneracy of the

genetic code. The virus has to avoid a global pressure, such as an innate immune receptor targeting a given nucleotide word or phrase, while keeping its capability to make both viable and fit proteins, and, at the same time, operating under a host codon bias that may differ from its own.

To quantify this selective pressure acting in a coding "context" we use a random codon model (RCM) with a given codon usage and fixed amino-acid sequence. The degeneracy of the genetic code allows a number of possible genomes (sequences of codons compatible with the fixed amino-acid sequence) to code for the same protein. We associate to this number an entropic force allowing multiple synonymous mutational paths to the viral sequences in the course of evolution. We then quantify the change in entropy associated with an alteration in the number of possible genomes once a reasonable set of biological and physical constraints are imposed on a virus, such alteration is the pressure associated with moving the virus from an entropically favored configuration to a less favored one due to the external pressure exerted by the innate immune system on nucleotide phrases. In this way, we can infer when a virus is operating under a significant external pressure, since it will be in a lower probability state than the maximum entropy configuration.

In the following we review the statistical physics approach we have introduced in Greenbaum et al. (2014) to characterize the pressure associated to the number of occurrence of small nucleotidic motifs. We will start by computing for the RCM the entropy of sequences as a function of the number of occurrences of one particular dinucleotide motif. Then we draw the occurrences of the motifs sampled on the true sequence, which will correspond to a point in the distribution. The corresponding entropy will tell us how much the set of sequences is reduced or constrained by the presence of the motifs. We will define a 'pressure', equal to the derivative of the distribution in that point, to quantify the degree of such a constraint. We will study the selective pressures on all the dinucleotidic motifs in influenza and HIV viruses of different subtypes for a set of coding regions. The characterization of a given genomic viral sequence in term of the selective pressure, which is an extensive parameter and in particular does not depend on the length of the sequence, will allow us to compare all such cases. Moreover, as detailed in Greenbaum et al. (2014) the selective pressure can be followed during the evolution of a virus which adapts to a human host, and it can be shown to evolve to reach an equilibrium value. We will finally focus on *CpG* motifs and compare the selective pressures on different viruses.

In a second part of the chapter which contain new results we will extend the approach in several directions: First we will introduce a technique based on Monte-Carlo simulation to evolve in silico a sequence, starting from an initial, non-equilibrium selective pressure, to the final equilibrium value. Secondly we will also extend the approach to more motifs. In this way we will obtain a surface in a multi-dimensional space. Finally we will discuss how a selective pressure alters the space of coding sequences, in particular the loss in entropy due to a selective pressure can be associated to an increase of homology between two random sequences under the same selective pressure.

2.2 Random Codon Model: Definitions and Notations

We review here the approach introduced in Greenbaum et al. (2014). The idea is to quantify the motif frequencies in a given sequence with respect to what is expected from a random model (RCM) where the only constraints are the fixed amino acid sequence and the codon bias. We start with particular coding sequence:

$$\bar{\mathbf{C}} = \{\bar{C}_1, \bar{C}_2, \ldots, \bar{C}_L\} \ , \tag{1}$$

where $\bar{C}_i$ the ith codon coding for the ith amino-acid $\bar{a}_i$, and L is the number of amino-acids in the sequence. $\bar{\mathbf{C}}$ can be seen as a sequence of $3 \times L$ nucleotides. Let $\bar{c}_{i,\ell}$ denote the ℓ nucleotide in codon i, with $\ell = 1, 2, 3$, i.e. $\bar{C}_i = \{\bar{c}_{i,1}, \bar{c}_{i,2}, \bar{c}_{i,3}\}$. In the following we will label a nucleotide c with two indices, e.g. $c_{i,\ell}$ to indicate the codon position i and the position ℓ of the nucleotide in the codon, or, alternatively, with only one index to refer to its absolute position along the sequence, e.g. c_j, $j = 1 \ldots 3L$. We therefore have:

$$\bar{\mathbf{C}} = \{\bar{c}_{1,1}, \bar{c}_{1,2}, \bar{c}_{1,3}, \bar{c}_{2,1}, \bar{c}_{2,2}, \bar{c}_{2,3} \ldots, \bar{c}_{L,1}, \bar{c}_{L,2}, \bar{c}_{L,3}\} = \{\bar{c}_1, \bar{c}_2, \ldots \bar{c}_{3L}\} \ . \tag{2}$$

We generate random sequences $C = \{C_1, C_2, \ldots, C_L\}$ coding for the same amino acids as $\bar{\mathbf{C}}$, such that each codon in the random sequence, $C_i = \{c_{i,1}, c_{i,2}, c_{i,3}\}$ (coding for a_i), has a probability equal to the codon bias $p(C_i|a_i)$. At most six codons C_i have a non-zero probability for a given a_i. Codons are drawn independently and at random, and the probability of C is simply the product of the probabilities of the codons,

$$p(C) = \prod_{i=1}^{L} p(C_i|a_i) \ . \tag{3}$$

A motif of length K is a sequence of K characters among $\{A, C, G, T\}$, which we denote by $m = (m_1, m_2, \ldots, m_K)$. We want to compare the number of occurrences of this motif in the natural sequence,

$$\bar{N}_m = \sum_{j=1}^{3L-K+1} \prod_{k=0}^{K-1} \delta_{\bar{c}_{j+k}, m_k} \ , \tag{4}$$

to the average number of occurrences of the same motif in the RCM model,

$$\langle N_m \rangle = \sum_{C} p(C) \sum_{j=1}^{3L-K+1} \prod_{k=0}^{K-1} \delta_{c_{j+k}, m_k} \ . \tag{5}$$

Here, $\delta_{c,m}$ is the Kronecker function: $\delta_{c,m} = 1$ if the nucleotides c and m are identical, 0 otherwise. The first sum in Eq. (5) is computed over all possible codon

sequences compatible with the amino-acid content. As this number is enormous (typically, exponential–in–L), Monte Carlo simulations were used to compute such average number in Li et al. (1985); in the following we will review the faster method introduced in Greenbaum et al. (2014), based on the transfer matrix approach (Onsager 1944). We will also need to determine whether any difference between $\hat{N}_m$ and $\langle N_m \rangle$ is statistically meaningful or not. To do so, we will consider

$$\langle N_m^2 \rangle = \sum_C p(C) \left(\sum_{j=1}^{3L-K+1} \prod_{k=0}^{K-1} \delta_{c_{j+k},m_k} \right)^2, \tag{6}$$

and compare $\langle N_m \rangle - \bar{N}_m$ to the statistical fluctuation $\sqrt{\langle N_m^2 \rangle - \langle N_m \rangle^2}$ within the random codon model.

2.3 Statistical Physics Approach: Partition Function

A way to calculate the moments of the distribution of the number of motifs in the random model, borrowed from statistical physics, is to introduce the so-called partition function:

$$Z(x) = \sum_C p(C) \exp \left(x \sum_{j=1}^{3L-K+1} \prod_{k=0}^{K-1} \delta_{c_{j+k},m_k} \right). \tag{7}$$

The derivative

$$N_m(x) = \frac{\partial \log Z(x)}{\partial x}, \tag{8}$$

gives the average number of occurrences of the motif for the fixed parameter x. In particular,

$$\langle N_m \rangle = \left. \frac{\partial \log Z(x)}{\partial x} \right|_{x=0} \tag{9}$$

is the average number of times the motif is found in the unbiased RCM, as can be verified by comparing with Eq. (5). Similarly, the second derivative of the partition function gives access to the variance of the number of motifs:

$$\langle N_m^2 \rangle - \langle N_m \rangle^2 = \left. \frac{\partial^2 \log Z(x)}{\partial x^2} \right|_{x=0}, \tag{10}$$

as can be verified by comparing with Eq. (6). More generally all the moments of the distribution of the number of motifs can be calculated from the derivatives of the partition function in $x = 0$.

2.4 Constrained Model, Maximum Entropy Approach, Legendre Transform and Selective Force

In this section the analogy with statistical physics is further developed, and we show that the partition function introduced above can be considered for arguments $x \neq 0$. Parameter x will play the role of a (selective) force, constraining the distribution of the codons in the RCM to have a given average number of occurrence of the motif under consideration. Following the maximal entropy principle introduced by Jaynes (1957) the least constrained, or maximal entropy distribution $P(C|x)$ capable of reproducing the average number $N_m(C)$ of occurrence of a motifs has an exponential form of the type

$$P(C|x) = \frac{1}{Z(x)} \prod_{i=1}^{L} p_i(C_i|a_i) \times \exp\left(xN_m(C)\right), \tag{11}$$

where, for simplicity, we have assumed that the codon biases are not much affected by the constraint. For $x = 0$ one recovers the unconstrained case of Eq. (3). Our aim is to find, for any given genomic sequence $\bar{C}$, the value of x for which the average number of the number of occurrences of a motif with the distribution $P(C|x)$ corresponds to the number of motifs $\bar{N}_m$ present in the sequence. Parameter x therefore satisfies the equation:

$$\sum_{C} P(C|x) \sum_{j=1}^{3L-K+1} \prod_{k=0}^{K-1} \delta_{c_{j+k},m_k} = \bar{N}_m \tag{12}$$

which is the generalization of Eq. (5) to the biased case, $x \neq 0$.

In statistical physics a Legendre transform allows one to change the description of a system containing a fixed number of particles (Canonical Ensemble) to a system in which the number of particle can fluctuate around an average value determined by the choice of the chemical potential (Grand Canonical Ensemble). Using the same description, here, we can describe the RCM by the free energy potential, i.e. minus the logarithm of the partition function, at fixed number of occurrence of a motifs N_m, or by the entropy at fixed value of the parameter x. x is an intensive parameter, similar to the chemical potential, which we call selective pressure. In the following we show how the Legendre transform relates the two potentials and how they are equivalent in the limit of long sequences. One can

rewrite the partition function in Eq. (7) by summing together all sequences having the same number of occurrences of a motif:

$$Z(x) = \sum_{N_m \geq 0} \Omega(N_m) \exp(x N_m). \tag{13}$$

where $\Omega(N_m)$ is the weighted number of nucleotide sequences (at fixed amino acid content) having N_m motifs, as each sequence is weighted by the product of the codon biases of its codons. We consider the logarithm of $\Omega(N_m)$, denoted by $\sigma(N_m) = \log \Omega(N_m)$. In the case of very long sequences the sum over N_m in (13) is dominated by its maximal contribution, obtained for the value of N_m such that

$$\frac{\partial \sigma(N_m(x))}{\partial N_m} = -x. \tag{14}$$

We therefore obtain

$$\log Z(x) \approx x N_m(x) + \sigma(N_m(x)). \tag{15}$$

or equivalently

$$\sigma(N_m(x)) = \log Z(x) - x N_m(x). \tag{16}$$

which expresses the Legendre relation between the function $\sigma(N_m)$ and minus the free energy, $\log Z(x)$.

What is the interpretation of $\sigma(N_m)$ defined above? If the sequences were not weighted by the product of their codon biases, Ω would a number of sequences, and σ would be an entropy. Due to the presence of the multiplicative weights, σ defined above is a relative entropy with respect to the unbiased distribution. Indeed, it is easy to check from Eq. (16) that σ vanishes for $x = 0$. We therefore introduce the absolute entropy of the unconstrained RCM,

$$\sigma_0 = -\sum_{i=1}^{L} \sum_{C_i} p_i(C_i) \log p_i(C_i) = \sum_{a=1}^{20} N_a \left(-\sum_{C_\alpha} p(C_\alpha|a) \log p(C_\alpha|a) \right) \tag{17}$$

where C_α are all the codons coding for the amino acid a, $\alpha = 1 \ldots deg(a)$, where $deg(a)$ is the degeneracy of the amino acid. A simple upper bound of σ_0 is obtained by considering all amino acids as having the maximal degeneracy of 6 and all the corresponding codons as equiprobable; in this case $p(C_\alpha|a) = 1/6$ and $\sigma_0 \leq L \log 6$. A more precise upper bound is to take into account the degeneracy of each amino acid $deg(a)$ but still considering each codon coding for the same amino acid as equiprobable; we then obtain the upper bound $\sigma_0 = \sum_a N_a \log deg(a)$.

The absolute entropy of sequences, defined as the logarithm of the typical number of sequences available under pressure x, is then given by

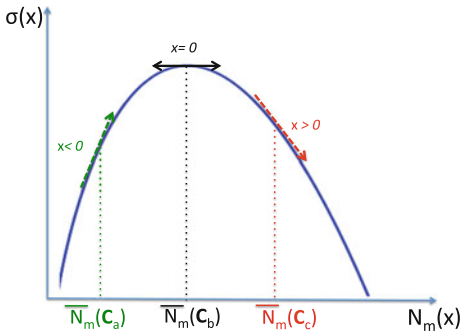


Fig. 1 Sketch of the entropy σ in the random codon model as a function of the number of occurrences of the motif, N_m. The selective pressure x associated to a given genomic sequence C with a number of motifs $\bar{N}_m$ is the derivative of the entropy σ in $N_m = \bar{N}_m$. Three cases are shown: a typical value $\bar{N}_m$ corresponding to the unconstrained case $x = 0$ (*black*, top of entropy curve); $\bar{N}_m$ atypically small, corresponding to a selective pressure $x < 0$; atypically large $\bar{N}_m$, corresponding to a selective pressure $x > 0$

$$\sigma_{tot}(x) = \sigma_0 + \sigma(N_m(x)) \ . \tag{18}$$

A sketch of the absolute entropy curve is plotted as a function of N_m in Fig. 1. The selective pressure x associated to a specific number of occurrence of motifs $\bar{N}_m$ is minus the derivative of the curve $\sigma(N_m)$ in $\bar{N}_m$, see Eq. (14). As shown in Fig. 1 the maximal value of the curve corresponds to the unconstrained case $x = 0$ and is the unconstrained entropy σ_0. Negative values of x constrain the distribution to a smaller number of occurrence of the motif with respect to the unconstrained case, while positive values of it constrain the distribution to a larger number of occurrences of the motif.

In the following section we show how to derive the curve sketched in Fig. 1 by computing, using the transfer matrix technique, the partition function and its derivative, the number of motifs, as a function of x and use Eqs. (16, 18) to obtain the entropy curve. The selective force $\bar{x}$ for a given genome is then obtained from minus the derivative of the entropy curve in $\bar{N}_m$.

2.5 *Practical Implementation with the Transfer Matrix Approach*

We calculate the normalization constant $Z(x)$, Eq. (7), using the transfer matrix formalism. We denote by $C[n : n + K - 1]$ the subsequence of K nucleotides in C, starting at position n and ending up at position $n + K - 1$. The number of occurrences of the motif $m = (m_1, m_2, \ldots, m_K)$ in a random sequence C, see Eq. (5), can be written as

$$N_m(C) = \sum_{n=1}^{3L-K+1} \delta_{C[n:n+K-1],m} \tag{19}$$

The subsequence $C[n : n+K-1]$ spreads over at most $K_c = \mathrm{Int}((K+1)/3)+1$ contiguous codons C_i in C, where Int denotes the integer part. Consider for instance the case of dinucleotide motifs m, for which $K = 2$ and $K_c = 2$ according to the formula above. The two nucleotides of such a motif can indeed be found

- at the positions 1, 2 of a single codon, say, C_i; then we have $m_1 = c_{i,1}$, $m_2 = c_{i,2}$.
- at the positions 2, 3 of codon C_i; then we have $m_1 = c_{i,2}$, $m_2 = c_{i,3}$.
- at the position 3 of codon C_i, and position 1 of codon C_{i+1}; then we have $m_1 = c_{i,3}$, $m_2 = c_{i+1,1}$.

For the sake of simplicity we assume that $K = 2$; the case of longer motifs can be treated similarly. According to the discussion above we can write

$$N_m(C) = \sum_{i=1}^{L-1} F(m, C_i, C_{i+1}) \ , \tag{20}$$

where

$$F(m, C_i, C_{i+1}) = \delta_{m_1,c_{i,1}}\delta_{m_2,c_{i,2}} + \delta_{m_1,c_{i,2}}\delta_{m_2,c_{i,3}} + \delta_{m_1,c_{i,3}}\delta_{m_2,c_{i+1,1}} \tag{21}$$

for all $i = 1, \ldots, L-2$ and

$$\begin{aligned} F(m, C_{L-1}, C_L) = {} & \delta_{m_1,c_{L-1,1}}\delta_{m_2,c_{L-1,2}} + \delta_{m_1,c_{L-1,2}}\delta_{m_2,c_{L-1,3}} + \delta_{m_1,c_{L-1,3}}\delta_{m_2,c_{L,1}} \\ & + \delta_{m_1,c_{L,1}}\delta_{m_2,c_{L,2}} + \delta_{m_1,c_{L,2}}\delta_{m_2,c_{L,3}} \ . \end{aligned} \tag{22}$$

The expression for F in the bulk of the sequence ($i \leq L-1$) avoids double counting of the motif occurrences.

We now rewrite $Z(x)$ as a sum over the possible codons corresponding to the same amino acids as in the viral sequence C_0:

$$Z(x) = \sum_C \left(\prod_{i=1}^{L} p_i(C_i|a_i) \right) \exp[x \sum_{i=1}^{L-1} F(m, C_i, C_{i+1})] \tag{23}$$

$$= \sum_C \prod_{i=1}^{L-1} (p_i(C_i|a_i) \exp[x F(m, C_i, C_{i+1})]) \, p_L(C_L|a_L), \tag{24}$$

where $p_i(C_i|a_i)$ is the codon bias for codon C_i (coding for the ith amino acid a_i). Let us now define L 'transfer' matrices M_i, $i = 1, \ldots, L$. The dimension of matrix M_i is $\deg(C_i) \times \deg(C_{i+1})$, where $\deg(C)$ is the degeneracy of codon C. The entries of M_i are given by, for all $i = 1, \ldots, L-2$,

$$M_i(C_i, C_{i+1}) = p_i(C_i|a_i) \exp[x F(m, C_i, C_{i+1})] \,, \tag{25}$$

and

$$M_{L-1}(C_{L-1}, C_L) = p_{L-1}(C_{L-1}|a_{L-1}) \exp[x F(m, C_{L-1}, C_L)] p_L(C_L|a_L) \,. \tag{26}$$

Then, we observe that

$$\begin{aligned} Z(x) &= \sum_{C_1, C_2, \ldots, C_{L-2}, C_{L-1}} M_1(C_1, C_2) M_2(C_2, C_3) \ldots M_{L-2}(C_{L-2}, C_{L-1}) M_{L-1}(C_{L-1}, C_L) \\ &= \sum_{C_1, C_L} (M_1 \times M_2 \times \ldots \times M_{L-2} \times M_{L-1})(C_1, C_L) \,, \end{aligned} \tag{27}$$

where $\times$ denotes the matrix product in the formula above. This formula shows that Z can be computed in a time growing linearly with L only. This is a huge gain compared to the original expression of Z, Eq. (7) in main text, which sums up an exponentially large–in–L number of codon configurations.

In practice we define the $\deg(C_L)$-dimensional vector v_L, with entries $v_L(C_L) = 1$ for all codons C_L coding for amino-acid a_L. Then we compute the vector

$$v_{L-1}(C_{L-1}) = \sum_{C_L} M_{L-1}(C_{L-1}, C_L) v_L(C_L) \,. \tag{28}$$

Then, we sum over all possible values for the $(L-1)$th codon, C_{L-1}:

$$v_{L-2}(C_{L-2}) = \sum_{C_{L-1}} M_{L-2}(C_{L-2}, C_{L-1}) \, v_{L-1}(C_{L-1}). \tag{29}$$

The process is iterated until the first codon:

$$v_1(C_1) = \sum_{C_2} M_1(C_1, C_2) \, v_2(C_2). \tag{30}$$

Finally, we obtain the value of the normalization constant through

$$Z(x) = \sum_{C_1} v_1(C_1). \tag{31}$$

When the motif is of longer length, and overlap with K_c contiguous codons, Eq. (20) has to be modified. In general one can write

$$N_m(C) = \sum_{i=1}^{L-K_c+1} F(m, C_i, C_{i+1}, \ldots, C_{i+K_c-1}) \,, \tag{32}$$

where the function F is an obvious extension of Eqs. (21) and (22). The transfer matrix method, shown above can still be used, but at a price of introducing larger transfer matrices M_i.

2.5.1 Example on Two Very Short Sequences

We will first apply the above framework on two simple examples: the derivation of the entropy associated to the number of motifs CpU (the letter p indicates that the nucleotide C and U are consecutive on the phosphate backbone) for the sequences $L = 2$ or $L = 3$ amino acid of type proline, which we will indicate as $C_1 = Pro - Pro$ and $C_2 = Pro - Pro - Pro$. The proline is a $\alpha = 1 \ldots deg(Pro) = 4$ time degenerate amino acid coded by the following codons: $C_1 = CCU$, $C_2 = CCC$, $C_3 = CCA$, $C_4 = CCG$. Considering an uniform codon bias $p(C_\alpha) = 1/4$ the average numbers of occurrence of the motif CpU in the unconstrained case is $\langle N_m \rangle = 0.5$ for C_1 and $\langle N_m \rangle = 0.75$ for C_2.

In Fig. 2 we plot the total entropy $\sigma_{tot}(N_m)$ versus the number N_m of occurrences of CpU for C_1 and C_2. The maximum of the entropy always corresponds to the unconstrained case $x = 0$, and we obtain $\sigma_0 = L \log(4)$ giving 2.77 and 4.16 for the two sequences. In Fig. 2 (left) we plot the entropy for C_1. The two extreme points of the entropy curve corresponds to $\langle N_m \rangle = 0$, $\sigma = 2.197$: there are $e^{2.197} = 9$ sequences compatible with ProPro without CpU, and for $\langle N_m \rangle = 2$, $\sigma = 0$: there is a single sequence compatible with ProPro and including 2 CpU. For $\langle N_m \rangle = 1$ we obtain $\sigma = 2.472$ and e^σ is larger than 6 (the number of sequences compatible with

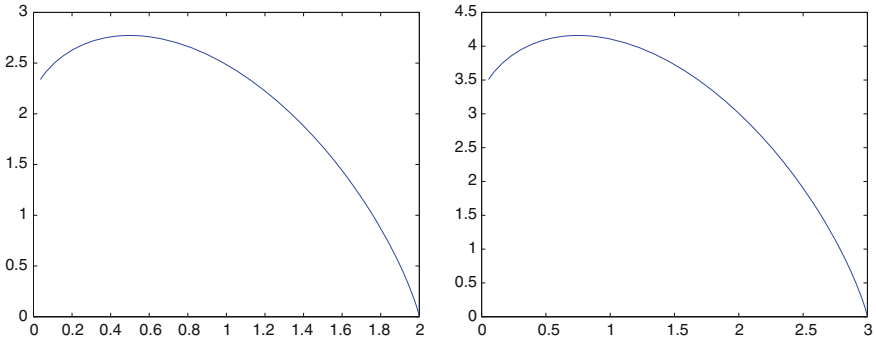


Fig. 2 Entropy σ_{tot} of sequences $C_1 = Pro - Pro$ (*left*) and $C_2 = Pro - Pro - Pro$ (*right*) as functions of the average number of occurrences of the motif CpU

ProPro with one CpU). This is because $\langle N_m \rangle$ does not coincide with N_m. As illustrated above we calculate the entropy of sequences that contain in average $\langle N_m \rangle$ repetitions of the motif, and not exactly N_m repetitions of the motif. Only for large values of N we expect that N_m will coincide with $\langle N_m \rangle$ up to negligible relative fluctuations. The entropy of sequences containing exactly 0 times the motif or two times the motif coincides with what we calculate because there is only one way to obtain zero time the motif (neither in the first nor in the second codons) or two times the motif (both in the first and in the second codons). In Fig. 2 (right) we plot the entropy curve for C_2. The total entropy of sequences with zero occurrence of the motifs is $\sigma \simeq 3.3$ and the number of sequences with zero occurrence of the motif is $e^{3.3} = 27$. The number of sequences with 3 times the motif is $\exp(\sigma)$, with $\sigma \simeq 0$.

2.5.2 Illustration on a Influenza B Sequence

In Fig. 3 we show the entropy curve obtained for an influenza B sequence with respect to the dinucleotide motifs *CpG* (left) and *ApC* (right) and with the segment codon bias. Influenza B is a virus for which humans have been a natural host for many centuries. As expected the number of CpG dinucleotides varies little over time. The green line correspond to the maximal unconstrained entropy $\sigma_0 \simeq \sum_a N_a deg(a)$ which is the same in the two cases. The red value correspond to the occurrence of number of *CpG* and *ApC* motifs in a typical sequence for Influenza B. For *ApC* the curve is quite flat (weak pressure x), hence the number of occurrences of ApC dinucleotides may largely and randomly vary. On the contrary for the *CpG* motif the selective force corresponding to the influenza B genomic sequence is large and negative, indicating that there is an important selective pressure to reduce the number of *CpG* in the sequence. The entropy of random sequences with the same number of *CpG* motifs and the same selective pressure is largely reduced with respect to the maximal, unconstrained value.

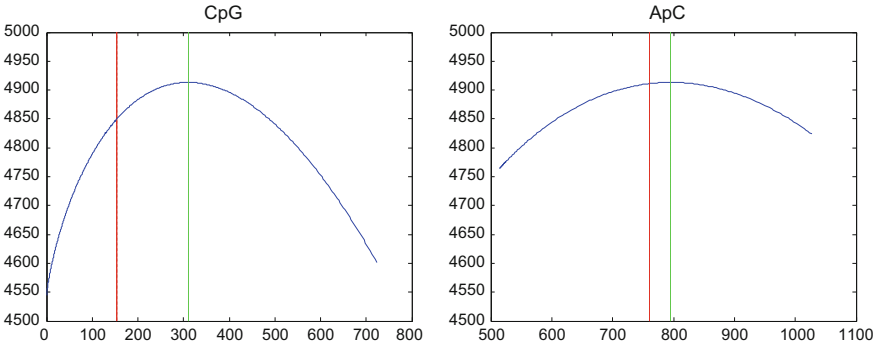


Fig. 3 *Left* Entropy σ of a influenza B isolate with its own codon bias for the dinucleotide CpG. *Right* Entropy σ of an influenza B isolate with its own codon bias for the dinucleotide ApC

2.5.3 Finding Quickly the Right Value for x

An important problem is to find the values of the entropy and of x, hereafter called $\bar{x}$, corresponding to the number $\bar{N}_m$ of occurrences of the motif in the real virus sequence. One way to do this is to compute the entropy, $\sigma(x)$, and the average number of occurrences, $N_m(x)$, for many values of x on a grid and try to be as close as possible to the data, i.e. choose $\bar{x}$ such that $N_m(x) \simeq \bar{N}_m$. A much faster procedure is the following. Consider the function

$$G(x) = \log Z(x) - x\bar{N}_m. \tag{33}$$

Two important facts about G are:

- G is a convex function of x, as its second derivative is positive:

$$\frac{d^2}{dx^2} G(x) = N_m^2(x) - N_m(x)^2 \geq 0. \tag{34}$$

- the first derivative of G vanishes when x takes the value we are looking for, since

$$\frac{d}{dx} G(\bar{x}) = N_m(\bar{x}) - \bar{N}_m = 0\ . \tag{35}$$

Hence, G has a unique minimum in $x = \bar{x}$, and we can find it very quickly with standard optimization techniques, e.g. the Newton-Raphson algorithm. Here is the procedure:

1. Start with $x = 0$
2. Compute the first and second derivatives of G in x, that is, $D_1 = N_m(x) - \bar{N}_m$ and $D_2 = N_m^2(x) - N_m(x)^2$.
3. compute the new value of x (which would be exact if G were a parabolic function)

$$x \to x - \frac{D_1}{D_2}\ . \tag{36}$$

4. Iterate step 2 until convergence is achieved.

As the parabolic approximation is generally good, we can expect that the procedure will converge very fast, in a few iterations.

2.6 *Results on Selective Pressures on Viral Sequences*

In Greenbaum et al. (2014) we have applied the above approach to influenza and HIV viral sequences. Here we recall some of the main results.

2.6.1 Influenza

We have first computed the selective force on all 16 possible dinucleotide motifs for the eight longest open reading frames from the lineage of H1N1 viruses that descend from the 1918 pandemic influenza. In Fig. 4 we show the results focusing on four dinucleotides most frequently found to be anomalous motifs and only on the PB2 gene influenza, which is the longest gene. We observe that

- The motif with the largest negative selective pressure is dinucleotide *CpG*; for this motif there is a clear evolution of the selective pressure from year 1918 when H1N1 entered the human population to much lower values, corresponding to influenza B, which has been in the human population since hundreds of years. The selective pressure has become more and more negative and the number of *CpG* dinucleotides has been lowered in the course of the viral evolution to adapt the viral sequence to the human host and avoid recognition by the immune system, which would recognize large numbers of *CpG* motifs.
- The vast majority of motifs, not represented in Fig. 4, see Fig. 2a of Greenbaum et al. (2014), have $x = 0$ when using the segment codon bias and x going from

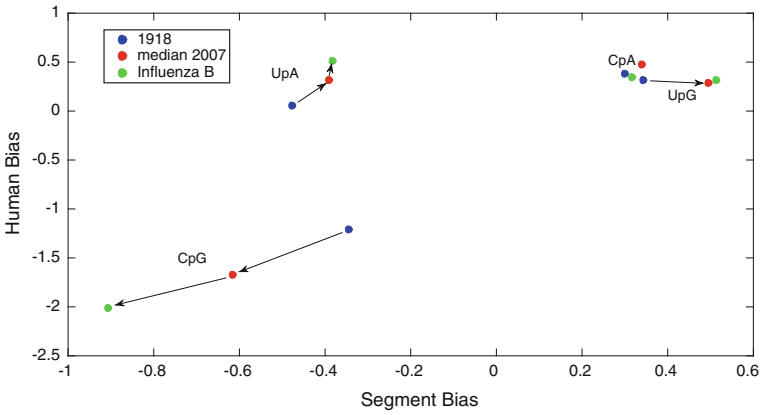


Fig. 4 A comparison of the selective pressures when calculated using the segment and human codon biases for the four dinucleotides CpA, CpG, UpA and UpA for the PB2 gene in influenza. These quantities are calculated for the 1918 H1N1, the H1N1 segments from 2007 and for Influenza B. In the later two cases the median values are shown. The arrows follow the evolution of the flu from the H1N1 1918 influenza through 2007 to influenza B (present in humans for a very long time)

$x = -1$ to $x = 1$ when using the human codon bias. This result shows that even if the virus codon bias is very similar to the one of the host it is not yet completely equivalent.

- The dependence of the selective force on the segment similarity is not very large, as shown here for PB2, it is only noticeable for CpG dinucleotides.

2.6.2 HIV

For HIV we show in Fig. 5 the selective force on six dinucleotide motifs for the Pol gene. Points of interest include:

- As for influenza sequences the motif with largest and negative pressure is CpG.
- Likewise, the vast majority of motifs have $x = 0$ when using the human codon bias and x going from $x = -1$ to $x = 1$ when using the human codon bias.
- There is some dependence on the type of protein and on the region of the sequence (not shown here, see Fig. 4d and Supplementary material in Greenbaum et al. (2014)), likely reflecting that HIVs genome codes for multiple proteins and, as a retrovirus, is targeting by many innate defense mechanisms (Vabret et al. 2016).
- There is not much dependence on the HIV subtype, showing that there is not a large evolutionary trend between different types of HIV virus which therefore seems to be already in equilibrium with respect to the small dinucleotide motif usage. This likely reflects that whereas influenza entered humans from avian and swine hosts, HIV came from primates, which are closer evolutionary species.

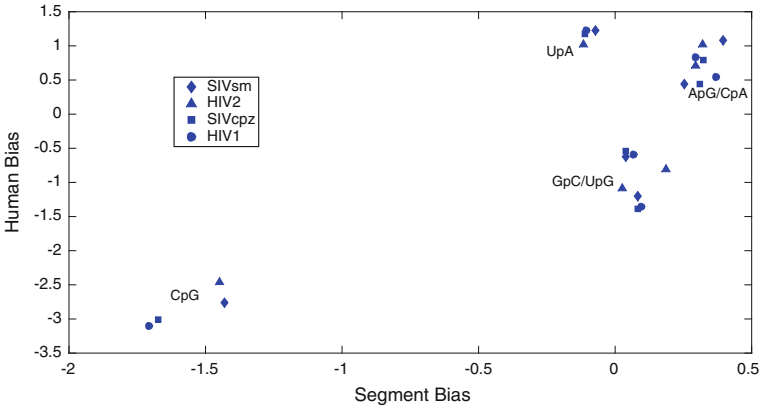


Fig. 5 A comparison of the selective pressures when calculated using the segment and human codon biases for six dinucleotides for the for the Pol genes in HIV. These quantities are calculated for the HIV1, HIV2, SIVcpz and SIVsm

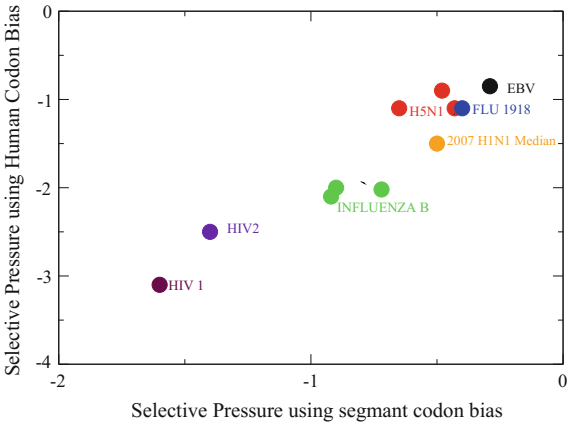


Fig. 6 Comparison of selective pressures for CpG dinucleotides using both segment codon bias and human codon bias for different viruses: influenza virus (segment PB2) with the 1918 H1N1 sequence, and the median values for all 2007 H1N1 and three influenza B segments. We also show results for Ebola virus and HIV pol (showing median values for HIV-1 and HIV-2)

2.6.3 Comparison of Different Viruses: Relationship Between the Selective Pressure and the Virulence of the Virus

The advantage of the approach presented here is that the forces associated with a given genomic sequence is an intensive variable; it is then independent of the length of the sequence and therefore different viral sequences can be compared. In Fig. 6 we compare the selective forces on *CpG* motifs for the 1918 H1N1 influenza sequence, for the median sequence from 2007 H1N1, and for the median sequence of recent Ebola virus and for the HIV1 and HIV2 median Pol sequences. Interestingly Ebola, 1918 H1N1 and 2007 H1N1 cluster together at values of the selective force which are weakly negative, while for influenza B and HIV they are much larger and negative. There is therefore a large correlation between a value of the selective pressure larger than the 'stationary' equilibrium value for influenza B and the degree to which these sequences have evolved in humans or closely related species, which may also be associated with an aberrant innate response.

3 Further Applications of the Statistical Physics Approach to Detect Anomalous Motif Usage

3.1 Monte Carlo Simulations of the Evolutionary Dynamics of Sequences

In Greenbaum et al. (2014) we have investigated a simple general dynamical model which describes the evolution of the selective pressure in the H1N1 flu virus to reach the equilibrium value:

$$\tau \frac{dN}{dt} = -x(N_m(t)) + x_{eq} \tag{37}$$

where $N(t)$ is the number of occurrences of motif m at time t. The underlying idea was directly inspired from the so-called Langevin relaxation equation of statistical physics: the dynamical variable (here, the number of motifs) relaxed to an equilibrium value where the forces acting on this variable (here, the selective and entropic pressures) balance each other. We assumed that influenza B is at equilibrium, given that the number of CpG motifs in that virus did not change much over the same time scales under which a substantial change was observed in H1N1. We therefore estimated the equilibrium pressure x_{eq} as the mean value of the pressures computed for the set of influenza B sequences. We chose for initial condition the H1N1 sequence from 1918, which had a well defined number of motifs, N_0, and the corresponding pressure, x_0.

We have solved Eq. (37) and obtained the instantaneous selective pressure $x(t) \equiv x(N(t))$, where t is the years of evolution from 1918. The time scale τ was tuned to make $x(t)$ fit best with H1N1 data over the available time range. As the pressures were (in absolute value) of the order of the unity, τ could be interpreted as the typical times it takes for the virus to decrease or increase its number of motifs by unity (see Fig. 3 in Greenbaum et al. (2014) and the values of x_B, x_0, and τ given in Table 1 of this reference).

Here we report new Monte Carlo (MC) simulations of a microscopic mutational model for the sequence of codons (with fixed amino-acid content) under constant selective pressure, denoted by x_s and supposed to be negative. The MC algorithm works in discrete time $T = \Delta t, 2\Delta t, 3\Delta t, \ldots$ as follows, from an initial sequence $C = (c_1, c_2, \ldots, c_L)$ of codons at time $T = 0$:

1. at each time step $T \to T + \Delta t$ a site i is chosen uniformly at random between 1 and L;
2. a codon C' corresponding to the ith amino acid a_i is chosen at random with probability $p_i(C'|a_i)$. If $C' = C_i$ the algorithm loops to step 1.
3. if $C' \neq C_i$ we compute the change in the number of motif occurrences ΔN_m. The move $C_i \to C'$ is always accepted if $\Delta N \leq 0$, and is accepted with probability $\exp(x_s \Delta N_m)$ if $\Delta N_m > 0$. The algorithm then loops to step 1.

This microscopic dynamics obeys detailed balance (i.e. corresponds to a general time-reversible process) and is guaranteed to converge to equilibrium at large enough times. We show in Fig. 7 typical runs of the MC algorithm for various values of the pressure (see caption). We compare the behaviour of $N_m(T)$ with the solution of (37), and observe a very good agreement of the two curves provided the elementary time-step is chosen to be $\Delta t \simeq \tau/250$.

The Monte Carlo algorithm can be used to artificially evolve sequences, starting from an initial sequence, say, the 1918 H1N1. As time goes on, the content in amino acids remains fixed, but the nucleotidic sequence changes. When the MC dynamics is stopped the resulting codon sequence may have very different

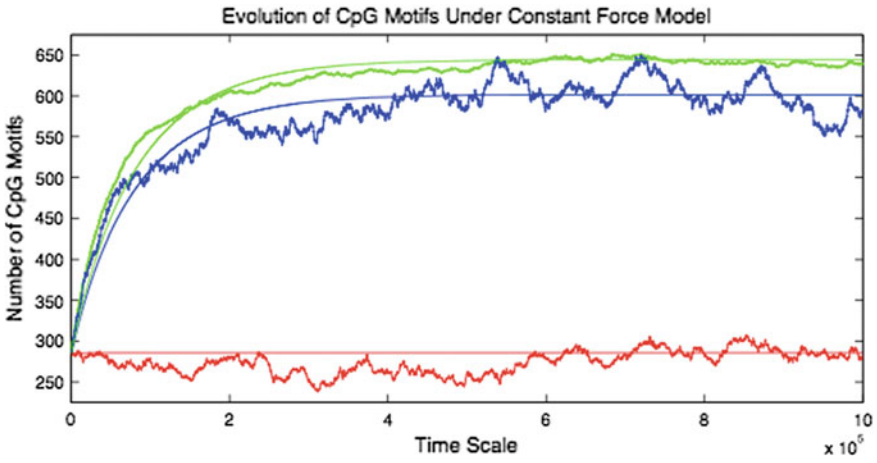


Fig. 7 Monte Carlo dynamics compared to average number of CpG motifs for three constant selective pressure values: 0, −0.119, and −1.19. These pressure values are shown in green, blue, and red respectively. In the last case the selective pressure was roughly the same as the one of the 1918 H1N1, which is the initial condition for all three trajectories

properties (compared to the initial sequence) in term of stimulation of the immune response, and can in particular be much less immuno-stimulatory, if the number of *CpG* motifs has been reduced under the action of the selective pressure.

3.2 *Entropy of Multiple Motifs*

To calculate the entropy associated with the number of occurrences of several motifs, one can extend the formalism of Sect. 2. As an example, for two dinucleotides the partition function will vary over two parameters (x_1, x_2) corresponding to dinucleotide motifs $m_1 = (m_{11}, m_{12})$ and $m_2 = (m_{21}, m_{22})$. The partition function naturally becomes

$$Z(x_1, x_2) = \sum_C p(C) \exp\left[x_1 \sum_{i=1}^{L-1} M_{1i}(C_i, C_{i+1}) + x_2 \sum_{i=1}^{L-1} M_{2i}(C_i, C_{i+1})\right], \quad (38)$$

where $M_{1i}(C_i, C_{i+1})$ is the previously defined matrix $M_i(C_i, C_{i+1})$ for the motif m_1., and M_{2i} its counterpart for motif m_2. The Legendre transformation will become

$$\sigma(x_1, x_2) = \log Z(x_1, x_2) - x_1 N_{m1}(x_1, x_2) - x_2 N_{m2}(x_1, x_2), \quad (39)$$

where

$$N_{m_1}(x_1, x_2) = \frac{\partial}{\partial x_1} \log Z(x_1, x_2) \tag{40}$$

and likewise for $N_{m2}(x_1, x_2)$. Then the average number of occurrences of motif m_1 can be computed from the partial derivative of Z with respect to x_1,

$$\langle N_{m_1} \rangle = \frac{\partial}{\partial x_1} \log Z(x_1, x_2)\Big|_{x_1=x_2=0}. \tag{41}$$

Similarly, the joint moments of the numbers of occurrences of m_1 and m_2 can be obtained from higher derivatives with respect to x_1 and x_2.

An application of the di-motif formalism is shown in Fig. 8, where we plot the entropy surface as a function of N_{UpA} and N_{CpG}. The value of the entropy constrained to the measured number of occurrence N_{UpA} and N_{CpG} in a particular sequence is smaller than the unconstrained, maximal value. The pressures $x_{ApC+CpG}$ and $x_{CpG+ApC}$ are the derivative of the entropy curve along the two axes.

An interesting question is if the selective pressures for multiple motifs are coupled, i.e. are different from the values obtained by considering one motif at a time. In Fig. 9 we compare the uncoupled (red dots) and coupled (blue) pressures for four motifs in PB1 segment. Results show that the UpA motif is essentially independent from the CpG one, as the values of the pressure for the uncoupled RCM are very similar to the one found for the coupled UpA + CpG RCM. On the

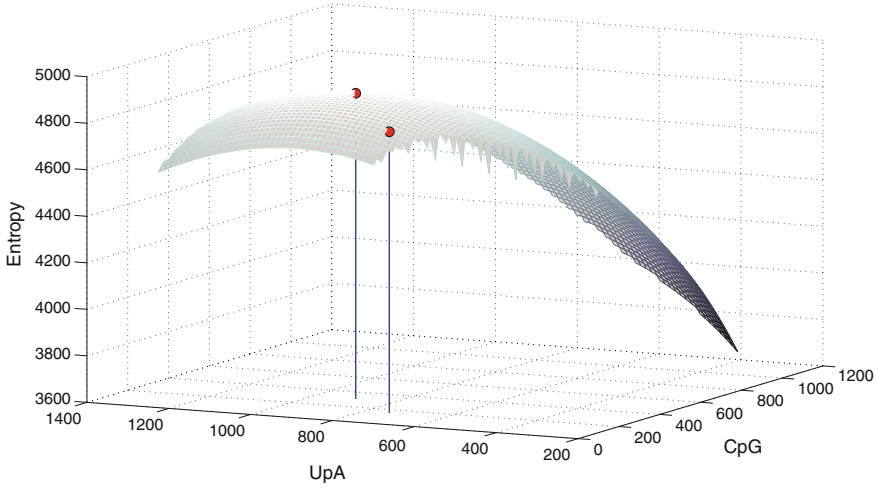


Fig. 8 Entropy σ of influenza sequences with their own codon bias for the dinucleotides CpG and UpA. Results were obtained from the eight longest coding regions of the influenza B virus (B/Cordoba/2979/1991)

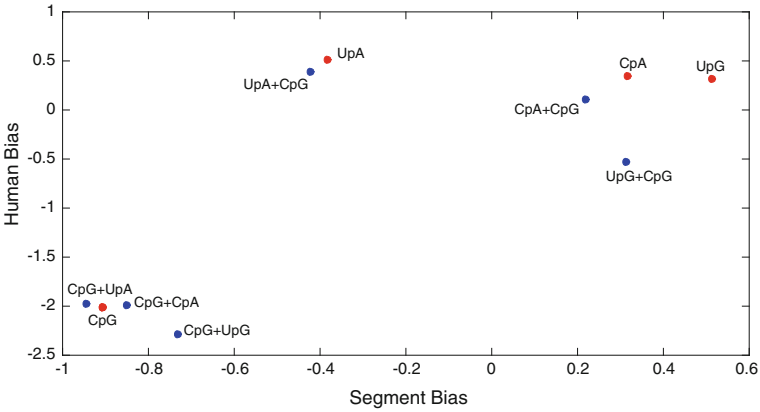


Fig. 9 Selective pressure calculated with the human codon bias and the segment codon bias for four dinucleotidic motifs in the PB1 segment of influenza B virus, calculated with RCM with two coupled motifs (*blue dots*) compared to the ones calculated with RCM with one single motif (*red dots*). For the two–motif model the selective pressure refers to the first motif in the label

contrary the selective pressures on CpA and UpG are not independent from the one of CpG. This coupling presumably originates from the fact that CpA and UpG are the mutational partners of CpG: diminishing the number of CpG motifs naturally increases the number of its mutational partners.

3.3 Geometrical Nature of the Sequence Space

So far, we have computed the entropy, that is, the log of the effective number of sequences (under some pressure). However, we do not have any information about the way those sequences are arranged in the configuration space. Are they spread over the whole configuration space or are they clustered in one tiny region? Our statistical physics formalism can however help us gain some intuition about the spatial organization of sequences as shown below.

3.3.1 Two-Sequence Formalism

Consider the following partition function, for a two-sequence system (instead of one-sequence system we have focused on so far):

$$Z_2(x,x',y) = \sum_{\{C,C'\}} \prod_{i=1}^{L} p_i(C_i|a_i)p_i(C_i'|a_i) \exp\left[x\sum_{i=1}^{L-1} M_i(C_i,C_{i+1}) + x'\sum_{i=1}^{L-1} M_i(C_i',C_{i+1}') + y\sum_{i=1}^{L}\delta_{C_i,C_i'}\right] \tag{42}$$

When $y = 0$, we simply have two independent sequences, one under pressure x and one under pressure x':

$$Z_2(x, x', y) = Z(x) \times Z(x') \ , \tag{43}$$

where $Z(.)$ is the partition function we have considered so far.

When y is not equal to zero, the two sequences are coupled according to their similarity. The weight associated to a set of two sequences is proportional to $\exp(yn_2)$; here n_2 is the number of codons equal on both sequences, it is also equal to $L - D$ where D is the Hamming distance between the two sequences (measured at the codon level, not at the base level).

We now define the average values of the number of motifs in each sequence, the average value of common codons, n_2, and a new entropy, σ_2:

$$\begin{aligned} N_m(x, x', y) &= \frac{\partial \log Z_2}{\partial x}(x, x', y), \quad N'_m(x, x', y) = \frac{\partial \log Z_2}{\partial x'}(x, x', y), \\ n_2(x, x', y) &= \frac{\partial \log Z_2}{\partial y}(x, x', y), \\ \sigma_2(x, x', y) &= \log Z_2(x, x', y) - xN_m(x, x', y) - x'N'_m(x, x', y) - yn_2(x, x', y) \ . \end{aligned} \tag{44}$$

If we choose the two pressures x and x', and we let y vary, then we can plot in a parametric way the entropy σ_2 as a function of n_2. This way, we will know how many pairs of sequences are located at a distance $d = L - n_2$. In the next paragraph we will see how this distance-dependent entropy changes as the pressures change. In general, we can choose $x = x'$ as both sequences are under the same pressure.

From a practical point of view, the calculation of Z_2 can be done along the same lines as the one of Z. The only difference is that the vectors v to be iterated are not functions of C_i only, but are now functions of both C_i, C'_i. So the maximal number of components of v is 36 instead of 6, making the computation only slightly slower.

3.3.2 Practical Implementation: Entropy as Function of Distance Between Sequences

We consider the following problem. We choose the codon bias, say, the human one, and one virus sequence, say, 1918 H1N1, and one motif, say, *CpG*. Let $\bar{N}_m$ be the number of motifs in the viral sequence, which defines the amino-acid set and the allowed codons, i.e. the probabilities $p_i(C_i|a_i)$ for all i. We want to know how many sequences (weighted by the codon bias) there are a that share n_2 codons. We consider the function

$$G(x, y; n_2) = \log Z_2(x, x, y) - 2x\bar{N}_m - yn_2. \tag{45}$$

Note that we have chosen $x = x'$ here and note also the presence of the factor 2. The variable n_2 is a positive parameter, smaller than the sequence length (measured in codons). Now, for any n_2, we can optimize G over x and y using Newton's method. The result is

$$\sigma_2(n_2) = \min_{x,y} G(x, y; n_2) \ . \tag{46}$$

The interpretation is that $\sigma_2(n_2)$ is the entropy of sequences with similarity (number of equal codons) n_2 (we neglect here the contributions coming from the fact that the average number of motifs depends on y). The maximum of the curve will be reached in n_2^*, corresponding to $y = 0$ and to the same value of x and the same entropy found in the standard one-sequence calculation. If $n_2 \neq n_2^*$, x will take a different value.

As an example of how one can interpret our results in terms of the geometry of a space of sequences, we calculate the sequence similarity for the genes of HIV and influenza. This measure shows the typical number of shared codons for two sequences drawn randomly from the distribution of possible sequences. In this case, the quantity is computed for each individual sequence when these sequences are under the derived entropic force. The average similarity (number of identical codons) between two random sequences drawn from the same codon distribution is defined as

$$n_2(x) = \sum_{C,C'} P(C|x)\ P(C'|x)\ \sum_{i=1}^{L} \delta_{C_i,C'_i} \tag{47}$$

where δ_{C_i,C'_i} equals one if the two codons at the i-th position are equal and is zero otherwise. Sequences with a large degree of similarity are close together in the space of possible sequences. In our case, for individual sequences, this would measure how close together sequences are with the same amino-acid distribution once a pressure is applied to a motif, or a set of motifs.

We plot the sequence similarity as a function of the entropy for the PB2 segment of the H1N1 virus in Fig. 10 and in Fig. 11 for the Pol gene in HIV. In much the same way as what was previously observed for the selective pressures, the similarity between sequences calculated with the RCM using the human codon bias are different to the ones obtained using the virus codon bias. The similarity is generally lower when the human codon bias is used for the background distribution rather than the bias for that segment. Overall while there is more similarity between random sequences when the segment bias is used, the difference in similarity

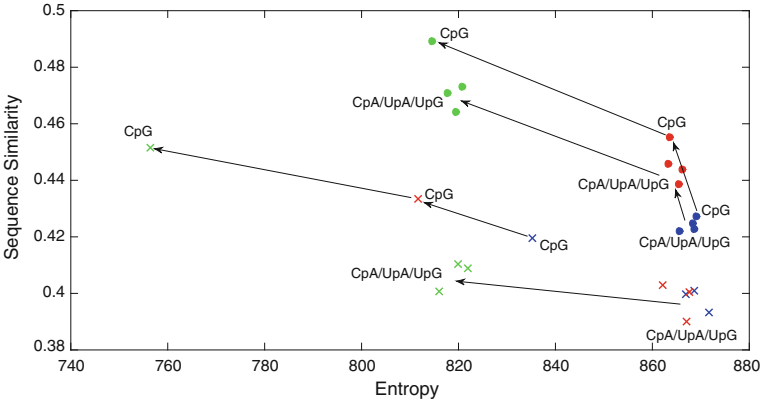


Fig. 10 Normalized sequence similarity n_2/L versus Entropy for PB2 from H1N1 flu virus (*blue* 1918 H1N1 sequence, *red* 2007 H1N1 sequence, *green* Flu B sequence for comparison). Crosses indicate the human codon bias while circles the segment codon bias

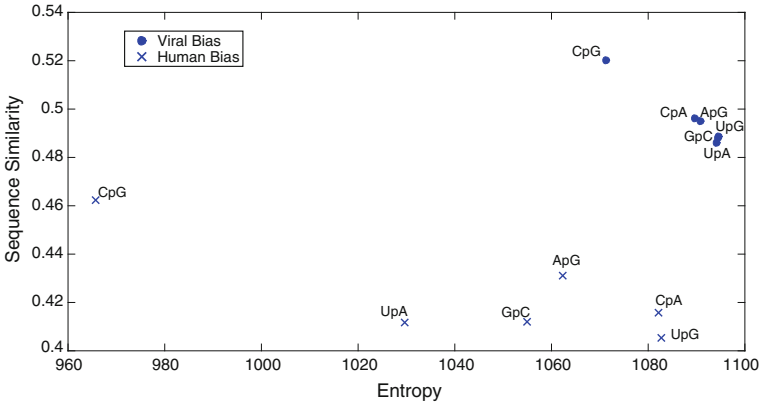


Fig. 11 Normalized sequence similarity n_2/L versus Entropy for the HIV genome from Pol HIV-1. Crosses indicate the human codon bias while circles the segment codon bias

between motifs is much larger when the human bias is used. In influenza B, with respect to the segment codon bias, the difference in similarity between CpG and other dinucleotides is much lower than the difference for the human bias.

As a general trend, for a fixed codon bias, large selective pressures lead to greater degree of similarity between sequences. The pressures, by making sequences less random, make the resulting distribution of sequences more concentrated. As expected, this effect is strong for *CpG*.

4 Out-of-Frame Stop Codons and the Ambush Hypothesis

4.1 The Ambush Hypothesis: Brief Review of Literature

Considering the deleterious effects of ribosome frame-shifts during translation Seligmann and Pollock (2004) introduced the Ambush Hypothesis according to which such deleterious effects can be avoided owing to the existence of off-frame STOP codons (OSC). This hypothesis was initially tested by Seligmann and Pollock in vertebrate mitochondrial genes (Seligmann 2010; Seligmann and Pollock 2004) and later extended to the case of prokaryotic genomes (Morgens et al. 2013; Tse et al. 2010; Wong et al. 2008). The latest study of the abundance of OSCs in prokaryotic genomes (Morgens et al. 2013) led to the conclusion that there was no statistical evidence for the existence of a correlation between a codon's usage and its propensity to form OSCs which would have been a strong evidence for the validity of the Ambush Hypothesis. Indeed, in all previous studies, the occurence of OSCs was largely dominated by the AT content of the studied genomes, and clear-cut conclusions were difficult to extract.

Here, we re-address this question along two different lines. First, we adopt a different approach in comparison with previous statistical studies. Our starting point is that apparition of an OSC involves 2 adjacent codons and thus measurement of their abundance should involve the use of the statistics of apparition of dicodons instead of mere single codons. We therefore introduce the notion of dicodon bias analogous to the well-known codon bias and refer this dicodon bias to a null model in which successive codons appear in a non-correlated way (Coleman et al. 2008; Long et al. 1998). We will adopt conventional notations for the frameshift of an OSC: within a dicodon an OSC is of type

$$\begin{cases} +1, & \text{if the OSC}'\text{S first nucleotide is the second nucleotide of the dicodon;} \\ -1, & \text{if the OSC}'\text{S first nucleotide is the third nucleotide of the dicodon.} \end{cases}$$

The study presented here is based on the use of the bacteria RefSeq database of NCBI, from which 1852 genomes of single chromosome bacterial species have been analyzed (the reduction in number of the RefSeq database was performed in order to avoid over-representation of specific bacterial species since for instance *Escherichia coli* species is represented by 173 strains in the initial database).

Secondly, since the outcome of the statistical analysis does not show any significant bias supporting the Ambush Hypothesis across all genomes, we ask whether modifying the statistics of nucleotides is actually necessary to have many OSC. To do so, we consider the random codon model of Sect. 4.2, and compute analytically within this model the distribution of distances to the first OSC after a frameshift equal to $+1$ or -1. We show that the distribution of distances decay very quickly as the distance increases, with an average distance of less than ten codons for both frameshifts. Note that this value is robust against the choice of the initial condition, i.e. also corresponds to the average distance to an OSC even if the

frameshift takes place at any location in the coding sequence (not necessarily at the beginning). Our theoretical result is corroborated by the statistical analysis of genomic sequences, and thus strongly suggests that the Ambush Hypothesis is not required to have many OSC.

4.2 Statistical Analysis of Dicodons Biases

4.2.1 Definitions and Notations

In order to quantitatively assess the occurrence of OSC within a genome we introduce the general notion of an average dicodon bias $\langle DCB_\alpha \rangle$ for dicodons belonging to a particular class α; this average dicodon bias is defined as:

$$\langle DCB_\alpha \rangle = \sum_{a,a'} p(a,a') \sum_{c,c'} (dcb(c,c') - cb(c)cb(c')) I_\alpha(c,c') \tag{48}$$

Here c (resp. c') stands for a codon and $cb(c)$ (resp. $cb(c')$) stands for the corresponding codon bias according to its usual definition, i.e. for a given amino acid a, if c codes for a, $cb(c)$ is the probability of c being chosen over all possible codons coding for a; (c,c') stands for the dicodon formed by c followed by c' and $dcb(c,c')$ stands for the dicodon bias of (c,c'). The notation a (resp. a') stands for the amino acid coded by c (resp. c'); $p(a,a')$ stands for the probability of occurence of the diamino acid (a,a'). $I_\alpha(c,c')$ is an indicator of the membership of dicodon (c,c') to a specific class α (to be specified below), and takes values 0 and 1 according to whether or not dicodon (c,c') belongs to class α. At fixed (a,a') the sum is performed over all codons c and c' coding respectively for a and a'. The definition of a dicodon bias is entirely analogous to the definition of a codon bias, i.e. for a given diamino acid (a,a') coded by (c,c') the dicodon bias for (c,c') is the probability for (c,c') to be chosen over all possible dicodons coding for (a ,a').

It should be pointed out that definition (48) of an average dicodon bias for dicodons belonging to a specific class α is a direct measure of the excess of appearance of dicodons belonging to class α with respect to the hypothesis of uncorrelated appearance of codons forming the dicodons. In addition, this estimator does not make any assumption about the statistics of di-amino acids, likely to be correlated in real coding sequences. $\langle DCB_\alpha \rangle$ can be conveniently rewritten as the dot product of 2 vectors $\vec{Y}$ and $\vec{C}_\alpha$:

$$\langle DCB_\alpha \rangle = \vec{Y} \cdot \vec{C}_\alpha \tag{49}$$

where the components of $\vec{Y}$ and $\vec{C}_\alpha$ are given by:

$$Y(c,c') = \sqrt{\omega(c,c')}X(c,c'), \quad C_\alpha(c,c') = \sqrt{\omega(c,c')}I_\alpha(c,c'), \tag{50}$$

with:

$$X(c,c') = \frac{dcb(c,c')}{cb(c)cb(c')} - 1, \quad \omega(c,c') = p(a,a')cb(c)cb(c'). \tag{51}$$

$\vec{Y}$ and $\vec{C}_\alpha$ are vectors of size $63 \times 63 = 3969$ corresponding to the formation of all possible dicodons once excluded the codon TAG which codes for non standard pyrrolysine amino acid only found in methanogenic archaea.

In order to calculate this average dicodon bias for each genome in the collection of the 1852 genomes selected from the RefSeq database, we have extracted the codon content of each CDS as well as its dicodon content; from those contents it is then easy to deduce the quantities of interest in our analysis: codon bias, dicodon bias and probability of appearance of (a,a'). In analyzing the CDS sequences the initial START codon and the sense STOP codon were excluded.

4.2.2 Statistical Significance of Calculated Values of $\langle DCB_\alpha \rangle$

Due to the limited number of codons belonging to a specific class α, it is of interest to be able to test the statistical significance of the calculated value of $\langle DCB_\alpha \rangle$. In order to perform such a test we adopt the following procedure. If n_α is the number of dicodons belonging to class α we perform **N** random permutations amongst the n_α non-zero values of the indicator $I_\alpha(c,c')$ and calculate the **N** obtained values of $\langle DCB_{\alpha,test} \rangle$; from this distribution of values of $\langle DCB_{\alpha,test} \rangle$ we then calculate a standard deviation and normalize the value of $\langle DCB_\alpha \rangle$ for the considered class with respect to this standard deviation (z-score). Following this normalization procedure a value of $\langle DCB_\alpha \rangle$ is considered as statistically significant if it is greater than 2 (in absolute value), which means away from the mean by more than twice the standard deviation of the distribution of n_α randomly chosen dicodons.

In the following we will introduce 4 classes of dicodons:

1. Class $+1$ for which (c,c') contains an OSC in the frame $+1$ associated to $\langle DCB_{+1} \rangle$;
2. Class -1 for which (c,c') contains an OSC in the frame -1 associated to $\langle DCB_{-1} \rangle$;
3. Class ± 1 for which (c,c') contains an OSC in any frame ($+1$ or -1) associated to $\langle DCB_{\pm 1} \rangle$;
4. Class *identical* for which $c = c'$ associated to $\langle DCB_{id} \rangle$.

The first (resp. second) class refers to all dicodons containing an OSC in the frame $+1$ (resp. in the frame -1); the third class refers to all dicodons containing

an OSC in whichever frame. The fourth class refers to all dicodons constituted of 2 identical codons. As a matter of example we give below the values of $I_{id}(c, c')$ for the fourth class:

$$I_{id}(c, c') = \begin{cases} 1, & \text{if } c = c'; \\ 0, & \text{otherwise.} \end{cases}$$

This fourth class is not related to the Ambush Hypothesis but will be used to validate our statistical analysis below.

In order to illustrate the statistical test explained above we present in Fig. 12 the probability density function of the $\mathbf{N}\,(=10,000)$ random permutations amongst the n_α non-zero values of the indicator $I_\alpha(c, c')$ ($n_{id} = 63$, $n_{+1} = n_{-1} = 192$ for, respectively, classes *identical*, $+1$ and -1) in the case of 2 specific genomes (*E. coli* and *Lysteria monocyogenes*).

4.2.3 Results

In a first step we report in Fig. 13 the normalized average dicodon biases for Class $+1$, Class -1, Class ± 1 and Class *identical* across all bacterial genomes. As explained above all values of $\langle DCB_\alpha \rangle$ for each genome are normalized by the standard deviation of similar distributions obtained for each studied genome and will be denoted by $\langle DCB_\alpha \rangle_{\text{norm}}$. The bottom panel of Fig. 13 refers to Class *identical*; for this class of dicodons the average bias is overall positive meaning that for the coding of 2 successive identical amino acids there is a bias towards choosing 2 identical codons. One should point out that this effect is rather weak and at the limit of being statistically significant.

The 2 upper panels refer to Class $+1$ and Class -1; quite obviously for a vast majority of genomes the Class $+1$ dicodons exhibit a bias that can be considered as showing no statistically significant deviation from 0. More interestingly the situation is quite different for Class -1 dicodons, which exhibit a statistically significant overall negative value. Grouping these 2 classes gives the third class Class ± 1, for which the overall tendency of dicodon bias values is negative (as shown on third panel from top on Fig. 13).

Before further discussing these first results we still have to test our estimator of the dicodon biases against any strong bias with respect to AT content of the considered sequences. We present in Fig. 14 the same quantities as above plotted against AT contents of the genomes used for our analysis. Quite obviously for the 4 classes of dicodons tested here there is no evidence of a strong bias of our estimator with respect to the AT content of the investigated genomes; this seems to justify our claim that our estimator for dicodon bias is a better estimator as compared to previously used estimators, see Sect. 4.1.

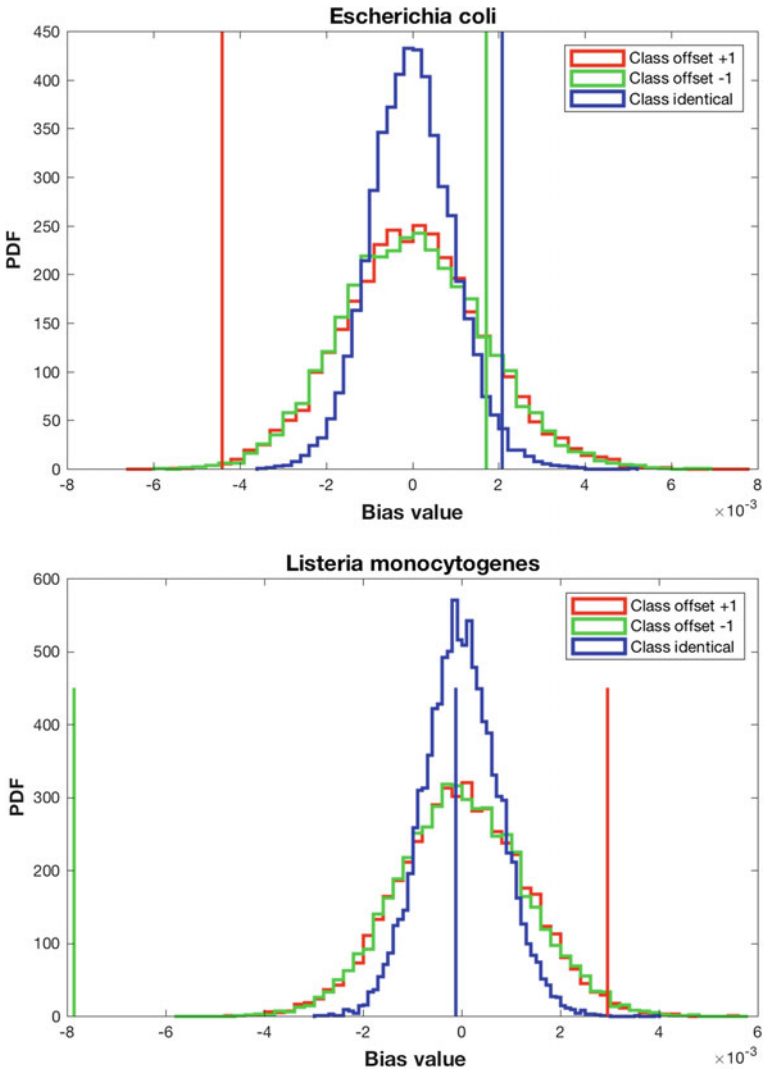


Fig. 12 Distributions of dicodon biases $\langle DCB_\alpha \rangle$ for the 3 classes *Identical*, $+1$, and -1 for two bacterial genomes, obtained by randomly reshuffling the components of vectors $\vec{C}_\alpha$, see text. *Vertical colored bars* give the values of $\langle DCB_\alpha \rangle$ for each class computed from the data. Clearly the measured value of $\langle DCB_{id} \rangle$ for *Lysteria monocyogenes* is statistically not meaningful (see position of the *vertical blue line* in the bottom panel), whereas for the same genome the value of $\langle DCB_{-1} \rangle$ is statistically meaningful (see *vertical green line* in the same bottom panel)

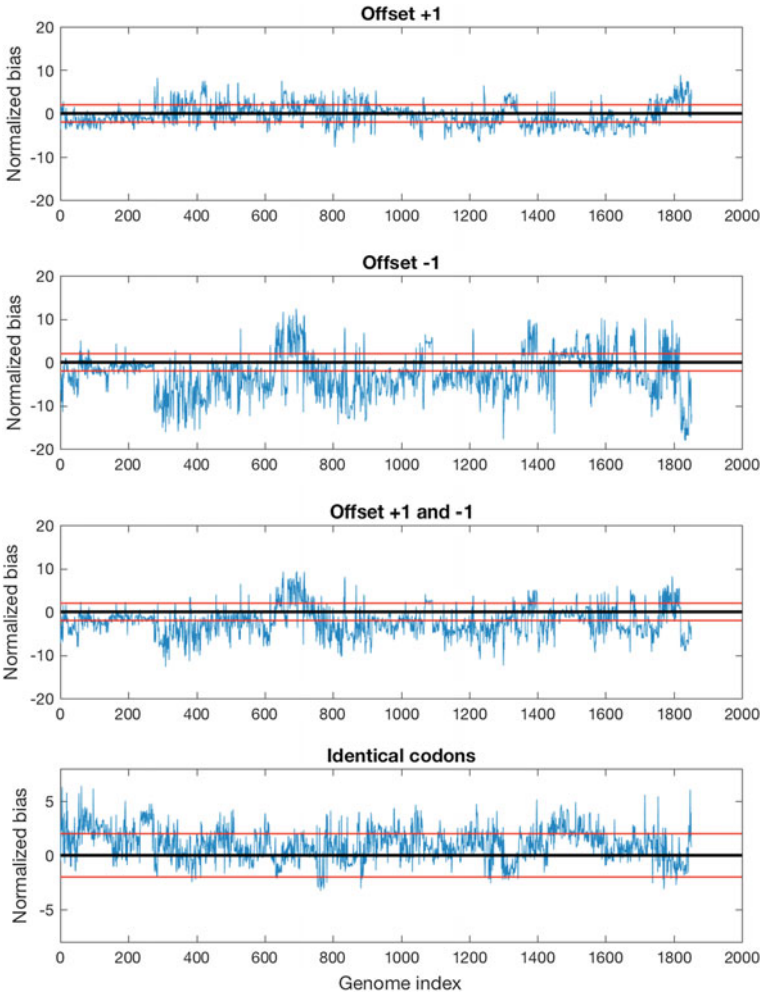


Fig. 13 Values of $\langle DCB_\alpha \rangle_{\text{norm}}$ for the 4 classes mentioned in the text. The abscissa refers to indexes of bacterial genomes in databases and *red horizontal lines* are given by $\langle DCB_\alpha \rangle_{\text{norm}} = \pm 2$; the *continuous blue lines* serve as guides to the eye. As explained in the text values of $\langle DCB_\alpha \rangle_{\text{norm}}$ above or below those *red lines* are statistically significant

We may sum up our results in the following way:

1. We have introduced an unbiased (with respect to genomic AT content) statistical indicator in which the deviations in the probability of having a stop codon out of frame are calculated with respect to the probability based on the dicodon frequencies at fixed codon bias and fixed diamino acids frequencies;
2. From this estimator we evidence a slight positive bias (at the limit of being statistically significant) for the presence of dicodons formed by identical codons

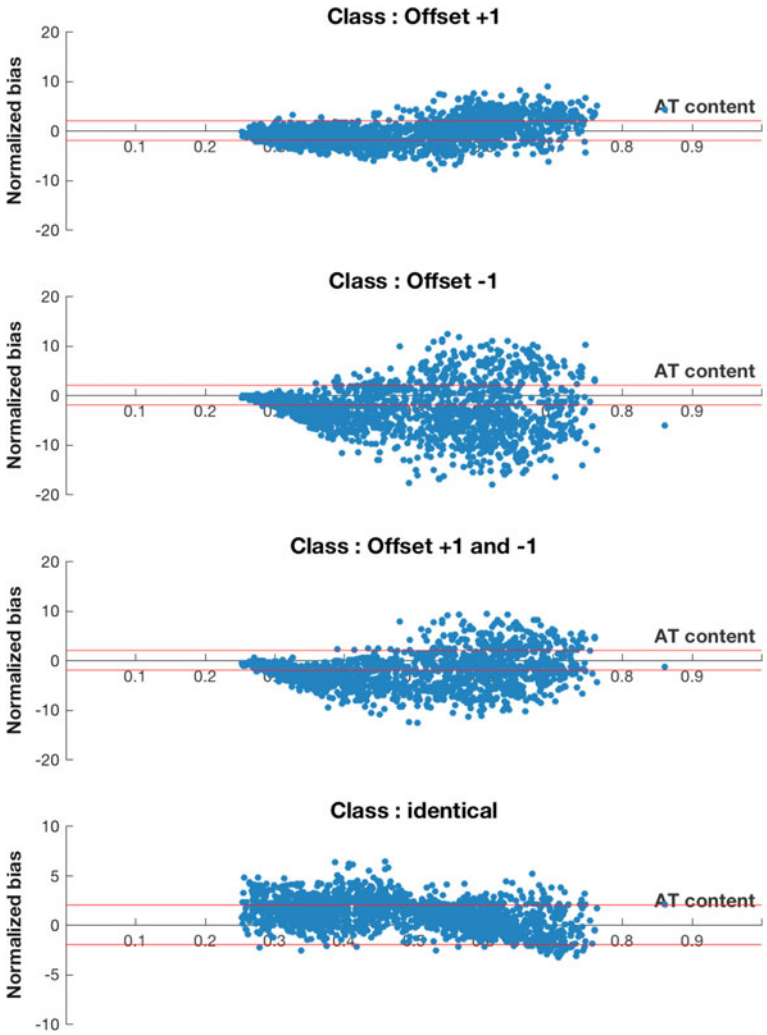


Fig. 14 Values of $\langle DCB_\alpha \rangle_{\text{norm}}$ for the 4 classes mentioned in the text versus AT content of genomes. Each point represents one bacterial genome. Again *red horizontal lines* are given by $\langle DCB_\alpha \rangle_{\text{norm}} = \pm 2$

for the coding of 2 successive identical amino acids. As the presence of correlations favoring identical successive codons was expected from literature (Shao et al. 2012), see Sect. 5, this finding shows that our approach is able to detect relevant statistical signals;

3. We also evidence an overall negative bias for the presence of dicodons containing an OSC (estimator $\langle DCB_{\pm 1} \rangle$ associated to Class ± 1). This result strongly suggest that the Ambush Hypothesis does not hold, at least for the bacterial genomes studied here;

4. This overall trend can be attributed mainly to Class -1 dicodons which present an overall negative bias, whereas Class $+1$ dicodons present an overall null bias.

4.3 Distribution of Distances to Off-Frame Stop Codons in the Random Codon Model

We analyze here whether the Ambush Hypothesis is actually necessary to prevent translation of long abnormal protein chains resulting from frameshift. In this regard, we compute the distance to the first encountered off-frame STOP after a frameshift to $+1$ or -1, starting from definiteness from the start AUG codon in the random codon model (RCM). In practice, we compute the codon usage from the genome of a given species, and draw random codons from this distribution, omitting any correlation between codons. This model therefore generates sequences of random codons. We then estimate the probabilities $Q(\ell)$ that this sequence, in frames $+1$ and -1, produces a STOP codon. To compute the distributions $Q_{+1}(\ell)$ and $Q_{-1}(\ell)$, we have to sum over sequences with ℓ off-frame codons ending up in one of the 3 possible STOPS. The summation over the exponential–in–ℓ number of compatible sequences can be easily carried out with the transfer-matrix formalism shown in Sect. 2.5. We do not report details here; note however that, as STOP codons are defined from 3 nucleotides only, the effective interaction between codons is short-range: only nearest neighbor codons interact along the sequence.

We show in Fig. 15 the outcome of this calculation for one specific bacterial species, *Thermodesulfobium-narugense*. Apart from differences at small ℓ reflecting the influence of the start codon (after the frameshift), both distribution apparently decay exponentially with ℓ. Actually the decay is not a pure exponential, as the transfer matrix is of dimension 4×4, and the number of exponentials is generically given by the size of the transfer matrix, minus one. We obtain that the average distance to the first OSC is about 8–9 in both frames. Hence, even *without any optimization* over the correlations between successive codons along the CDS, OSCs are very quickly found after a frameshift. This result raises doubts about the necessity of selecting codons to make the distance even smaller, as postulated by the Ambush Hypothesis.

5 Discussion and Perspectives

5.1 Nucleotide Motif Usage and Selective Pressures

Viruses have a rapid evolutionary rate, relatively small genomes, and, in many cases, databases of both genomic and phenotypic data that one can use to test

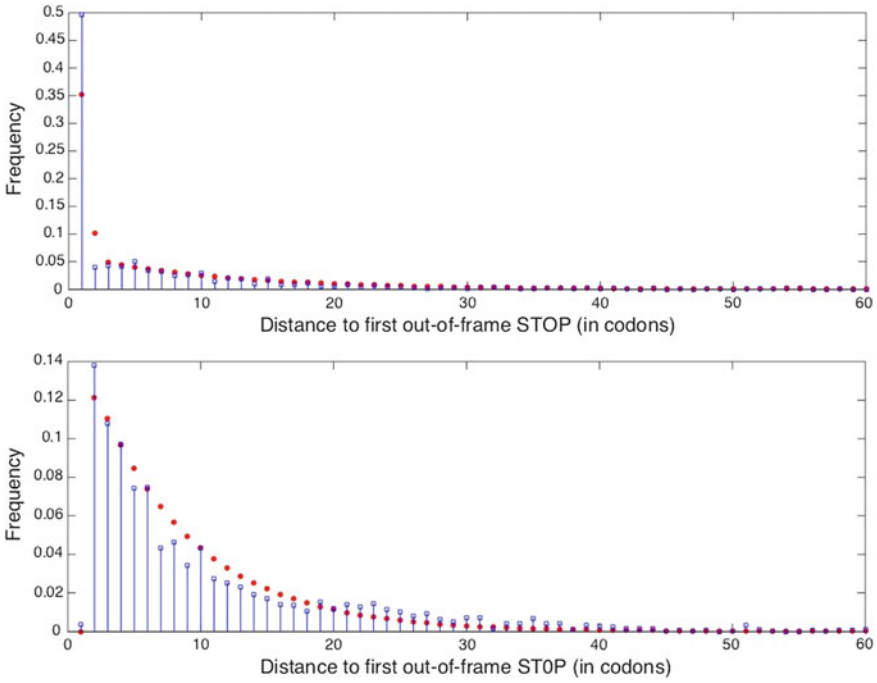


Fig. 15 Distributions of distances to first out-of-frame STOP codon after the start AUG codon and a frameshift equal to $+1$ (top panel) and -1 (bottom panel), measured in codons. *Blue* impulses and *squares* show the experimental distributions computed from all CDS of *Thermodesulfobium-narugense-DSM-14796*. *Red full circles* show the predictions from the random codon model (RCM), obtained with the transfer-matrix formalism, with codon usage estimated from the CDS of the same species (in frame). The average distances are: $\ell_{+1} \simeq 7.9$ and $\ell_{-1} \simeq 9.0$ codons

theoretical approaches. In this work we introduce a mathematical framework, inspired by an analogy with statistical physics, for a class of problems related to the evolution of viruses. The notions of entropy and pressure (or force) evoke the classical concepts of mutation-selection balance in population genetics. A major advantage of our approach is that these notions can be made quantitatively precise, with a very limited computational effort (scaling linearly with the sequence length). This approach is quite versatile, and could be extended to other evolutionary problems. Note that, while we have concentrated here on short nucleotidic motifs, our formalism can be extended to deal with longer motifs. If the motif contains from 2 to 4 nucleotides the transfer matrix M is given by Eqs. (25, 26). There are 63×63 possible matrices, which can be calculated once for all prior to the calculation of $Z(x)$ for several values of x. If the motif contains from 5 to 7 nucleotides the matrix M is $M(C_1, C_2, C_3)$ is "tridimensional", and there are 63^3 possible matrices. The vectors v_i are now functions of two codons. The calculation is slightly more complicated but can be done anyway.

While we have shown applications mainly to Influenza and HIV, many other viruses could be studied. An example is provided by Dengue virus, which goes back and forth between humans and insects. The time scales involved its evolution and the possible presence of mixed pressures acting on different motifs would be worth being studied.

A potentially interesting issue is whether the presence of pressures limits the accessibility of sequences through random mutations in the sequence space. In the absence of pressure codons are independent in our model, and may rapidly evolve under single nucleotide mutations. Hence, any possible sequence can be easily reached from another sequence. When a pressure acting on one motif is considered neighboring codons along the chain start to interact, as the motif may cover two or more contiguous codons, depending on its length. The resulting model is therefore a particular case of the short-range one-dimensional Potts model (Wu 1982), which is known in statistical physics to quickly thermalize. Therefore, as in the independent codon case, the sequence space is sampled efficiently by local moves (such as point mutations). We have checked this statement by running Monte Carlo simulations, and have verified that the relaxation times to the average values of various quantities, such as similarity between sequences and number of motifs, are independent of the value of the pressure. It is however possible that multiple pressures may lead to more complex sequence space structures, less efficiently sampled by local moves. Further studies of this point would be interesting to characterize how much pressures dynamically constrain the evolution of the virus sequence.

Another important application of our formalism is the case of non-coding sequences. In a related work (Tanne et al. 2015) we have extended our approach to non coding RNA, overexpressed in cancer cells compared to healthy tissues. Our analysis has allowed us to show that those overexpressed sequences, such as GSAT and HSATII, correspond to abnormal values of the forces acting on CpG and UpA motifs, and are likely to trigger a large auto-immune response. This prediction was confirmed experimentally, both in human and murine cells (Tanne et al. 2015).

5.2 Ambush Hypothesis

In the present work, we have analyzed Coding DNA Sequence (CDS) regions in all bacterial genomes to better investigate the validity of the so-called Ambush Hypothesis. We have introduced a statistical indicator in which the deviations in the probability of having a stop codon one or two nucleotides (1nt or 2nt shift) out of frame are calculated with respect to the probability based on the dicodon frequencies at fixed codon bias and fixed di-amino acids frequencies. With this unbiased indicator we found no systematic deviation across bacterial genomes favoring out of frame stop codons. On the contrary some significant statistical deviations are found for 2nt shifts, in which the probability of out frame stop is smaller than what expected in random sequences.

Our study has focused on four specific classes of di-codons. We will first discuss our result concerning Class *identical*, consisting of pairs of identical codons. Though the effect may seem weak, there is little doubt that there is a slight positive bias $\langle DCB_{\mathrm{id}} \rangle_{\mathrm{norm}}$ which means that translation of a pair of successive identical amino acids slightly favors the use of identical successive codons. This observation can be related to previously reported importance of synonymous codon ordering in yeast (Cannarozzi et al. 2010) and in bacteria (Shao et al. 2012); furthermore a recent study of archaeal aminoacyl-tRNA synthetases (aaRS) has shown that there was evidence for interactions between aaRS and the ribosome thus allowing to recycle tRNAs (Godinic-Mikulcic et al. 2014). Altogether these observations support a mechanism in which, due to colocalization of some aminoacyl-tRNA synthetases and ribosomes, in case of translation of 2 identical successive codons the ribosome, once the first codon translated, may use the same aaRS to translate the next codon.

Concerning our results for the 3 other classes (Class $+1$, Class -1, Class ± 1) one may first observe that the net result for Class ± 1 is at odds with previous results which may have seemed to support the Ambush Hypothesis, though this support was already questioned (Morgens et al. 2013). Indeed the overall negative values of $\langle DCB_{\pm 1} \rangle_{\mathrm{norm}}$ show that presence of dicodons containing an OSC is rather disadvantaged; furthermore comparison of $\langle DCB_{+1} \rangle_{\mathrm{norm}}$ and $\langle DCB_{-1} \rangle_{\mathrm{norm}}$ shows that these overall negative values can be mainly attributed to Class -1 dicodons, Class $+1$ dicodons exhibiting no specific trend in term of signed bias.

One may get further insight into our results examining Figs. 16 and 17. Figure 16 clearly shows the overall negative trend for $\langle DCB_{-1} \rangle_{\mathrm{norm}}$ and also shows that there is no obvious grouping of the genomes as characterized by their values of $\langle DCB_{+1} \rangle_{\mathrm{norm}}$ and $\langle DCB_{-1} \rangle_{\mathrm{norm}}$. Such an observation prompts to examine our results taking into account the phylogeny of our database which has been performed

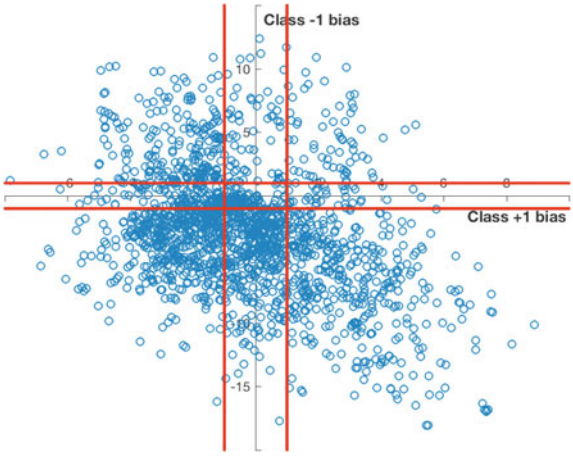


Fig. 16 Two-dimensional plot of values of $\langle DCB_{+1} \rangle_{\mathrm{norm}}$ and $\langle DCB_{-1} \rangle_{\mathrm{norm}}$ for the 1852 studied genomes. Again *red horizontal* and *vertical lines* are given by $\langle DCB_{\alpha} \rangle_{\mathrm{norm}} = \pm 2$ and define regions of statistical significance as explained in the text

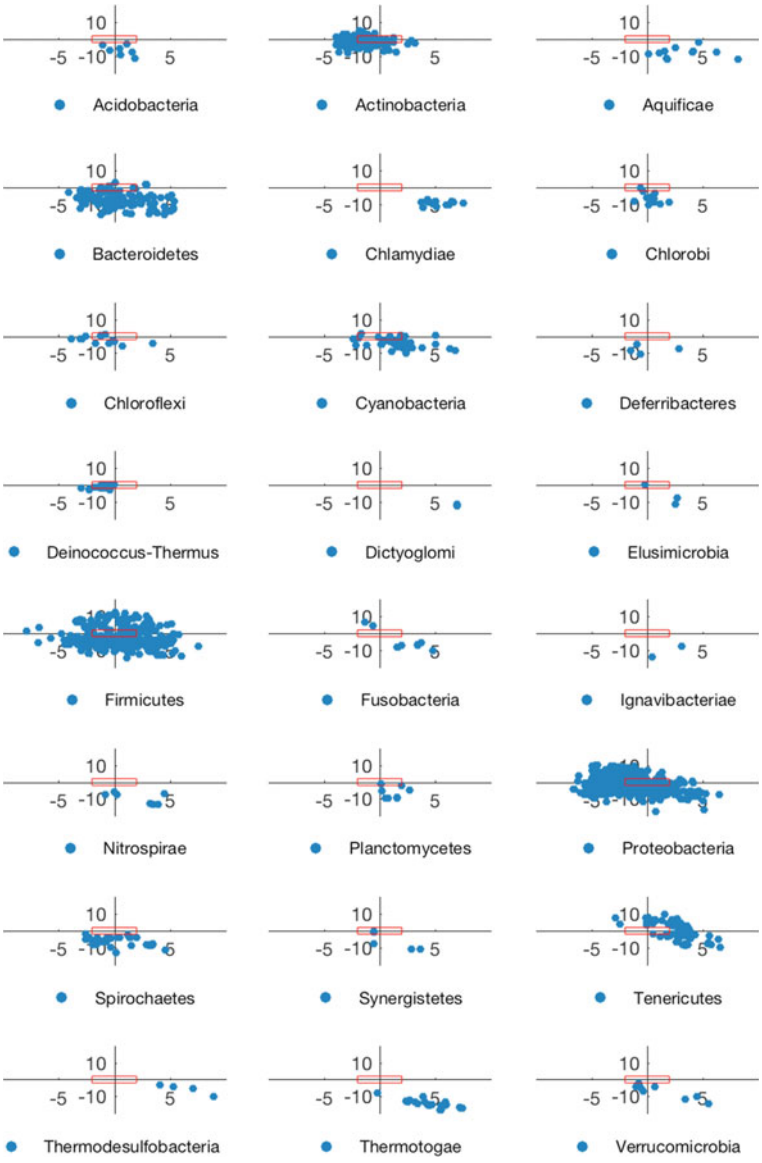


Fig. 17 Two-dimensional plot of values of $\langle DCB_{+1}\rangle_{\text{norm}}$ and $\langle DCB_{-1}\rangle_{\text{norm}}$ for the 1852 studied genomes grouped by phylum. *Red boxes* define regions of statistical significance as explained in the text

in Fig. 17. Indeed, Fig. 17 clearly shows that most phyla exhibit a negative value of $\langle DCB_{-1}\rangle_{\text{norm}}$ with the notable exception of the phyla *Actinobacteria*, *Firmicutes*, *Proteobacteria* and *Tenericutes*.

Quite obviously our results deserve further future analysis. Indeed, at this stage we can reject the Ambush Hypothesis as a general rule for prokaryotic genomes; nevertheless, refining the analysis as shown in Fig. 17, one reaches the conclusion that the situation is somehow more complex and specific phyla would deserve more detailed analysis (see the data in Fig. 17 concerning *Firmicutes* which show that within the same phylus one observes classes of opposite signs for $\langle DCB_{-1} \rangle_{\text{norm}}$). Furthermore, at the present level of analysis, we did not take into account the status of each OSC (TAA, TAG and TGA) which would also deserve more detailed analysis as previously suggested (Morgens et al. 2013); indeed such analysis is probably needed if, as in the case of the observed positive values of $\langle DCB_{\text{id}} \rangle_{\text{norm}}$, one wishes to give a meaningful interpretation in terms of biological processes to the measured values of the various $\langle DCB_{\alpha} \rangle_{\text{norm}}$.

Acknowledgements We are grateful to A. Levine for many enlightening discussions. This work was partly funded by the ANR Coevstat project (ANR-13-BS04-0012-01).

References

Cannarozzi G, Schraudolph NN, Faty M, von Rohr P, Friberg MT, Roth AC, Gonnet P, Gonnet G, Barral Y (2010) A role for codon order in translation dynamics. Cell 141:355–367

Coleman JR, Papamichail D, Skiena S, Futcher B, Wimmer E, Mueller S (2008) Virus attenuation by genome-scale changes in codon pair bias. Science 320(5884):1784–1787

Godinic-Mikulcic V, Jaric J, Greber BJ, Franke V, Hodnik V, Anderluh G, Ban N, Weygand-Durasevic I (2014) Archaeal aminoacyl-trna synthetases interact with the ribosome to recycle trnas. Nucleic Acids Res 42(8):5191

Greenbaum BD, Cocco S, Levine AJ, Monasson R (2014) Quantitative theory of entropic forces acting on constrained nucleotide sequences applied to viruses. Proc Natl Acad Sci 111(13): 5054–5059

Greenbaum BD, Levine AJ, Bhanot G, Rabadan R (2008) Patterns of evolution and host gene mimicry in influenza and other rna viruses. PLoS Pathog 4(6):e1000079

Hemmi H, Takeuchi O, Kawai T, Kaisho T, Sato S, Sanjo H, Matsumoto M, Hoshino K, Wagner H, Takeda K et al (2000) A toll-like receptor recognizes bacterial dna. Nature 408(6813):740–745

Jaynes ET (1957) Information theory and statistical mechanics. Phys Rev 106(4):620

Jimenez-Baranda S, Greenbaum B, Manches O, Handler J, Rabadán R, Levine A, Bhardwaj N (2011) Oligonucleotide motifs that disappear during the evolution of influenza in humans increase ifn-α secretion by plasmacytoid dendritic cells. J Virol

Li W-H, Wu C-I, Luo C-C (1985) A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. Mol Biol Evol 2(2):150–174

Long M, De Souza SJ, Rosenberg C, Gilbert W (1998) Proc Natl Acad Sci USA 95(1):219–223

Medzhitov R, Janeway C Jr (2000) Innate immunity. N Engl J Med 343(5):338–344

Morgens DW, Chang CH, Cavalcanti ARO (2013) Ambushing the ambush hypothesis: predicting and evaluating off-frame codon frequencies in prokaryotic genomes. BMC Genomics 14(1): 1–8

Nei M, Gojobori T (1986) Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol Biol Evol 3(5):418–426

Onsager L (1944) Crystal statistics I: two dimensional model with an order disorder transition. Phys Rev 65:117
Plotkin JB, Kudla G (2011) Synonymous but not the same: the causes and consequences of codon bias. Nat Rev Genet 12(1):32–42
Seligmann H (2010) The ambush hypothesis at the whole-organism level: off frame, 'hidden' stops in vertebrate mitochondrial genes increase developmental stability. Comput Biol Chem 34(2): 80–85
Seligmann H, Pollock DD (2004) The ambush hypothesis: hidden stop codons prevent off-frame gene reading. DNA Cell Biol 23(10):701–705
Shao Z-Q, Zhang Y-M, Feng X-Y, Wang B, Chen J-Q (2012) Synonymous codon ordering: a subtle but prevalent strategy of bacteria to improve translational efficiency. PLoS One 7(3): e33547
Sharp PM, Li W-H (1987) The codon adaptation index-a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res 15(3):1281–1295
Tanne A, Muniz LR, Puzio-Kuter A, Leonova KI, Gudkov AV, Ting DT, Monasson R, Cocco S, Levine AJ, Bhardwaj N et al (2015) Distinguishing the immunostimulatory properties of noncoding rnas expressed in cancer cells. Proc Natl Acad Sci 112(49):15154–15159
Tse H, Cai JJ, Tsoi HW, Lam EP, Yuen KY (2010) Natural selection retains overrepresented out-of-frame stop codons against frameshift peptides in prokaryotes. BMC Genomics 11(1): 491
Vabret N, Bhardwaj N, Greenbaum BD (2016) Sequence-specific sensing of nucleic acids. Trends Immunol 38(1):53–65
Wong TY, Fernandes S, Sankhon N, Leong PP, Kuo J, Liu JK (2008) Role of premature stop codons in bacterial evolution. J Bacteriol 190 (20):6718–6725
Wu FY (1982) The potts model. Rev Mod Phys 54(1):235–268

Case Studies of Seven Gene Families with Unusual High Retention Rate Since the Vertebrate and Teleost Whole-Genome Duplications

Frédéric G. Brunet, Thibault Lorin, Laure Bernard, Zofia Haftek-Terreau, Delphine Galiana, Manfred Schartl and Jean-Nicolas Volff

Abstract In the course of their evolution, the genomes of vertebrates have been subject to several events of whole-genome duplication (WGD): two rounds at the base of the vertebrates, a third at the base of the teleostean fish, and a few others in specific lineages like the salmonid fish or the *Xenopus* frogs among amphibians. Among the genes that were kept as duplicates long after the rediploidization process occurred, those that are involved in development, cell, and tissue diversification have the highest retention rates. In these categories fall gene families of different sizes that we previously investigated, namely genes that are implicated in the extracellular matrix formation that are the lectican, hapln, and adamts, or inducing a transcription or a cellular cascades: the transcription factors sox gene family, the nuclear receptors, the melanocortin receptors, cytoplasmic tyrosine kinases, and receptor tyrosine kinases. Here, we present an update of the expansion of these gene families due to the major vertebrate WGDs operated. Since the first occurrence of these WGDs, only six events of single gene duplication are reported, one being teleost fish specific and the second being *sry* in therians. This contrasts with the 71 ancestral genes that expanded in the jawed vertebrates to a total of 192 extant genes,

F.G. Brunet (✉) · T. Lorin · L. Bernard · Z. Haftek-Terreau · D. Galiana · J.-N. Volff
Univ Lyon, Institut de Genomique Fonctionnelle de Lyon, CNRS UMR 5242,
Ecole Normale Superieure de Lyon, Universite Claude Bernard Lyon 1,
46 allée d'Italie, 69364 Lyon, France
e-mail: frederic.brunet@ens-lyon.fr

M. Schartl
Physiologische Chemie, Biozentrum, University of Würzburg, Am Hubland,
97074 Würzburg, Germany

M. Schartl
Comprehensive Cancer Center, University Clinic Würzburg, Josef Schneider Straße 6,
97074 Würzburg, Germany

M. Schartl
Texas Institute for Advanced Study and Department of Biology,
Texas A&M University, College Station, USA

P. Pontarotti (ed.), *Evolutionary Biology: Self/Nonself Evolution, Species and Complex Traits Evolution, Methods and Concepts*,
DOI 10.1007/978-3-319-61569-1_19

and 75 additional genes in the teleost fish. We also discuss the ambiguities observed with the usual procedures of calculating gene retention rates. Interestingly, we observe that some of the gene families expanded in an intriguing regular manner. Those are also the ones that show very restricted lineage-specific losses. This feature is compatible and thus supportive of the autosomal-dominant deleterious mutation hypothesis that hampers the loss of the duplicated genes and explains their retention. Overall, the WGDs are shown to have amply participated in expansion of the gene families studied here, hence providing new potential for gene diversification, and by extension to the enhancement of molecular and cellular components of vertebrates.

Acronyms

WGD Whole-genome duplication
MYA Million years ago
MYR Million years

1 Introduction

1.1 Whole-Genome Duplications as a Major Evolutionary Force

Genome plasticity has been observed on Drosophila polytene chromosomes, revealing losses and gains of large chunks of DNA that are the deletions, insertions, or duplications, or conservative events such as fissions, fusions, or translocations. More than a century ago, tetraploidization has been reported in the maize (*Zea mays*) (Kuwada 1911), and several cases of duplication events were reported (see review in Taylor and Raes 2004) before Susumo Ohno published his milestone book that has remained up to now a reference regarding gene duplications (Ohno 1970). The numerous completed genomes of eukaryotes being sequenced allow the analysis of the variation in chromosome structure and number of genes and reveal many past events of duplications, ranging from small-scale duplications (SSD) of DNA segments to whole-genome duplications (WGDs), or polyploidy. SSDs are generated by several mechanisms: for example, tandem duplications via unequal crossing-over; segmental duplications vectorized by transposable elements (Yang et al. 2008); or retrogenes, some of them being chimeric like the *jingwei* gene (Wang et al. 2000; Zhang et al. 2004). New gain of function can also come from retroposon domestication (Volff 2006, 2009).

Being either autopolyploids (intraspecific WGDs) or allopolyploids (genome doubling involving interspecific hybridization), these events lead to individuals with identical or similar duplicated chromosome sets, respectively. After each tetraploidization event, genomes are returning to a diploid state. During this

process, the major fate of the duplicated genes, whether it is of SSD or WGD origin, is loss (Li 1983; Maere et al. 2005) by pseudogenization, keeping occasionally another loose functional role, and/or deletion of large chunks of chromosome. In this case, illegitimate or unequal intrastrand homologous recombinations occur during the rediploidization process, often involving repetitive sequences (simple repeats and transposable elements, being DNA transposons and retrotransposons) (Devos et al. 2002). Other duplicates can remain and the two co-orthologs are maintained due to various mechanisms that range, in short, from subfunctionalization to neofunctionalization. During subfunctionalization, previous functions of the ancestral gene are split between the two duplicates, whereas in neofunctionalization, one gene keeps the original function and the duplicate is free to evolve toward a new function (Force et al. 1999; Hughes et al. 2007; Conant and Wolfe 2008; Gout and Lynch 2015).

1.2 Recurrence of Whole-Genome Duplications in the Course of Vertebrate Evolution

WGDs have shaped the genomes of many taxonomic groups. In plants, many forms are described as being triploid, tetraploid, or with even higher ploidy levels like the octoploid cultivated strawberry (Lee et al. 2012; Vanneste et al. 2014; Solti et al. 2015). In other organisms, fewer events of WGDs occurred. It has been amply described in the yeast lineage (Wolfe and Shields 1997; Kellis et al. 2004; Fischer et al. 2006), in the Paramecium species (Aury et al. 2006; McGrath et al. 2014a), and recently in the horseshoe crabs (Kenny et al. 2016). At the base of the vertebrates, two WDG events have been described, called 1R and 2R for first and second rounds of WGDs (Ohno 1999; Li et al. 2001; McLysaght et al. 2002; Panopoulou et al. 2003; Dehal and Boore 2005; Holland et al. 2008; Smith and Keinath 2015; Panopoulou and Poustka 2017). The first event has probably occurred before the divergence of the jawed vertebrates. The second WGD most likely happened after the emergence of the agnathans (lampreys and hagfish), although other hypotheses have been proposed including loss of duplication or independent WGD in these lineages (Smith et al. 2013) and see discussion in (Panopoulou and Poustka 2017). This second event certainly occurred in a far ancestor of the cartilaginous fish (Ravi et al. 2009). The time period between these two WGDs is estimated to be between 30 myr (Inoue et al. 2015) to 100 myr apart (Wang and Gu 2000). Genome sequencing of more species of lampreys and hagfishes should help to solve those questions. A third WGD has occurred at the base of the teleosts (Ts3R-WGD, for teleost-specific) (Meyer and Schartl 1999; Taylor et al. 2003; Jaillon et al. 2004; Naruse et al. 2004; Woods et al. 2005; Brunet et al. 2006; Kasahara et al. 2007; Nakatani et al. 2007; Kassahn et al. 2009). With the recent sequencing of the rainbow trout (*Onchorynchus mykiss*) and the Atlantic salmon (*Salmo salar*), a fourth one (Ss4R-WGD, salmonid-specific) was described on the whole-genome

level for the Salmoniformes (Lien et al. 2016; Berthelot et al. 2014; Macqueen and Johnston 2014). Another 4R-WGD has been described in the common carp (Li et al. 2015). In vertebrates, a 3R-WGD was described for the allotetraploïd genome of the African claw frog, *Xenopus laevis* (Hugues et al. 1993; Uno et al. 2013; Session et al. 2016), and polyploidy is recurrent in amphibians (Evans et al. 2004; Schmid et al. 2015). Multiple cases of polyploidy are described in the sturgeon lineage (Rajkov et al. 2014), and it has been suggested that the genome of the lampreys could have experienced some WGDs independent from that of the other gnathostomes (Smith et al. 2013). Paralogous genes as footprints of WGDs on homologous chromosomes have been coined as "ohnologs" (Wolfe 2000; Turunen et al. 2009) in reference to Susumu Ohno and his seminal work on the genome duplication hypothesis (Ohno 1970, 1999). Those ohnologous genes are part of an orthogroup where each of them is orthologous to the gene in the closest lineages in which the WGD considered did not occur.

1.3 Gene Categories Retained After WGDs and How to Visualize Them

Gene ontology searches of duplicated ohnologs retained preferentially after WGDs showed a higher proportion of those involved in developmentally regulated signaling processes, DNA-binding proteins, transcription regulation, signal transduction, or cell cycle regulation pathways, either in vertebrates, including fish (Bertrand et al. 2004; Blomme et al. 2006; Brunet et al. 2006; Steinke et al. 2006; Hufton et al. 2008; Kassahn et al. 2009; Sato et al. 2009), yeast (Davis and Petrov 2005), or plants (Blanc and Wolfe 2004; Seoighe and Gehring 2004; Maere et al. 2005; Paterson et al. 2006). During the past decade, we have been interested in the evolution of several gene families among these gene categories.

Here, we present an up-to-date version of our previous analyses allowing us to visualize how the first three rounds of WGDs along the various vertebrate lineages have impacted the different members of multigene families. These genes are either involved in the extracellular matrix formation: lectican and hapln (Brunet et al. 2012), adamts (Brunet et al. 2015); transcription factors: sox (Voldoire et al. 2017) and nuclear receptors (Bertrand et al. 2004; Schubert et al. 2008); or trigger cellular cascade responses: the melanocortin receptors (Volff et al. 2013), and receptor and cytoplasmic tyrosine kinases (Schartl et al. 2015; Brunet et al. 2016). To this end, we analyzed a total of 192 genes in these families among vertebrate genomes that are available in the most recent Ensembl release 87—Dec 2016 (Bronwen et al. 2016). We focused our attention on whether or not these genes were present in major lineages that include: primates, other placental mammals, marsupials and/or the prototherian *Ornithorynchus anatinus*, birds, non-avian sauropsids, the amphibian *Xenopus tropicalis*, coelacanth, in fish the non-teleost Actinopterygii (the gar, *Lepisosteus oculatus*), Ostariophysi (zebrafish and cave fish) and

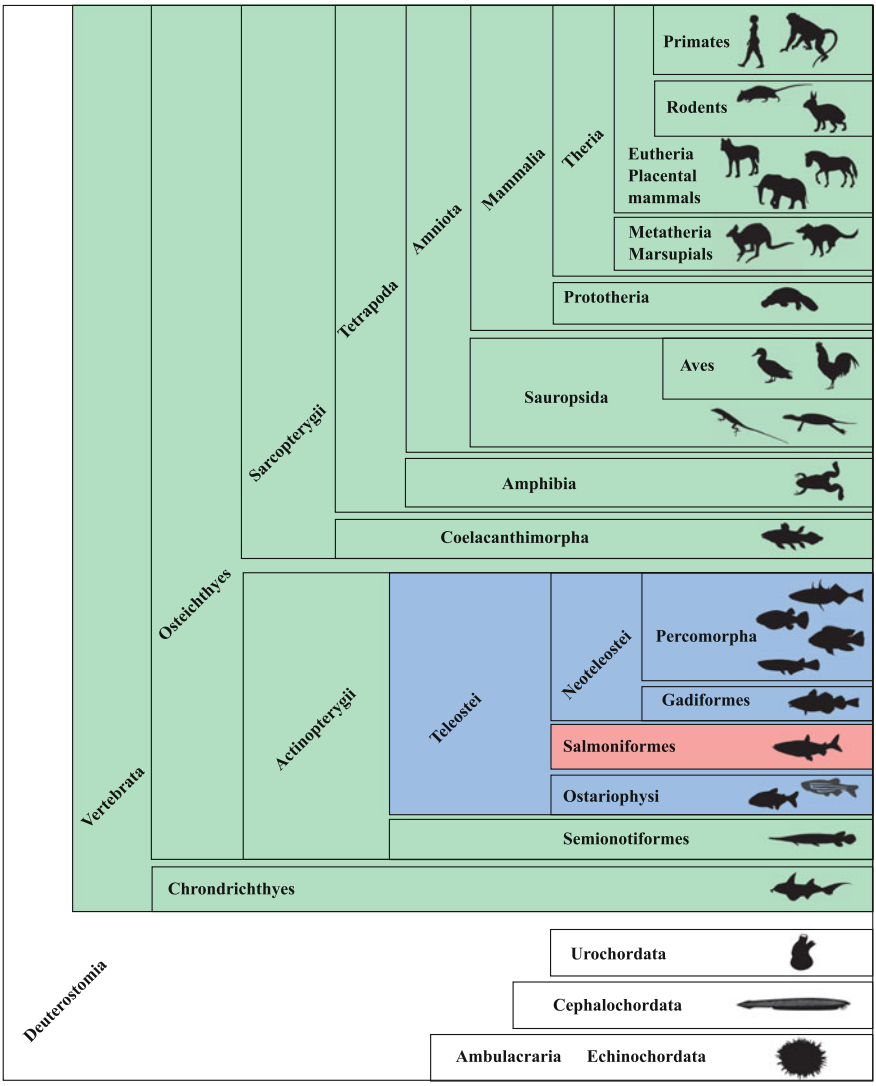


Fig. 1 Classification of the species showing the different taxonomic names used in this article. The *green* background shows the vertebrate species that went through the 1R/2R WGDs; *blue* background shows the teleostean species that went through the Ts3R WGD; *red* background shows the Salmoniformes species that went through the Ss4R WGD

Neoteleostei (Fig. 1). In addition, we double-checked the origin of the duplicated genes from the teleost-specific 3R-WGD using both the Genomicus synteny viewer (Louis et al. 2012) (Genomicus, database version: 86.01) and a tool (see Fig. 2) that we designed to facilitate the visualization of the ohnologous synteny information

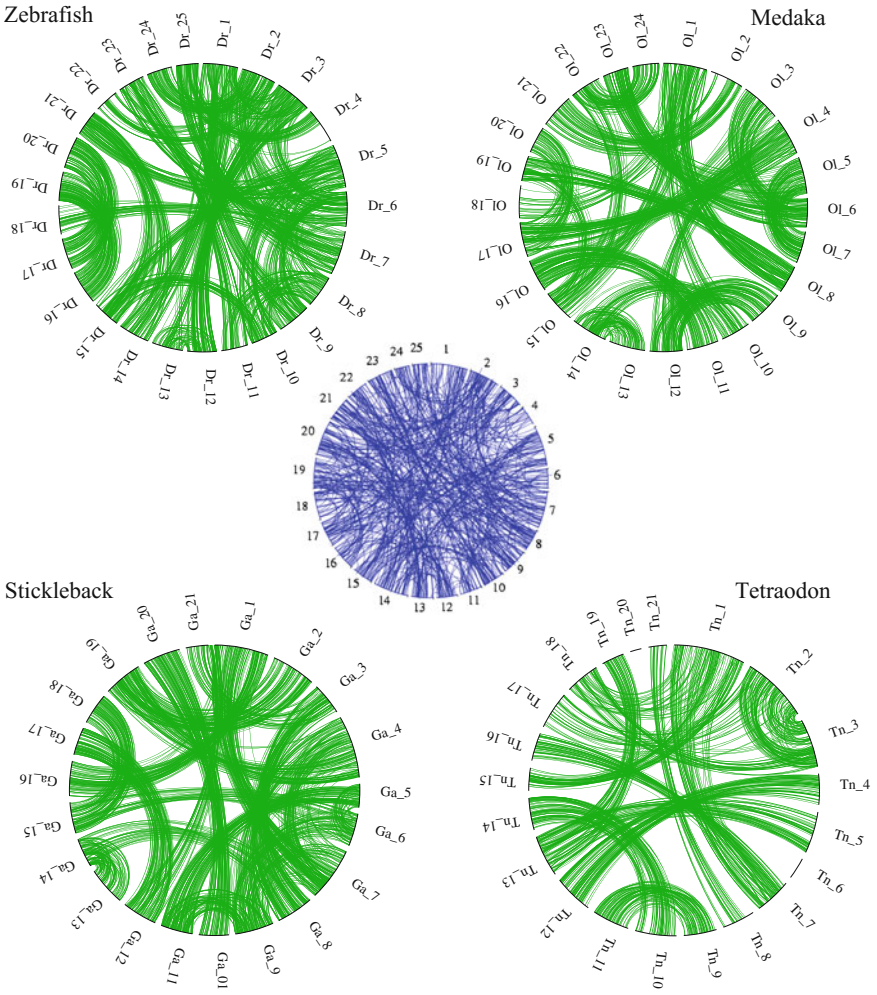
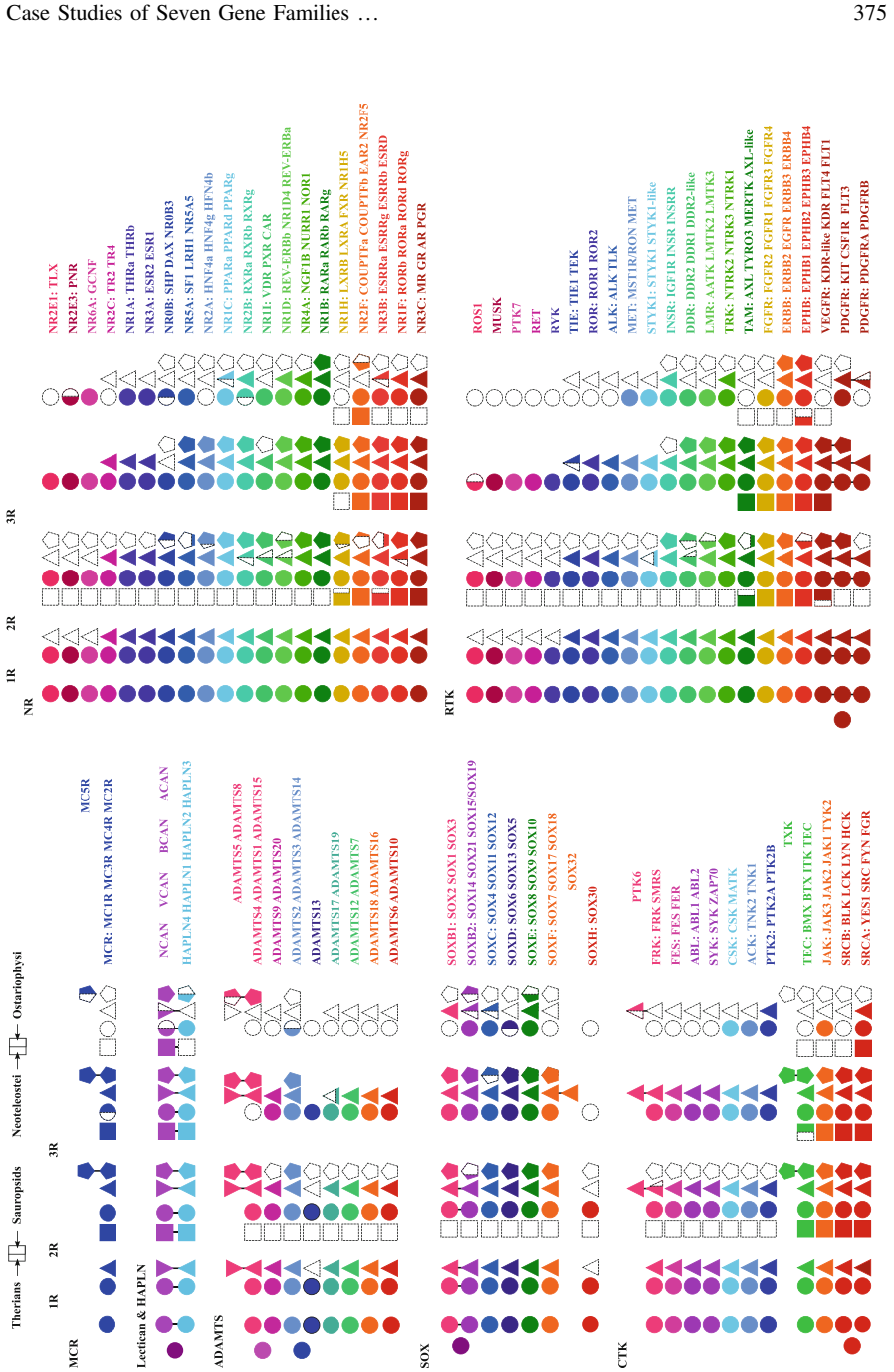


Fig. 2 Circular representation of the Ts3R-WGD of 4 fish species, the zebra fish *Danio rerio* (version Zv8) (Cypriniformes, Ostariophysi) and three Percomorpha (Acanthomorpha) that are the *green* spotted puffer *Tetraodon nigroviridis* (v. TETRAODON8), the medaka *Oryzias latipes* (v. MEDAKA1), and the stickleback *Gasterosteus aculeatus* (v. BROADS1). Their chromosomes are positioned in *circle* (in *black*) and the ohnologs linked by *green* arcs. In the *middle* (in *blue*) is a representation of a random distribution of the paralogues. The non-random distribution of these links allowed the visualization of the ohnology remaining from the Ts3R for the first time in the tetraodon (Jaillon et al. 2004)

(paralogy of homologous chromosomes with WGD origin) and used systematically in our previous searches for ohnologous genes (e.g., Jaillon et al. 2004; Brunet et al. 2016).


Therians
Sauropsids
Neoteleostei
Ostariophysi
1R
2R
3R
MCR
MC5R
MCR: MC1R MC3R MC4R MC2R
Lectican & HAPLN
NCAN VCAN BCAN ACAN
HAPLN4 HAPLN1 HAPLN2 HAPLN3
ADAMTS
ADAMTS5 ADAMTS8
ADAMTS4 ADAMTS1 ADAMTS15
ADAMTS9 ADAMTS20
ADAMTS2 ADAMTS3 ADAMTS14
ADAMTS13
ADAMTS17 ADAMTS19
ADAMTS12 ADAMTS7
ADAMTS18 ADAMTS16
ADAMTS6 ADAMTS10
SOX
SOXB1: SOX2 SOX1 SOX3
SOXB2: SOX14 SOX21 SOX15/SOX19
SOXC: SOX4 SOX11 SOX12
SOXD: SOX6 SOX13 SOX5
SOXE: SOX8 SOX9 SOX10
SOXF: SOX7 SOX17 SOX18
SOX32
SOXH: SOX30
CTK
PTK6
FRK: FRK SMRS
FES: FES FER
ABL: ABL1 ABL2
SYK: SYK ZAP70
CSK: CSK MATK
ACK: TNK2 TNK1
PTK2: PTK2A PTK2B
TXK
TEC: BMX BTK ITK TEC
JAK: JAK3 JAK2 JAK1 TYK2
SRCB: BLK LCK LYN HCK
SRCA: YES1 SRC FYN FGR
NR
RTK
NR2E1: TLX
NR2E3: PNR
NR6A: GCNF
NR2C: TR2 TR4
NR1A: THRa THRb
NR3A: ESR2 ESR1
NR0B: SHP DAX NR0B3
NR5A: SF1 LRH1 NR5A5
NR2A: HNF4a HNF4g HFN4b
NR1C: PPARa PPARd PPARg
NR2B: RXRa RXRb RXRg
NR1I: VDR PXR CAR
NR1D: REV-ERBb NR1D4 REV-ERBa
NR4A: NGFIB NURR1 NOR1
NR1B: RARa RARb RARg
NR1H: LXRB LXRA FXR NR1H5
NR2F: COUPTFa COUPTFb EAR2 NR2F5
NR3B: ESRRa ESRRg ESRRb ESRRD
NR1F: RORb RORa RORd RORg
NR3C: MR GR AR PGR
ROS1
MUSK
PTK7
RET
RYK
TIE: TIE1 TEK
ROR: ROR1 ROR2
ALK: ALK TLK
MET: MST1R/RON MET
STYK1: STYK1 STYK1-like
INSR: IGF1R INSR INSRR
DDR: DDR2 DDR1 DDR2-like
LMR: AATK LMTK2 LMTK3
TRK: NTRK2 NTRK3 NTRK1
TAM: AXL TYRO3 MERTK AXL-like
FGFR: FGFR2 FGFR1 FGFR3 FGFR4
ERBB: ERBB2 EGFR ERBB3 ERBB4
EPHB: EPHB1 EPHB2 EPHB3 EPHB4
VEGFR: KDR-like KDR FLT4 FLT1
PDGFR: KIT CSF1R FLT3
PDGFR: PDGFRA PDGFRB

◂**Fig. 3** Gene distribution after the vertebrate 1R and 2R, and the Ts3R WGDs described in seven multigenic families harboring a high retention rate. The vertebrate genes amount here for about 1% of the estimated 19,000 coding sequences of the human genome. Filled motifs represent the occurrence of a gene within tetrapods (after the 2R) and fish (after the 3R). After the 1R and 2R WGDs, the motifs can be partially filled whenever a gene is found only among mammals (placentals and/or marsupials) on the *left* part, or only in sauropsids (lizards, turtles, birds) on the *right* part. The same distinction is made for genes that are only found in the Neoteleostei on the *left* or only found in the Ostariophysi on the *right*. Other distinctions are made, e.g., *adamts19* is lost in teleostean fish but found in the gar; *ddr1* is lost in birds; *kdr*-like is lost in placental mammals. Considered here but not named: *car* is named *crx* in birds; *fgr* is named *yrk* in fish and sauropsids

2 Results and Discussion

2.1 An Overview of the 7 Gene Families Under Investigation

Figure 3 illustrates the results of the 1R/2R and 3R WGDs on the seven families we previously scrutinized. The first two vertebrate WGDs generated 192 genes that constitute 71 subfamilies. The Ts3R-WGD added 75 more genes for the teleost fish genomes. The two first families are relatively small and are presented as simple case studies. The few SSD events that occurred either after the WGDs or right before are also shown.

2.1.1 The Melanocortin Receptor Family

The *melanocortin receptor* genes (MCRs) encode for G-protein-coupled receptors that are involved in key biological functions such as metabolic regulation and energy homeostasis. Five paralogs named mc1r to mc5r are observed among vertebrates in the gene family (Takahashi and Kawauchi 2006). Although no orthologue could be detected in non-vertebrate deuterostome species and beyond (Västermark and Schiöth 2011), this gene family presents the pattern of an ancestral gene duplicated with full retention through the 1R/2R WGDs, with an additional SSD leading to *mc2r* and *mc5r* genes.

2.1.2 The Lectican and Hapln Families

The hyalectan gene family commonly known as lectican family is part of the large proteoglycan genes encoding for some of the many macromolecules that are components of the extracellular matrix. There are four members of the lectican family: aggrecan (*Acan*), brevican (*Bcan*), neurocan (*Ncan*), and versican (*Vcan*). These proteins bind to hyaluronan proteins (HA) to form rigid proteoglycan structure of high resistance and strain which characterizes their function in tissues such as cartilage and tendon (Binder et al. 2017).

There are four hyaluronan and proteoglycan link proteins (Hapln) that are also part of the extracellular matrix. These hyaluronans bind to lecticans, and together, they make a significant contribution to cell migration and proliferation. These *lectican* and *hapln* genes originated from a head-to-head tandem SSD, and this syntenic arrangement has been retained after the 1R/2R: *ncan* with *hapln4*, *vcan* with *hapln1*, *bcan* with *hapln2,* and *acan* with *hapln3* (Spicer et al. 2003). This tight synteny is disrupted in only one of the eight pairs of Ts3R ohnologs (Brunet et al. 2012).

2.1.3 The Adamts Family

The A Disintegrin-Like and Metalloproteinsae Domain with Thrombospondin-1 Repeats (*adamts*) proteoglycanase family shares some common domains with the large ADAM family (A Disintegrin And Metalloproteinase). These are transmembrane and secreted metalloendopeptidases involved in major extracellular matrix remodeling events during proliferation, morphogenesis, and cell fusion (Brocker et al. 2009; McCulloch et al. 2009; Stupka et al. 2013). Both gene families are part of a larger superfamily comprising the snake venom metalloproteases (SVMPs) and the matrix metalloproteases (MMPs) (Seals and Courtneidge 2003). The hyalectan proteoglycans are regulated by the proteoglycanase activity of the Adamts. Alteration of this regulation could be proposed to play an important role in disruption of the extracellular matrix leading to cancer progression (Binder et al. 2017). There are 19 *adamts* genes in human, labeled from *Adamts1* to *Adamts20*, where *Adamts11* is the same gene as *Adamts5*. These genes are grouped in 8 subfamilies, and the encoded proteins share a common structure with domains being a propeptide, a metallopeptidase M12B, a disintegrin-like, and a thrombospondin (TSP) type-1 followed by a spacer. Additional domains characterize each family (Nicholson et al. 2005; Brunet et al. 2015; Kelwick et al. 2015). The evolutionary history of this family can be traced back to 600 mya with one copy in the sponge *Amphimedon queenslandica* (Srivastava et al. 2010) that gradually increased to 6 genes found in the Lophotrochozoans and Ecdysozoa (Brunet et al. 2015) and raised to 8 in the deuterostomes after two events of SSDs. The first one occurred in the ancestral genes giving the subfamilies adamts2, -3, -14 and -13, and the second is the ancestral gene of adamts9, -20 that duplicated by retroposition into the actual orthogroup of adamts4, -1, -5, -8, -15. This case is deduced from the traces of loss of the introns that characterized these genes and those newly acquired since then (Nicholson et al. 2005). The 1R/2R provided a sudden burst to the preexisting genes raising this number up to 19. This gene family is peculiar as the Ts3R-WGD did not lead to an increase on average of the total number of *adamts* genes among teleosts.

2.1.4 The Sox Family

Sox proteins are DNA-binding proteins of the high-mobility group (HMG) box superfamily. The HMG box superfamily is found already in the deep branches of the tree of life, as it appears in animals, yeasts, sponges, and plants. In contrast, *sox* genes are found only in deuterostomes and protostomes, as well as in non-bilaterian metazoans (e.g., jellyfish) where they play key roles in developmental processes such as germ layer formation, organ development, and cell-type specification (Laudet et al. 1993; Soullier et al. 1999; Le Gouar et al. 2004; Jager et al. 2006; Larroux et al. 2006; Jager et al. 2011; Heenan et al. 2016). Based on sequence similarity in the HMG domain, the 20 *sox* genes found in mammals are subdivided into seven groups (Heenan et al. 2016). A paradigmatic *sox* gene is the mammalian testis-determining gene *Sry* that induces testis differentiation from the bipotential gonad and subsequent male development (Foster et al. 1994). The **S**ry-related HMG b**ox** (Sox) is a highly conserved DNA-binding motif of about 70 amino acids (aa) onto which a strong selective pressure is acting due to biophysical interactions between the concave binding surface of these proteins matching over a large surface the minor groove of their target base-specific DNA sequences (Bianchi and Beltrame 1998; Wegner 1999).

The sox gene family is divided into 7 subgroups, from soxB to soxH, with soxB being subdivided into soxB1 and soxB2, soxA derived from soxB1 and soxG from soxF (see below). A tandem duplication occurring prior to the 1R/2R WGD duplications is still observed between *sox2* and *sox1* genes (soxB1) and s*ox14* and *sox21* (soxB2). Some loose links can still be observed among the ohnologous chromosomes of the zebra fish. All these soxB genes are intronless, indicating that their common ancestral gene may have originated from a former retroposition event. The soxB2 subfamily contains also two ohnologous Ts3R *sox19* genes in fish, but no direct ortholog is found in tetrapods. Inversely, *sox15*, initially assigned to the soxG group, is only present in mammals (not in marsupials, monotremes, or sauropsids). The region around human *Sox15* on chromosome Hs_17 harbors several syntenic links with the regions in which *sox19a* and *sox19b* reside, on chromosomes Dr_5 and Dr_7, respectively (Okuda et al. 2006; Voldoire et al. 2017). In addition, *sox15* genes have an intron that shares the exact same position as those of the *sox19* genes. These features suggest that their ancestral gene was acquired before the split of the Sarcopterygians and the Actinopterigyans and altogether that *sox15* belong, like *sox19* genes, to the soxB2 subfamily. Group A is made only of *sry*, the male sex-determining gene located on the Y chromosome of mammals and marsupials (but absent from the monotremes). This intronless gene, like all members of the soxB1 and soxB2 subfamilies, originated from *sox3* before the differentiation of the protosex chromosomes into X and Y (Lahn and Page 1999; Veyrunes et al. 2008; Sekido and Lovell-Badge 2009).

2.1.5 The Nuclear Receptor Family

The nuclear hormone receptors (NRs) are DNA-binding transcription factors involved in biological processes as wide as development, cellular processes or the control of the homeostasis (Germain et al. 2006). Most of them are ligand-binding receptor, but a few are not, thus called orphan receptors, like *Nurr1* (Wang et al. 2003). NRs are characterized by a common structure of 5 domains: the N-terminal and the C-terminal domains that surround the two main functional domains: the DNA-binding-domain (DBD), which is highly conserved due its function to bind to the uniform structure of the DNA double helix, and the ligand-binding domain (LBD). These two domains linked by the hinge domain. NRs generally bind to DNA either as homodimers or as heterodimers with one member of the RXR NR family. In the absence of a ligand, NRs act as transcriptional repressors in corepressor complexes. These complexes dissociate in the presence of their specific ligand with the recruitment of coactivators mediating the initiation of the transcription process of the neighboring genes they regulate (Germain et al. 2006; Markov and Laudet 2011). The NRs arose very early in the evolutionary history of the metazoan lineage, with two NRs being present in the sponge *Amphimedon queenslandica*. 33 NR genes are in the amphioxus genome and 48 in human (Bertrand et al. 2011), as reviewed in (Lecroisey et al. 2012), and even more in teleosts due to the Ts3R-WGD (Bertrand et al. 2004). In vertebrates, this large gene family is subdivided into 7 subfamilies, from NR0 to NR7. NR1 is thyroid hormone receptor-like; NR2, retinoid X receptor-like; NR3, estrogen receptor-like; NR4, nerve growth factor IB-like; NR5, steroidogenic factor-like; NR6, germ cell nuclear factor-like; and NR0, the miscellaneaous. Each of these subfamilies has another level of subdivision. For example, NR1A is the thyroid receptors; NR1B, the retinoic acid receptors; and NR1I, the vitamin D receptor-like. Here, we see the impact of the 1R/2R WGDs that led to a total of 56 NRs in non-teleost vertebrates and their 22 inferred duplicates in the teleostean fish after the Ts3R-WGD.

2.1.6 The Tyrosine Kinases Family

Protein kinases (PKs) are one of the largest superfamilies among eukaryotic proteins (Hanks 2003) with a total of 518 PKs found in human (Manning et al. 2002). These enzymes phosphorylate specific tyrosine, serine, or threonine residues in substrate proteins using the gamma phosphate of adenosine triphosphate (Lemmon and Schlessinger 2010). Those that specifically phosphorylate tyrosine residues are the protein tyrosine kinases (PTKs) that are subdivided into cytoplasmic non-receptor proteins (CTKs) and receptor tyrosine kinases (RTKs). CTKs relay intracellular signals, and RTKs are cell surface receptors that transduce extracellular signals to the cytoplasm by activating several downstream signaling cascades. PTKs have major roles in maintenance of homeostasis and development, growth, cellular differentiation, and apoptosis in multicellular organisms. Mutations in many

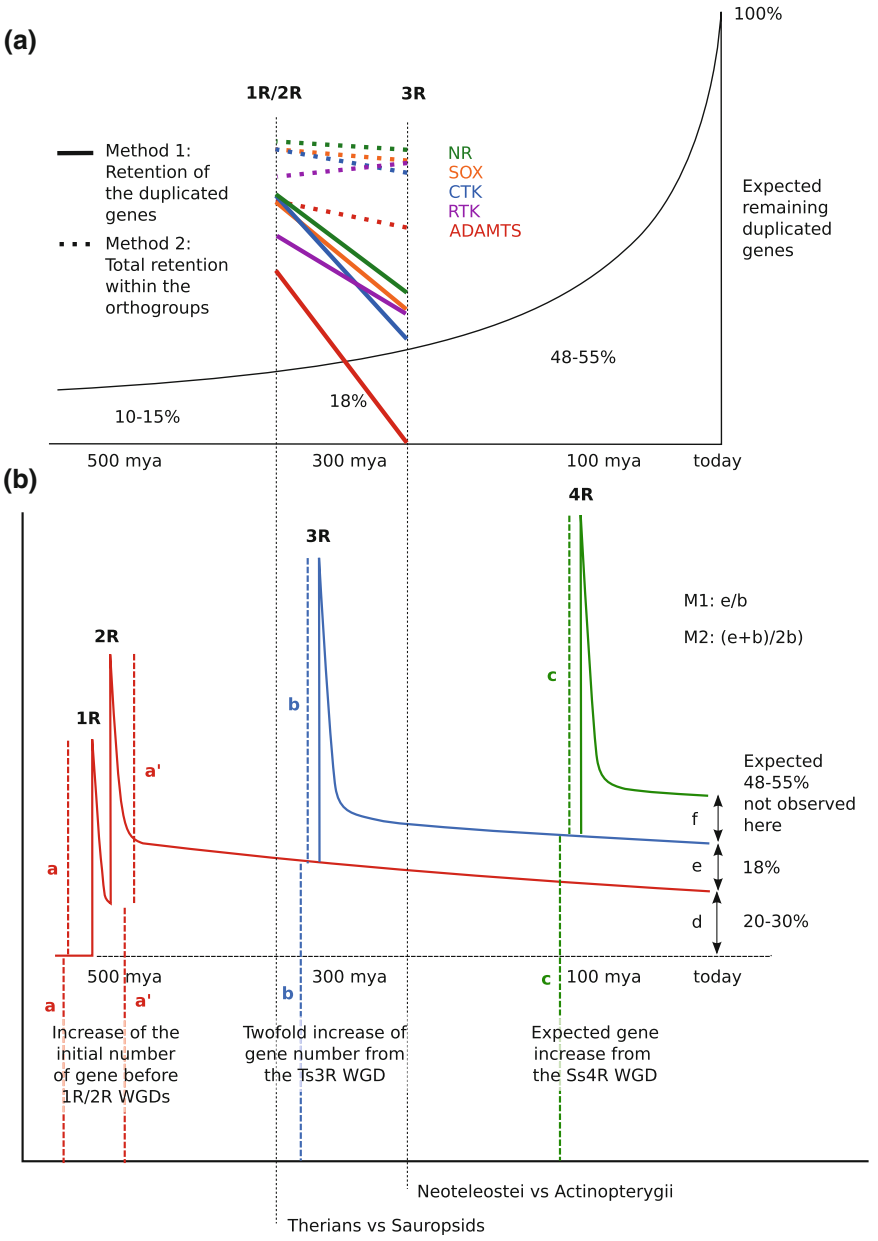
of these genes or their deregulation lead to a wide spectrum of pathologies including cancer (Zwick et al. 2001; Gschwind et al. 2004; Schlessinger 2014).

Up to 58 RTK and 32 CTK genes have been described in the human and mouse genomes (Manning et al. 2002; Suga et al. 2008; Robinson et al. 2000). They share an intracellular C-terminal tyrosine kinase domain (TKD), which is connected in the RTK proteins, to an alpha helical transmembrane domain and extracellular N-terminal part composed of a variety of different domains. RTKs are subdivided into 20 subfamilies and have an N-terminal part with modular structural domains in the outer cell surface for 18 of them. The Lmr (Lmtk/Aatk) and Styk1/Nok subfamilies are phylogenetically the most distantly related subfamily from all other PTKs. Structurally, they have only a short amino-terminus domain on the outer cellular surface attached to their transmembrane domain. Comparing the position and phase of the introns in the TK domain of these PTKs revealed that secondary losses of their external domains could have occurred for Styk1 subfamily that would have derived from Vegfr/Pdgfr/Ret/Fgfr/Tie subfamilies, and the (Lmtk/Aatk) subfamily that most likely derived from Alk/Ros1/Insr (Brunet et al. 2016). Each phylogenetic tree of the 11 CTK subfamilies observed in vertebrates is rooted by amphioxus and sea squirt. Src genes are subdivided into *srcA* and *srcB* and derived from an SSD in an ancestor of vertebrates (D'Aniello et al. 2008).

2.2 Gene Retention After Whole-Genome Duplications

The 1R/2R WGDs have impacted these gene families differently. The whole set of the expected four ohnologs have been retained in the small families of MCR, lectican, and hapln genes (without counting the SSD events). Larger gene families show very different evolutionary histories. A quite regular distribution is observed among the RTKs (orthogroups/subfamilies of 1/2/3/4 genes: 5/6/6/4) or the NRs (orthogroups/subfamilies of 1/2/3/4 genes: 3/3/9/5). The three other gene families of medium size exhibit some very interesting patterns. The sox family has only a single orthogroup of 1 gene, and the 6 other orthogroups have 3 orthologs each (1/0/6/0). Adamts family has one orthogroup of 1 gene, 5 orthogroups of 2 genes, and one orthogroup of 3 genes (1/5/1/0; plus the last orthogroup of 5 genes). The pattern of the CTK is also very interesting as there are only two categories: 7 orthogroups of 2 ohnologs and 4 orthogroups of 4 (0/7/0/4).

In the two large RTK and NR gene families, with a total of 108 genes, 19 lineage-specific losses have been reported, implying a constant loss process. In contrast, only two genes (*sox15* and *ptk6*) present some lineage-specific losses among the 84 other genes. These are as follows: *nr1h5/fxrB* in primates (Otte et al. 2003); *nr2a2/hnf4b, nr1f4/rord,* and *kdr*-like in eutherians; *nr0b3* and *ddr2*-like in therians; *adamts19* and *adamts20* in marsupials; *nr1h2/lxrb, nr3b1/esrr-A, lmtk3*, *ddr1*, *ptk6* in birds; and *nr1i2/pxr* and *nr2b2/rxrb* in sauropsids. *Nr3b4/esrrd* is found in coelacanth and *Xenopus*, but not in other Sarcopterygii. *Styk1*-like is only found in *Xenopus* and a turtle. *Axl* is only found in the lizard among the Sauropsids.


(a)
100%
1R/2R
3R
Method 1: Retention of the duplicated genes
Method 2: Total retention within the orthogroups
NR
SOX
CTK
RTK
ADAMTS
Expected remaining duplicated genes
48-55%
18%
10-15%
500 mya
300 mya
100 mya
today
(b)
4R
3R
2R
1R
M1: e/b
M2: (e+b)/2b
a
a'
b
c
f
e
d
Expected 48-55% not observed here
18%
20-30%
500 mya
300 mya
100 mya
today
Increase of the initial number of gene before 1R/2R WGDs
Twofold increase of gene number from the Ts3R WGD
Expected gene increase from the Ss4R WGD
Neoteleostei vs Actinopterygii
Therians vs Sauropsids

◂**Fig. 4** Representation of the gene retention after the whole-genome duplications. **a** expected one-parameter exponential decrease of the duplicated genes from the values observed after the vertebrate 1R/2R, Ts3R, and Ss4R. Observed values of the remaining duplicated genes for the largest gene families that are considered in this work (M1, *bold lines*), and gene retention of the orthogroups using another mode of calculation (M2, *dashed lines*). These lines begin and end where the calculation of gene retention is made in the last common ancestor of the mammals and sauropsids for the 1R/2R, and the last common ancestor of the Neoteleostei and the Ostariophysi for the Ts3R. **b** hypothesis of the evolution of the number of genes in the genome of the vertebrates after the WGDs using the two-parameter model proposed by (Inoue et al. 2015). The number of genes in the ancestor of the vertebrate (*a*) was duplicated through the first round of WGD and followed by massive losses. The second round of WGD duplicated the number of genes (*a'*), and the pattern observed for the Ts3R by Inoue et al. is applied here. This model applied to the number of genes (*b*) in the ancestor of the teleosts that duplicated through the Ts3R and for the Salmoniformes (*c*) that duplicated through the Ss4R

Axl-like is only found in the coelacanth among Sarcopterygian. Several losses in independent multiple lineages can also occur: *nr0b3* is lost in both therians and fish; *nr1h2/lxrB* in birds and fish; *ptk6* in birds and Neoteleostei; and *nr1d1/rev-erbalpha* and *nr2f5* in birds and therians.

Gene losses after Ts3R also produced some peculiar pattern: 11 gene losses of one ohnolog are observed only in the Ostariophysi (Otocephala) (*vcan*, *bcan, hapln3, adamts2, sox6, sox10, nr2e3/pnr, nr2f2/coup-tf2, nr3b2/essrb, pdgfrb, ephb1*) and 11 in the Neoteleostei (*mc5r*, *adamts8, sox19, sox21, sox11, nr0b2/shp, nr1c2/ppard, nr2b2/rxrb, tek, ptk6, bmx*). Losses of both ohnologs are observed either in one lineage (*mc3r, sox12, ros1* in Neoteleostei, no full loss is observed in the Ostariophysi) or completely lost in teleosts (*adamts19*) or in Actinopterygii *(adamts4, sox30, nr2e1/tlx, nr0b3, nr1i3/car*). On average, the gain is zero after the Ts3R for two families: the MCR family gained one gene from the Ts3R in the Neoteleostei, but lost one gene in the Ostariophysi; and the adamts family lost two genes in fish since the Ts3R (*adamts13* and *adamts4*), a number that is compensated by the retention of two teleost duplicates: one in all fish (*adamts15*), one in Neoteleostei (*adamts2*), and one in Ostariophysi (*adamts8*).

Those numerous examples show that some genes are more prone to dispensability/expendability than others and raise question about the importance of their function. Singh et al. (Singh et al. 2012) proposed that purifying selection could more strongly retain genes that are prone to autosomal-dominant deleterious mutations. They suggested that the loss of one of the ohnolog would thus have a detrimental effect. RTKs and CTKs are on top of their list. Interestingly, we found that nuclear receptors and sox genes have a higher retention gene, no matter the method of calculation used (Fig. 4A). Adamts, sox, and CTK gene families that harbor the unusual distribution of genes (mostly 2 in adamts, 3 in sox, 2 or 4 only in CTK) are also the ones with the fewer examples of lineage-specific losses, a feature that favors the hypothesis of autosomal-dominant deleterious mutations.

2.3 Very Few SSDs After WGDs

While the vast majority of gene family expansion is due to the WGDs, a few occurrences of SSDs have occurred within these 7 gene families after the 1R/2R and 3R WGD episodes. A head-to-head tandem SSD occurred post-1R/2R leading to *mc2r* and *m5r* (Cortés et al. 2014). In the adamts4, -1, -5, -8, -15 subfamily, a parsimonious hypothesis suggests that a head-to-head duplication occurred for one of the duplicates between the 1R and 2R WGDs. *Adamts4* was left with no counterpart, and after the 2R WGD, the syntenic organization of the tandem genes *adamts1-adamts5* and *adamts15-adamts8* is conserved, even among the Ts3R duplicates. *Sry* is a retroposed copy of *sox3* found in therian species only. *Sox32*, the sole member of the soxG subfamily, is a head-to-tail tandem repeat of *sox17*. It is only found in fish and shows a similar size and same intron–exon features as its progenitor, suggesting that it appeared at the base of this lineage only and should belong to soxF group (Heenan et al. 2016; Voldoire et al. 2017). Among the CTK, *smrs* and *ptk6* are a head-to-tail tandem repeat that raised the number of FRK subfamily genes to three.

Several SSDs have been detected among the Eph receptor genes that are divided into EphA and EphB subfamilies. They are not phylogenetically clearly distinguished from one another, and they share the same domain architecture, but EphA genes differ from EphBs by an additional phase-1 intron. These SSDs preceded the vertebrate lineage expansion as many genes are found in the genome of the two sea squirts. *EphB6* is very divergent from *ephB1, -2, -3, -4* and is phylogenetically placed by sequence similarity among other *ephA* genes, reflecting a probable duplication prior to 1R/2R as there are several members of this gene in the two sea squirts, *C. intestinalis* and *C. savignyi*. Genes of the Vegfr and Pdgfr subfamilies arose by two successive tandem SSDs before the 1R/2R WGDs. This can be inferred from their actual phylogenetic position, their syntenic locations, the evolution of the intron positions and phases, and because all these genes are rooted by only one gene in the non-vertebrate deuterostomes (Brunet et al. 2016).

2.4 Different Methods to Calculate Gene Retention

Gene retention after WGDs has been studied in different organisms. It can be highly variable from one phylogenetic branch to another. For example, it was estimated to be about 48% in Paramecium species after a WGD that happened ∼320 mya (McGrath et al. 2014a, b). A retention fraction of 56% was calculated in *Xenopus laevis* since the allotetraploidization that occurred 17–18 mya (Session et al. 2016). In vertebrates, after the double events of 1R/2R WGDs, the retention rate has been estimated to be between 20 and 30% (Makino and McLysaght 2010). The 3R WGD pending of the approach and dataset is estimated between 12 and 24% (Postlethwait et al. 2000; Jaillon et al. 2004; Woods et al. 2005; Brunet et al. 2006; Steinke et al.

2006; Kassahn et al. 2009; Inoue et al. 2015) with a timing occurring between 225 and 333 mya (Hurley et al. 2007; Santini et al. 2009; Near et al. 2012; Betancur-R et al. 2013). As for the Ss4R, 55% of the duplicates are retained as two functional copies in the Atlantic salmon (Lien et al. 2016) and 48% in the rainbow trout (Berthelot et al. 2014) with an event estimated to be between 80 and 100 mya (Lien et al. 2016), 90 and 102 mya (Berthelot et al. 2014), and 88 and 103 mya (Macqueen and Johnston 2014).

The most usual method of calculating gene retention after a WGD usually considers only the duplicated copy, being either kept (100% retention) or lost (0% retention) (Brunet et al. 2006). After two rounds of WGDs, one ancestral gene leads to an orthogroup from 1 up to 4 ohnologs. With this method of calculation (M1), the retention rate would become 0%, 33%, 66%, and 100%, respectively, which is more a fraction deduced from the previous simple case, rather than a retention rate per se. However, this method has some skews. Whenever only one out of four ohnologs has disappeared, the timing of its loss can be easily deduced. In contrast, an orthogroup of two ohnologs can be the result of one loss after the 1R or two losses after the 2R when the two ohnologs were kept after the 1R. More losses imply even more possible scenarios. Also, this method does not consider cases for which none of the ohnologs are retained, either after a single round of WGD (Ts3R), or after two rounds (1R/2R). This case is observed in teleosts for *adamts4*, *adamts19*, *sox30*, *nr0b3*, *dax*, *insrr*, or when the two ohnologs are lost in one fish lineage, such as *mc3r*, *sox12*, *bmx*, *ros1,* and *shp*. It is also the case for *nr1e*, *nr5b,* and *nr7* that are observed in non-vertebrate deuterostomes but lost in vertebrates (*nr0a* is only found in Drosophila) (Lecroisey et al. 2012), although one cannot tell whether these genes were lost before or after the 1R/2R WGDs. For these cases, the loss of all members of an orthogroup would give a retention rate of −100% after one round of WGD and −33% after two rounds of WGDs, which seems counterintuitive as the loss of 4 genes should be more meaningful than the loss of 2 genes. To avoid dealing with these negative values, the retention rate should be zero whenever all members of an orthogroup are lost, and 100% when they are all kept whether it occurs after one or two WGDs, 2 or 4 ohnologs being expected (method 2, M2). This implies that if, on average, only half of the genes are kept in the orthogroups, the retention rate will be 50%, whenever there is only 1 gene after a single round of WGD or 2 genes left after two rounds. Method 1 rather considers the retention of the duplicated genes, while method 2 compares the retention within an orthogroup. The duplicated gene retention (M1) calculated following the 1R/2R WGDs and the Ts3R WGDs shows that the large gene families of NR, CTK, RTK, sox, and adamts harbor a higher gene retention than average, to the exception of the adamts gene family that presents on average no further expansion in fish after the Ts3R-WGD (Figs. 3 and 4A). Using M1, the retention rate after the Ts3R drops sharply for all other gene families, in comparison with the 1R/2R. Using M2, the retention rates decrease slowly.

Table 1 Detail of the retention rates calculated for each gene family

Families, subfamilies: genes	1R/2R	Out of	Met 2	Met 1		3R a	3R b	Out of	Met 2	Met 1	
Melanocortine receptor (MCR)											
MC1R MC3R MC4R MC5R	4	4		1.00		3.5	0.5	8		0.00	0
MC2R						1	0	2		0.00	0
sum	4	4	1.00	1.00		4.5	0.5	8	0.63		0
Hyaluronan (HA) and HAPLN											
HA: NCAN VCAN BCAN ACAN	4	4		1.00		4	3	8		0.75	6
HAPLN[1–4]	4	4		1.00		4	1.5	8		0.38	3
sum	8	8	1.00	1.00		8	4.5	16	0.78		0.56
ADAMTS											
ADAMTS[1, 4, 15]	3	4		0.67	2.67	2	1	6		0.00	0
ADAMTS[5, 8]	2	2		1.00	2.00	2	0.5	4		0.25	1
ADAMTS[9, 20]	2	4		0.33	1.33	2	0	4		0.00	0
ADAMTS[2, 3, 14]	3	4		0.67	2.67	3	0.5	6		0.17	1
ADAMTS[13]	1	4		0.00	0.00	1	0	2		0.00	0
ADAMTS[17, 19]	2	4		0.33	1.33	1	0	4		−0.50	−2
ADAMTS[7, 12]	2	4		0.33	1.33	2	0	4		0.00	0
ADAMTS[16, 18]	2	4		0.33	1.33	2	0	4		0.00	0
ADAMTS[6, 10]	2	4		0.33	1.33	2	0	4		0.00	0
sum	19	34	0.56		0.41	17	2	38	0.50		0.00
SOX											
SOXB1: SOX[1, 2, 3]	3	4		0.67		3	1	6		0.33	2
SOXB2: SOX[14, 21, 15/19]	3	4		0.67		3	2	6		0.67	4
SOXC: SOX[4, 11, 12]	3	4		0.67		2.5	1.5	6		0.33	2
SOXD: SOX[5, 6, 13]	3	4		0.67		3	0.5	6		0.17	1
SOXE: SOX[8, 9, 10]	3	4		0.67		3	2.5	6		0.83	5
SOXF: SOX[7, 17, 18]	3	4		0.67		3	0	6		0.00	0
SOXH: SOX[30]	1	4		0.00		0	0	2		−1.00	−2
sum	19	28	0.68	0.57		17.5	7.5	38	0.66		0.32
Cytosol tyrosine kinase (CTK)											
FRK:SRMS						1	0	2		0.00	0
FRK: FRK PTK6	2	4		0.33		2	0.5	4		0.25	1
FES: FES FER	2	4		0.33		2	0	4		0.00	0
ABL: ABL1 ABL2	2	4		0.33		2	0	4		0.00	0
SYK: SYK ZAP70	2	4		0.33		2	0	4		0.00	0
CSK: CSK MATK	2	4		0.33		2	1	4		0.50	2
ACK: TNK2 TNK1	2	4		0.33		2	1	4		0.50	2
PTK2: PTK2A PTK2B	2	4		0.33		2	2	4		1.00	4
TEC: BMX BTX ITK TEC	4	4		1.00		3.5	0	8		−0.13	−1

(continued)

Table 1 (continued)

Families, subfamilies: genes	1R/2R	Out of	Met 2	Met 1		3R a	3R b	Out of	Met 2	Met 1	
TEC: TXK						1	0	2		0.00	0
SRCB: BLK LCK LYN HCK	4	4		1.00		4	0	8		0.00	0
JAK: JAK3 JAK2 JAK1 TYK2	4	4		1.00		4	1	8		0.25	2
SRCA: YES1 SRC FYN FGR	4	4		1.00		4	3	8		0.75	6
sum	30	44	0.68	0.58		31.5	8.5	64	0.63		0.25
Receptor tyrosine kinase (RTK)											
ROS1	1	4		0.00		0.5	0	2		−0.50	−1
MUSK	1	4		0.00		1	0	2		0.00	0
PTK7	1	4		0.00		1	0	2		0.00	0
RET	1	4		0.00		1	0	2		0.00	0
RYK	1	4		0.00		1	0	2		0.00	0
TIE: TIE1 TEK	2	4		0.33		1.5	0	4		−0.25	−1
ROR: ROR1 ROR2	2	4		0.33		2	0	4		0.00	0
ALK: ALK TLK	2	4		0.33		2	0	4		0.00	0
MET: MST1R MET	2	4		0.33		2	1	4		0.50	2
STYK1: STYK1 STYK1-like	2	4		0.33		2	1	4		0.50	2
INSR: IGF1R INSR INSRR	3	4		0.67		2	2	6		0.33	2
DDR: DDR2 DDR1 DDR2-like	3	4		0.67		3	1	6		0.33	2
LMR: AATK LMTK2 LMTK3	3	4		0.67		3	1	6		0.33	2
TRK: NTRK2 NTRK3 NTRK1	3	4		0.67		3	2	6		0.67	4
TAM: AXL TYRO3 MERTK AXL-like	4	4		1.00		4	0	8		0.00	0
FGFR: FGFR2 FGFR1 FGFR3 FGFR4	4	4		1.00		4	1	8		0.25	2
ERBB: ERBB2 EGFR ERBB3 ERBB4	4	4		1.00		4	3	8		0.75	6
EPHB: EPHB1 EPHB2 EPHB3 EPHB4	4	4		1.00		4	3.5	8		0.88	7
VEGFR: KDR-like KDR FLT4 FLT1	4	4		1.00		4	0	8		0.00	0
PDGFR: KIT CSF1R FLT3	3	4		0.67		3	2	6		0.67	4
PDGFR: PDGFRA PDGFRB	2	4		0.33		2	0.5	4		0.25	1
sum	52	84	0.62	0.49		50	18	104	0.65		0.31

(continued)

Table 1 (continued)

Families, subfamilies: genes	1R/2R	Out of	Met 2	Met 1		3R a	3R b	Out of	Met 2	Met 1	
Nuclear receptor (NR)											
NR2E1: TLX	1	4		0.00		1	0	2		0.00	0
NR2E3: PNR	1	4		0.00		1	0.5	2		0.50	1
NR6A: GCNF	1	4		0.00		1	1	2		1.00	2
NR2C: TR2 TR4	2	4		0.33		2	0	4		0.00	0
NR1A: THRa THRb	2	4		0.33		2	1	4		0.50	2
NR3A: ESR2 ESR1	2	4		0.33		2	1	4		0.50	2
NR0B:SHP DAX NR0B3	3	4		0.33		1	0.5	6		−0.50	−3
NR5A: SF1 LRH1 NR5A5	3	4		0.67		3	1	6		0.33	2
NR2A: HNF4a HNF4g HFN4b	3	4		0.67		3	0	6		0.00	0
NR1C: PPARa PPARd PPARg	3	4		0.67		3	1.5	6		0.50	3
NR2B: RXRa RXRb RXRg	3	4		0.67		3	1.5	6		0.50	3
NR1I: VDR PXR CAR	3	4		0.67		2	1	6		0.00	0
NR1D: REV-ERBb NR1D4 REV-ERBa	3	4		0.67		3	2	6		0.67	4
NR4A: NGF1B NURR1 NOR1	3	4		0.67		3	2	6		0.67	4
NR1B: RARa RARb RARg	3	4		0.67		3	3	6		1.00	6
NR1H: LXRB LXRA FXR NR1H5	4	4		1.00		3	0	8		−0.25	−2
NR2F: COUPTFa COUPTFb EAR2 NR2F5	4	4		1.00		4	2.5	8		0.63	5
NR3B: ESRRa ESRRg ESRRb ESRD	4	4		1.00		4	1.5	8		0.38	3
NR1F: RORb RORa RORd RORg	4	4		1.00		4	2	8		0.50	4
NR3C: MR GR AR PGR	4	4		1.00		4	2	8		0.50	4
sum	56	80	0.70	0.58		52	24	112	0.68		0.36

2.5 The Two-Gene Model of Gene Loss Following a WGD

According to (Inoue et al. 2015), after a WGD, the process and mechanism of gene loss match a two-phase model. Based on a calculation made from teleost fish genomes, the first phase corresponds to a massive and rapid loss of genes immediately after the WGD, from 72 to 82% of the duplicates within about 10–20 myr. This intense process stalls rapidly to reveal the second one, which is the slow and regular decay of the duplicated genes (see Fig. 4B). The first mechanism

corresponds to the deletion of large chromosomal segments or contiguous clusters, or even losses of enhancer regulating genes dispersed throughout the genome, that would subsequently trigger their pseudogenization process. Ohnologous genes becoming subfunctionalized or neofunctionalized (Force et al. 1999) would therefore be under selective pressure and could reduce the decay and loss process. Inoue et al. (Inoue et al. 2015) suggested that the species diversification process observed for the teleost fish could occur only during the second phase. A similar observation was also made in the Salmoniformes, for which the Ss4R took place about 20 myr before the species diversification started (Berthelot et al. 2014; Macqueen and Johnston 2014; Lien et al. 2016). Based on the parameters they calculated to get their two-phase model equation (Inoue et al. 2015) matching the Ts3R data collected, we reproduced this graph for the 1R + 2R occurrences. Pending that the estimation of 100 myr between these two events of WGDs (Wang and Gu 2000) is correct, then close to 90% of duplicated genes could have been lost from the 1R-WGD before the 2R-WGD occurrence. Of note, this could be misleading as it gives some semblance of multiple successive events of large segmental duplications, as proposed as an alternative evolutionary hypothesis to the 1R/2R WGDs (Smith and Keinath 2015).

In Fig. 4B, we took the two-phase model gene loss proposed by (Inoue et al. 2015) for the Ts3R WGD (in blue) and used it for the 1R/2R WGDs (in red), then the Ss4R WGD (in green). Their model used for the Ts3R could match the 1R/2R WGDs with the 20–30% remaining genes calculated by (Makino and McLysaght 2010). However, it does not work for the Ss4R in which remaining duplicates are expected to range within 48–55%. Either this value is overestimated with several duplicated genes in the process of pseudogenization, or the parameters of the Ts3R are different from one WGD to another. In addition, it may also be that the duplicated genomes could first experience a stasis of variable duration before they engage in severe gene losses by shuffling and recombination events. Perhaps the end of the stasis may be triggered by a change in the environmental stressful conditions that could enhance those processes (e.g., ectopic recombinations following transposable element expansion). There is a discrepancy regarding the occurrence of this stasis hypothesis between the surveys done on the rainbow trout (Berthelot et al. 2014) and the Atlantic salmon (Lien et al. 2016) genomes. In the study of the specific common carp Cca4R allotetraploidization (Li et al. 2015), the authors found a very low rate of evolution between the duplicates and no evidence of severe gene losses since more than 92% of genes remained duplicated after only 8 mya, which is in favor of the existence of this stasis. More efforts will have to be made to reconstruct the processes involved after WGDs and find out whether these patterns could be repeated after any round of WGD.

3 Conclusion

The WGDs at the base of vertebrates and teleostean fish are key events that were instrumental for the diversification and evolutionary innovations observed in these lineages (Cañestro et al. 2013). The maintenance of duplicates from specific gene categories has been correlated with additional levels of cell complexity and consequently with a subsequent gain in organismal diversity (Maere et al. 2005; Freeling and Thomas 2006; Sémon and Wolfe 2007; Huminiecki and Heldin 2010). Also, and very intriguingly, the composition of the orthogroups within the gene families studied here seems to follow a non-random distribution: adamts have mostly 2 ohnologs, sox have mostly 3, and CTK has either 2 or 4 after the 1R/2R WGDs, and show a strong resilience to loss in every lineage. More investigations to look for those kinds of unusual pattern appear worthwhile to be done systematically in other large gene families involved in development, cell, and tissue diversification of vertebrates (Table 1).

References

Aury JM, Jaillon O, Duret L, Noel B, Jubin C, Porcel BM, Ségurens B, Daubin V, Anthouard V, Aiach N, Arnaiz O, Billaut A, Beisson J, Blanc I, Bouhouche K, Câmara F, Duharcourt S, Guigo R, Gogendeau D, Katinka M, Keller AM, Kissmehl R, Klotz C, Koll F, Le Mouël A, Lepère G, Malinsky S, Nowacki M, Nowak JK, Plattner H, Poulain J, Ruiz F, Serrano V, Zagulski M, Dessen P, Bétermier M, Weissenbach J, Scarpelli C, Schächter V, Sperling L, Meyer E, Cohen J, Wincker P (2006) Global trends of whole-genome duplications revealed by the ciliate *Paramecium tetraurelia*. Nature 444:171–178

Berthelot C, Brunet F, Chalopin D, Juanchich A, Bernard M, Noël B, Bento P, Da Silva C, Labadie K, Alberti A, Aury JM, Louis A, Dehais P, Bardou P, Montfort J, Klopp C, Cabau C, Gaspin C, Thorgaard GH, Boussaha M, Quillet E, Guyomard R, Galiana D, Bobe J, Volff JN, Genêt C, Wincker P, Jaillon O, Roest Crollius H, Guiguen Y (2014) The rainbow trout genome provides novel insights into evolution after whole-genome duplication in vertebrates. Nat Commun 5:3657. doi:10.1038/ncomms4657

Bertrand S, Belgacem MR, Escriva H (2011) Nuclear hormone receptors in chordates. Mol Cell Endocrinol 334:67–75

Bertrand S, Brunet F, Escriva H, Parmentier G, Laudet V, Robinson-Rechavi M (2004) Evolutionary genomics of nuclear receptors: from 25 ancestral genes to derived endocrine systems. Mol Biol Evol 21:1923–1937

Betancur-R R, Broughton RE, Wiley EO, Carpenter K, López JA, Li C, Holcroft NI, Arcila D, Sanciangco M, Cureton Ii JC, Zhang F, Buser T, Campbell MA, Ballesteros JA, Roa-Varon A, Willis S, Borden WC, Rowley T, Reneau PC, Hough DJ, Lu G, Grande T, Arratia G, Ortí G (2013) The tree of life and a new classification of bony fishes. PLoS Curr 5 pii: ecurrents.tol.53ba26640df0ccaee75bb165c8c26288

Bianchi ME, Beltrame M (1998) Flexing DNA: HMG-box proteins and their partners. Am J Hum Genet 63:1573–1577

Binder MJ, McCoombe S, Williams ED, McCulloch DR, Ward CA (2017) The extracellular matrix in cancer progression: role of hyalectan proteoglycans and ADAMTS enzymes. Cancer Lett 28:55–64

Blanc G, Wolfe KH (2004) Functional divergence of duplicated genes formed by polyploidy during Arabidopsis evolution. Plant Cell 16:1679–1691

Blomme T, Vandepoele K, De Bodt S, Simillion C, Maere S, Van de Peer Y (2006) The gain and loss of genes during 600 million years of vertebrate evolution. Genome Biol 7:R43

Brocker CN, Vasilis Vasiliou V, Nebert DW (2009) Evolutionary divergence and functions of the ADAM and ADAMTS gene families. Human Genomics 4:43

Bronwen LA, Ayling S, Barrell D, Clarke L, Curwen V, Fairley S, Fernandez Banet J, Billis K, Garcín Girón C, Hourlier T, Howe K, Kähäri A, Kokocinski F, Martin FJ, Murphy DN, Nag R, Ruffier M, Schuster M, Amy Tang Y, Vogel J-H, White S, Zadissa A, Flicek P, Searle SMJ (2016) The Ensembl gene annotation system. Database, baw093. doi:10.1093/database/baw093

Brunet F, Fraser FW, Binder MJ, Smith AD, Kintakas C, Dancevic CM, Ward AC, McCulloch DR (2015) The evolutionary conservation of the A Disintegrin-like and Metalloproteinase domain with Thrombospondin-1 motif metzincins across vertebrate species and their expression in teleost zebrafish. BMC Evol Biol 15:22

Brunet F, Kintakas C, Smith A, McCulloch DR (2012) The function of the hyalectan class of proteoglycans and their binding partners during vertebrate development. In: Berhardt LV (ed) Advances in medicine and biology, vol 52. ISBN: 978-1-62081-339-3

Brunet F, Roest Crollius H, Paris M, Aury JM, Gibert P, Jaillon O, Laudet V, Robinson-Rechavi M (2006) Gene loss and evolutionary rates following whole genome duplication in teleost fishes. Mol Biol Evol 23:1808–1816

Brunet F, Volff J-N, Schartl M (2016) Whole genome duplications shaped the receptor tyrosine kinase repertoire of jawed vertebrates. Genome Biol Evol 8:1600–1613

Cañestro C, Albalat R, Irimia M, Garcia-Fernàndez J (2013) Impact of gene gains, losses and duplication modes on the origin and diversification of vertebrates. Semin Cell Dev Biol 24:83–94. doi:10.1016/j.semcdb.2012.12.008

Conant GC, Wolfe KH (2008) Turning a hobby into a job: how duplicated genes find new functions. Nature Rev Genet 9:938–950

Cortés R, Navarro S, Agulleiro MJ, Guillot R, García-Herranz V, Sánchez E, Cerdá-Reverter JM (2014) Evolution of the melanocortin system. Gen Comp Endocrinol 209:3–10. doi:10.1016/j.ygcen.2014.04.005

D'Aniello S, Irimia M, Maeso I, Pascual-Anaya J, Jiménez-Delgado S, Bertrand S, Garcia-Fernàndez J (2008) Gene expansion and retention leads to a diverse tyrosine kinase superfamily in amphioxus. Mol Biol Evol 25:1841–1854

Davis JC, Petrov DA (2005) Do disparate mechanisms of duplication add similar genes to the genome? Trends Genet 21:548–551

Dehal P, Boore JL (2005) Two rounds of whole genome duplication in the ancestral vertebrate. PLoS Biol 3:e314

Devos KM, Brown JKM, Bennetzen JL (2002) Genome size reduction through illegitimate recombination counteracts genome expansion in Arabidopsis. Genome Res 12:1075–1079

Evans BJ, Kelley DB, Tinsley RC, Melnick DJ, Cannatella DC (2004) A mitochondrial DNA phylogeny of African clawed frogs: phylogeography and implications for polyploid evolution. Mol Phylogenet Evol 33:197–213

Fischer G, Rocha EP, Brunet F, Vergassola M, Dujon B (2006) Highly variable rates of genome rearrangements between Hemiascomycetous yeast lineages. PLoS Genet 2:e32

Force A, Lynch M, Pickett FB, Amores A, Yan YL, Postlethwait J (1999) Preservation of duplicate genes by complementary, degenerative mutations. Genetics 151:1531–1545

Foster JW, Graves JA (1994) An SRY-related sequence on the marsupial X chromosome: implications for the evolution of the mammalian testis-determining gene. Proc Natl Acad Sci USA 91:1927–1931

Freeling M, Thomas BC (2006) Gene-balanced duplications, like tetraploidy, provide predictable drive to increase morphological complexity. Genome Res 16:805–814

Germain P, Staels B, Dacquet C, Spedding M, Laudet V (2006) Overview of nomenclature of nuclear receptors. Pharmacol Rev 58:685–704

Gout J-F, Lynch M (2015) Maintenance and loss of duplicated genes by dosage subfunctionalization. Mol Biol Evol 32:2141–2148

Gschwind A, Fischer OM, Ullrich A (2004) The discovery of receptor tyrosine kinases: targets for cancer therapy. Nat Rev Cancer 4:361–370

Hanks SK (2003) Genomic analysis of the eukaryotic protein kinase superfamily: a perspective. Genome Biol 4:111

Heenan P, Zondag L, Wilson MJ (2016) Evolution of the Sox gene family within the chordate phylum. Gene 575:385–392

Holland LZ, Albalat R, Azumi K, Benito-Gutiérrez E, Blow MJ, Bronner-Fraser M, Brunet F, Butts T, Candiani S, Dishaw LJ, Ferrier DE, Garcia-Fernàndez J, Gibson-Brown JJ, Gissi C, Godzik A, Hallböök F, Hirose D, Hosomichi K, Ikuta T, Inoko H, Kasahara M, Kasamatsu J, Kawashima T, Kimura A, Kobayashi M, Kozmik Z, Kubokawa K, Laudet V, Litman GW, McHardy AC, Meulemans D, Nonaka M, Olinski RP, Pancer Z, Pennacchio LA, Pestarino M, Rast JP, Rigoutsos I, Robinson-Rechavi M, Roch G, Saiga H, Sasakura Y, Satake M, Satou Y, Schubert M, Sherwood N, Shiina T, Takatori N, Tello J, Vopalensky P, Wada S, Xu A, Ye Y, Yoshida K, Yoshizaki F, Yu JK, Zhang Q, Zmasek CM, de Jong PJ, Osoegawa K, Putnam NH, Rokhsar DS, Satoh N, Holland PW (2008) The amphioxus genome illuminates vertebrate origins and cephalochordate biology. Genome Res 8:1100–1111

Hufton AL, Groth D, Vingron M, Lehrach H, Poustka AJ, Panopoulou G (2008) Early vertebrate whole genome duplications were predated by a period of intense genome rearrangement. Genome Res 18:1582–1591

Hughes MK, Hughes AL (1993) Evolution of duplicate genes in a tetraploid animal, *Xenopus laevis*. Mol Biol Evol 10:1360–1369

Hughes T, Ekman D, Ardawatia H, Elofsson A, Liberles DA (2007) Evaluating dosage compensation as a cause of duplicate gene retention in Paramecium tetraurelia. Genome Biol 8:213

Huminiecki L, Heldin CH (2010) 2R and remodeling of vertebrate signal transduction engine. BMC Biol 8:146

Hurley IA, Mueller RL, Dunn KA, Schmidt EJ, Friedman M, Ho RK, Prince VE, Yang Z, Thomas MG, Coates MI (2007) A new time-scale for ray-finned fish evolution. Proc Biol Sci 274:489–498

Inoue J, Yukuto S, Sinclair R, Tsukamoto K, Nishida M (2015) Rapid genome reshaping by multiple-gene loss after while-genome duplication in teleost fish suggested by mathematical modeling. Proc Natl Acad Sci 112:14918–14923

Jager M, Queinnec E, Houliston E, Manuel M (2006) Expansion of the SOX gene family predated the emergence of the Bilateria. Mol Phylogenet Evol 39:468–477

Jager M, Quéinnec E, Le Guyader H, Manuel M (2011) Multiple Sox genes are expressed in stem cells or in differentiating neuro-sensory cells in the hydrozoan Clytia hemisphaerica. Evodevo 2:12

Jaillon O, Aury JM, Brunet F, Petit JL, Stange-Thomann N, Mauceli E, Bouneau L, Fischer C, Ozouf-Costaz C, Bernot A, Nicaud S, Jaffe D, Fisher S, Lutfalla G, Dossat C, Segurens B, Dasilva C, Salanoubat M, Levy M, Boudet N, Castellano S, Anthouard V, Jubin C, Castelli V, Katinka M, Vacherie B, Biémont C, Skalli Z, Cattolico L, Poulain J, De Berardinis V, Cruaud C, Duprat S, Brottier P, Coutanceau JP, Gouzy J, Parra G, Lardier G, Chapple C, McKernan KJ, McEwan P, Bosak S, Kellis M, Volff JN, Guigó R, Zody MC, Mesirov J, Lindblad-Toh K, Birren B, Nusbaum C, Kahn D, Robinson-Rechavi M, Laudet V, Schachter V, Quétier F, Saurin W, Scarpelli C, Wincker P, Lander ES, Weissenbach J, Roest Crollius H (2004) Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-karyotype. Nature 431:946–957

Kasahara M (2007) The 2R hypothesis: an update. Curr Opin Immunol 19:547–552

Kassahn KS, Dang VT, Wilkins SJ, Perkins AC, Ragan MA (2009) Evolution of gene function and regulatory control after whole-genome duplication: comparative analyses in vertebrates. Genome Res 19:1404–1418

Kellis M, Birren BW, Lander ES (2004) Proof and evolutionary analysis of ancient genome duplication in the yeast Saccharomyces cerevisiae. Nature 428:617–624

Kelwick R, Desanlis I, Wheeler GN, Edwards DR (2015) The ADAMTS (A Disintegrin and Metalloproteinase with Thrombospondin motifs) family. Genome Biol 16:113. doi:10.1186/s13059-015-0676-3

Kenny NJ, Chan KW, Nong W, Qu Z, Maeso I, Yip HY, Chan TF, Kwan HS, Holland PW, Chu KH, Hui JH (2016) Ancestral whole-genome duplication in the marine chelicerate horseshoe crabs. Heredity 116:90–199

Kuwada Y (1911) Meiosis in the pollen mother cells of Zea Mays L. Bot Mag 25:163

Lahn BT, Page DC (1999) Four evolutionary strata on the human X chromosome. Science 286:964–967

Larroux C, Fahey B, Liubicich D, Hinman VF, Gauthier M, Gongora M, Green K, Worheide G, Leys SP, Degnan BM (2006) Developmental expression of transcription factor genes in a demosponge: insights into the origin of metazoan multicellularity. Evol Dev 8:150–173

Laudet V, Stehelin D, Clevers H (1993) Ancestry and diversity of the HMG box superfamily. Nucleic Acids Res 21:2493–2501

Le Gouar M, Guillou A, Vervoort M (2004) Expression of a SoxB and a Wnt2/13 gene during the development of the mollusc *Patella vulgata*. Dev Genes Evol 214:250–256

Lecroisey C, Laudet V, Schubert M (2012) The cephalochordate amphioxus: a key to reveal the secrets of nuclear receptor evolution. Brief Funct Genomics 11:156–166. doi:10.1093/bfgp/els008

Lee TH, Tang H, Wang X, Paterson AH (2012) PGDD: a database of gene and genome duplication in plants. Nucleic Acids Res 41:D1152–D1158

Lemmon MA, Schlessinger J (2010) Cell signaling by receptor tyrosine kinases. Cell 141:1117–1134

Li JT, Hou GY, Kong XF, Li CY, Zeng JM, Li HD, Xiao GB, Li XM, Sun XW (2015) The fate of recent duplicated genes following a fourth-round whole genome duplication in a tetraploid fish, common carp (*Cyprinus carpio*). Sci Rep 5:8199. doi:10.1038/srep08199

Li WH (1983) Evolution of duplicate genes and pseudogenes. In: Nei M, Koehn RK (eds) Evolution of genes and proteins. Sinauer Associates, Sunderland, MA, pp 14–37

Li WH, Gu Z, Wang H, Nekrutenko A (2001) Evolutionary analyses of the human genome. Nature 409:847–849

Lien S, Koop BF, Sandve SR, Miller JR, Kent MP, Nome T, Hvidsten TR, Leong JS, Minkley DR, Zimin A, Grammes F, Grove H, Gjuvsland A, Walenz B, Hermansen RA, von Schalburg K, Rondeau EB, Di Genova A, Samy JK, Olav Vik J, Vigeland MD, Caler L, Grimholt U, Jentoft S, Våge DI, de Jong P, Moen T, Baranski M, Palti Y, Smith DR, Yorke JA, Nederbragt AJ, Tooming-Klunderud A, Jakobsen KS, Jiang X, Fan D, Hu Y, Liberles DA, Vidal R, Iturra P, Jones SJ, Jonassen I, Maass A, Omholt SW, Davidson WS (2016) The Atlantic salmon genome provides insights into rediploidization. Nature 533:200–205. doi:10.1038/nature17164

Louis A, Muffato M, Roest Crollius H (2012) Genomicus: five genome browsers for comparative genomics in Eukaryota. Nucleic Acids Res 41:D700–D705

Macqueen DJ, Johnston IA (2014) A well-constrained estimate for the timing of the salmonid whole genome duplication reveals major decoupling from species diversification. Proc R Soc. B 281:20132881

Maere S, De Bodt S, Raes J, Casneuf T, Van Montagu M, Kuiper M, Van de Peer Y (2005) Modeling gene and genome duplications in eukaryotes. Proc Natl Acad Sci USA 102:5454–5459

Makino T, McLysaght A (2010) Ohnologs in the human genome are dosage balanced and frequently associated with disease. Proc Natl Acad Sci 107:9270–9274

Manning G, Whyte DB, Martinez R, Hunter T, Sudarsanam S (2002) The protein kinase complement of the human genome. Science 298:1912–1934

Markov GV, Laudet V (2011) Origin and evolution of the ligand-binding ability of nuclear receptors. Mol Cell Endocrinol 334:21–30

McCulloch DR, Nelson CM, Dixon LJ, Silver DL, Wylie JD, Lindner V, Sasaki T, Cooley MA, Argraves WS, Apte SS (2009) ADAMTS metalloproteases generate active versican fragments that regulate interdigital web regression. Dev Cell 17:687–698

McGrath CL, Gout J-F, Johri P, Doak TG, Lynch M (2014a) Differential retention and divergent resolution of duplicate genes following whole genome duplication. Genome Res 24:1665–1675

McGrath CL, Gout JF, Doak TG, Yanagi A, Lynch M (2014b) Insights into three whole-genome duplications gleaned from the *Paramecium caudatum* genome sequence. Genetics 197:1417–1428

McLysaght A, Hokamp K, Wolfe KH (2002) Extensive genomic duplication during early chordate evolution. Nat Genet 31:200–204

Meyer A, Schartl M (1999) Gene and genome duplications in vertebrates: the one-to-four (-to-eight in fish) rule and the evolution of novel gene functions. Curr Opin Cell Biol 11:699–704

Nakatani Y, Takeda H, Kohara Y, Morishita S (2007) Reconstruction of the vertebrate ancestral genome reveals dynamic genome reorganization in early vertebrates. Genome Res 17:1254–1265

Naruse K, Tanaka M, Mita K, Shima A, Postlethwait J, Mitani H (2004) A medaka gene map: the trace of ancestral vertebrate proto-chromosomes revealed by comparative gene mapping. Genome Res 14:820–828

Near TJ, Eytan RI, Dornburg A, Kuhn KL, Moore JA, Davis MP, Wainwright PC, Friedman M, Smith WL (2012) Resolution of ray-finned fish phylogeny and timing of diversification. Proc Natl Acad Sci USA 109:13698–13703

Nicholson AC, Malik SB, Logsdon JM Jr, Van Meir EG (2005) Functional evolution of ADAMTS genes: evidence from analyses of phylogeny and geneorganization. BMC Evol Biol 5:11

Ohno S (1970) Evolution of gene duplication. Springer, New-York

Ohno S (1999) Gene duplication and the uniqueness of vertebrate genomes circa 1970–1999. Semin Cell Dev Biol 10:517–522

Okuda Y, Yoda H, Uchikawa M, Furutani-Seiki M, Takeda H, Kondoh H, Kamachi Y (2006) Comparative genomic and expression analysis of group B1 sox genes in zebrafish indicates their diversification during vertebrate evolution. Dev Dyn 235:811–825

Otte K, Kranz H, Kober I, Thompson P, Hoefer M, Haubold B, Remmel B, Voss H, Kaiser C, Albers M, Cheruvallath Z, Jackson D, Casari G, Koegl M, Pääbo S, Mous J, Kremoser C, Deuschle U (2003) Identification of farnesoid X receptor beta as a novel mammalian nuclear receptor sensing lanosterol. Mol Cell Biol 23:864–872. doi:10.1128/mcb.23.3.864-872.2003

Panopoulou G, Hennig S, Groth D, Krause A, Poustka AJ, Herwig R, Vingron M, Lehrach H (2003) New evidence for genome-wide duplications at the origin of vertebrates using an amphioxus gene set and completed animal genomes. Genome Res 13:1056–1106

Panopoulou G, Poustka AJ (2017) Timing and mechanism of ancient vertebrate genome duplications—the adventure of a hypothesis. Trends Genetics in press

Paterson AH, Chapman BA, Kissinger JC, Bowers JE, Feltus FA, Estill JC (2006) Many gene and domain families have convergent fates following independent whole-genome duplication events in *Arabidopsis*, *Oryza, Saccharomyces* and *Tetraodon*. Trends Genet 22:597–602

Postlethwait JH, Woods IG, Ngo-Hazelett P, Yan YL, Kelly PD, Chu F, Huang H, Hill-Force A, Talbot WS (2000) Zebrafish comparative genomics and the origins of vertebrate chromosomes. Genome Res 10:1890–1902

Rajkov J, Shao Z, Berrebi P (2014) Evolution of polyploidy and functional diploidization in sturgeons: microsatellite analysis in 10 sturgeon species. J Hered 105:521–531

Ravi V, Lam K, Tay BH, Tay A, Brenner S, Venkatesh B (2009) Elephant shark (*Callorhinchus milii*) provides insights into the evolution of Hox gene clusters in gnathostomes. Proc Natl Acad Sci USA 106:16327–16332. doi:10.1073/pnas.0907914106

Robinson DR, Wu YM, Lin SF (2000) The protein tyrosine kinase family of the human genome. Oncogene 19:5548–5557

Santini F, Harmon LJ, Carnevale G, Alfaro ME (2009) Did genome duplication drive the origin of teleosts? A comparative study of diversification in ray-finned fishes. BMC Evol Biol 9:94

Sato Y, Hashiguchi Y, Nishida M (2009) Temporal pattern of loss/persistence of duplicate genes involved in signal transduction and metabolic pathways after teleost-specific genome duplication. BMC Evol Biol 9:127

Schartl M, Volff J-N, Brunet F (2015) Evolution of receptor tyrosine kinases. In: Yarden Y (ed) Receptor tyrosine kinases: structure, functions and role in human disease handbook

Schlessinger J (2014) Receptor tyrosine kinases: legacy of the first two decades. Cold Spring Harb Perspect Biol 6:a008912

Schmid M, Evans BJ, Bogart JP (2015) Polyploidy in Amphibia. Cytogenet Genome Res 145:315–330

Schubert M, Brunet F, Paris M, Bertrand S, Benoit G, Laudet V (2008) Nuclear hormone receptor signaling in amphioxus. Dev Genes Evol 218:651–665

Seals DF, Courtneidge SA (2003) The ADAMs family of metalloproteases: multidomain proteins with multiple functions. Genes Dev 17:7–30

Sekido R, Lovell-Badge R (2009) Sex determination and SRY: down to a wink and a nudge? Trends Genet 25:19–29

Sémon M, Wolfe KH (2007) Consequences of genome duplication. Curr Opin Genet Dev 17:505–512

Session AM, Uno Y, Kwon T, Chapman JA, Toyoda A, Takahashi S, Fukui A, Hikosaka A, Suzuki A, Kondo M, van Heeringen SJ, Quigley I, Heinz S, Ogino H, Ochi H, Hellsten U, Lyons JB, Simakov O, Putnam N, Stites J, Kuroki Y, Tanaka T, Michiue T, Watanabe M, Bogdanovic O, Lister R, Georgiou G, Paranjpe SS, van Kruijsbergen I, Shu S, Carlson J, Kinoshita T, Ohta Y, Mawaribuchi S, Jenkins J, Grimwood J, Schmutz J, Mitros T, Mozaffari SV, Suzuki Y, Haramoto Y, Yamamoto TS, Takagi C, Heald R, Miller K, Haudenschild C, Kitzman J, Nakayama T, Izutsu Y, Robert J, Fortriede J, Burns K, Lotay V, Karimi K, Yasuoka Y, Dichmann DS, Flajnik MF, Houston DW, Shendure J, DuPasquier L, Vize PD, Zorn AM, Ito M, Marcotte EM, Wallingford JB, Ito Y, Asashima M, Ueno N, Matsuda Y, Veenstra GJ, Fujiyama A, Harland RM, Taira M, Rokhsar DS (2016) Genome evolution in the allotetraploid frog Xenopus laevis. Nature 538:336–343. doi:10.1038/nature19840

Seoighe C, Gehring C (2004) Genome duplication led to highly selective expansion of the Arabidopsis thaliana proteome. Trends Genet 20:461–464

Singh PP, Affeldt S, Cascone I, Selimoglu R, Camonis J, Isambert H (2012) On the expansion of "dangerous" gene repertoires by whole-genome duplications in early vertebrates. Cell Rep 2:1387–1398

Smith JJ, Keinath MC (2015) The sea lamprey meiotic map improves resolution of ancient vertebrate genome duplications. Genome Res 25:1081–1090

Smith JJ, Kuraku S, Holt C, Sauka-Spengler T, Jiang N, Campbell MS, Yandell MD, Manousaki T, Meyer A, Bloom OE, Morgan JR, Buxbaum JD, Sachidanandam R, Sims C, Garruss AS, Cook M, Krumlauf R, Wiedemann LM, Sower SA, Decatur WA, Hall JA, Amemiya CT, Saha NR, Buckley KM, Rast JP, Das S, Hirano M, McCurley N, Guo P, Rohner N, Tabin CJ, Piccinelli P, Elgar G, Ruffier M, Aken BL, Searle SM, Muffato M, Pignatelli M, Herrero J, Jones M, Brown CT, Chung-Davidson YW, Nanlohy KG, Libants SV, Yeh CY, McCauley DW, Langeland JA, Pancer Z, Fritzsch B, de Jong PJ, Zhu B, Fulton LL, Theising B, Flicek P, Bronner ME, Warren WC, Clifton SW, Wilson RK, Li W (2013) Sequencing of the sea lamprey (*Petromyzon marinus*) genome provides insights into vertebrate evolution. Nature Genet 45:415–421

Soltis PS, Marchant DB, Van de Peer Y, Soltis DE (2015) Polyploidy and genome evolution in plants. Curr Opin Genet Dev 35:119–125

Soullier S, Jay P, Poulat F, Vanacker J-M, Berta P, Laudet V (1999) Diversification pattern of the HMG and SOX family members during evolution. J Mol Evol 48:517–527

Spicer AP, Joo A, Bowling RA Jr (2003) A hyaluronan binding link protein gene family whose members are physically linked adjacent to chondroitin sulfate proteoglycan core protein genes: the missing links. J Biol Chem 278:21083–21091

Srivastava M, Simakov O, Chapman J, Fahey B, Gauthier ME, Mitros T et al (2010) The *Amphimedon queenslandica* genome and the evolution of animal complexity. Nature 466:720–726

Steinke D, Salzburger W, Braasch I, Meyer A (2006) Many genes in fish have species-specific asymmetric rates of molecular evolution. BMC Genom 7:20

Stupka N, Kintakas C, White JD, Fraser FW, Hanciu M, Aramaki-Hattori N, Martin S, Coles C, Collier F, Ward AC, Apte SS, McCulloch DR (2013) Versican processing by a disintegrin-like and metalloproteinase domain with thrombospondin-1 repeats proteinases-5 and -15 facilitates myoblast fusion. J Biol Chem 288:1907–1917

Suga H, Sasaki G, Kuma K, Nishiyori H, Hirose N, Su ZH, Iwabe N, Miyata T (2008) Ancient divergence of animal protein tyrosine kinase genes demonstrated by a gene family tree including choanoflagellate genes. FEBS Lett 582:815–818

Takahashi A, Kawauchi H (2006) Evolution of melanocortin systems in fish. Gen Comp Endocrinol 148:85–94

Taylor JS, Braasch I, Frickey T, Meyer A, Van de Peer Y (2003) Genome duplication, a trait shared by 22000 species of ray-finned fish. Genome Res 13:382–390

Taylor JS, Raes J (2004) Duplication and divergence: the evolution of new genes and old ideas. Annu Rev Genet 38:615–643

Turunen O, Seelke R, Macosko J (2009) *In silico* evidence for functional specialization after genome duplication in yeast. FEMS Yeast Res 9:16–31

Uno Y, Nishida C, Takagi C, Ueno N, Matsuda Y (2013) Homoeologous chromosomes of *Xenopus laevis* are highly conserved after whole-genome duplication. Heredity 111:430–436

Vanneste K, Baele G, Maere S, Van de Peer Y (2014) Analysis of 41 plant genomes supports a wave of successful genome duplications in association with the Cretaceous-Paleogene boundary. Genome Res 24:1334–1347

Västermark Å, Schiöth HB (2011) The early origin of melanocortin receptors, agouti-related peptide, agouti signalling peptide, and melanocortin receptor-accessory proteins, with emphasis on pufferfishes, elephant shark, lampreys, and amphioxus. Eur J Pharmacol 660:61–69

Veyrunes F, Waters PD, Miethke P, Rens W, McMillan D, Alsop AE, Grutzner F, Deakin JE, Whittington CM, Schatzkamer K, Kremitzki CL, Graves T, Ferguson-Smith MA, Warren W, Marshall Graves JA (2008) Bird-like sex chromosomes of platypus imply recent origin of mammal sex chromosomes. Genome Res 18:965–973

Voldoire E, Brunet FG, Naville M, Volff J-N, Galiana D (2017) Expansion by whole genome duplication and evolution of the sox gene family in teleost fish. PLoS ONE (PONE-D-17-05926R1)

Volff J-N (2006) Turning junk into gold: domestication of transposable elements and the creation of new genes in eukaryotes. BioEssays 28:913–922

Volff J-N (2009) Cellular genes derived from Gypsy/Ty3 retrotransposons in mammalian genomes. Ann NY Acad Sci 1178:233–243

Volff JN, Selz Y, Hoffmann C, Froschauer A, Schultheis C, Schmidt C, Zhou Q, Bernhardt W, Hanel R, Böhne A, Brunet F, Ségurens B, Couloux A, Bernard-Samain S, Barbe V, Ozouf-Costaz C, Galiana D, Lohse MJ, Schartl M (2013) Gene amplification and functional diversification of melanocortin 4 receptor at an extremely polymorphic locus controlling sexual maturation in the platyfish. Genetics 195:1337–1352

Wang W, Zhang J, Alvarez C, Llopart A, Long M (2000) The origin of the Jingwei gene and the complex modular structure of its parental gene, *yellow emperor*, in *Drosophila melanogaster*. Mol Biol Evol 17:1294–1301

Wang Y, Gu X (2000) Evolutionary patterns of gene families generated in the early stage of vertebrates. J Mol Evol 51:88–96

Wang Z, Benoit G, Liu J, Prasad S, Aarnisalo P, Liu X, Xu H, Walker NP, Perlmann T (2003) Structure and function of Nurr1 identifies a class of ligand-independent nuclear receptors. Nature 423:555–560

Wegner M (1999) From head to toes: the multiple facets of Sox proteins. Nucleic Acids Res 27:1409–1420

Wolfe K (2000) Robustness—it's not where you think it is. Nat Genet 25:3–4

Wolfe KH, Shields DC (1997) Molecular evidence for an ancient duplication of the entire yeast genome. Nature 387:708–713

Woods IG, Wilson C, Friedlander B, Chang P, Reyes DK, Nix R, Kelly PD, Chu F, Postlethwait JH, Talbot WS (2005) The zebrafish gene map defines ancestral vertebrate chromosomes. Genome Res 15:1307–1314

Yang S, Arguello JR, Li X, Ding Y, Zhou Q, Chen Y, Zhang Y, Zhao R, Brunet F, Peng L, Long M, Wang W (2008) Repetitive element-mediated recombination as a mechanism for new gene origination in Drosophila. PLoS Genet 4(1):e3

Zhang J, Dean AM, Brunet F, Long M (2004) Evolving protein functional diversity in new genes of Drosophila. Proc Natl Acad Sci USA 101:16246–16250

Zwick E, Bange J, Ullrich A (2001) Receptor tyrosine kinase signaling as a target for cancer intervention strategies. Endocr Relat Cancer 8:161–173