Pierre Pontarotti *Editor*

Evolution, Origin of Life, Concepts and Methods

Springer

Evolution, Origin of Life, Concepts and Methods

Pierre Pontarotti
Editor

Evolution, Origin of Life, Concepts and Methods

Springer

Editor
Pierre Pontarotti
CNRS and IHU
Aix-Marseille University
Marseille, France

ISBN 978-3-030-30362-4 ISBN 978-3-030-30363-1 (eBook)
https://doi.org/10.1007/978-3-030-30363-1

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

Preface

For the twelfth year, we publish a book on evolutionary biology concept and application.

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

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

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

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

The articles are organised in the following categories

Origin of Life (Chaps. 1–3)
Concepts and Methods (Chaps. 4–10)
Genomic Phenotype Evolution (Chaps. 11–15)

June 2019

Pierre Pontarotti
AEEB and CNRS
Marseille, France

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, ECCOREV Federation, Conseil Départemental 13, ITMO, Ville de Marseille.

We wish to thank the AEEB team for the organisation of the meeting.

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

Contents

Part I
Origin of Life

Chapter 1
A Non-paradoxical Pathway for the Chemical Evolution Toward the Most Primitive RNA-Based Life-like System

Kunio Kawamura

Abstract Although the RNA world hypothesis is an important hypothesis for the origin-of-life study, it involves drawbacks that should be evaluated. These drawbacks involve unknown points. First, several steps from inorganic materials to the functional RNA molecules are not yet clarified. Second, the simulation experiments for the prebiotic accumulation of RNA seem to be incompatible with the Hadean Earth environments. Third, the actual feature of the RNA world has not been identified. Here, we carried out possible simulation experiments for the chemical evolution of RNA using our hydrothermal flow reactor systems. We recently proposed the two-gene hypothesis for the emergence of life-like systems from simple chemical networks. Following the same methodology, here, we attempted to combine the knowledge obtained from experimental data on the chemical evolution of RNA and the theoretical work to deduce a realistic feature of the RNA-based life-like system.

1.1 Importance of the RNA World Hypothesis for the Origin-of-Life Problem

The RNA world hypothesis is important for the origin-of-life study. It provided a concept of how simple chemical networks could have evolved to a most primitive life-like system that we consider alive. The origin-of-life problem continues to exist as a great scientific question, and many scientific studies have been continuously undertaken, adopting different approaches. The origin-of-life problem has been well accepted as a practical scientific problem through the achievements by Pasteur, Oparin, and Miller (Pasteur 1861; Oparin 1924; Miller 1953). After Miller's experiment, various experimental approaches were proposed to investigate the structures and functions of biomolecules formed under primitive conditions by simulating the primitive Earth environments. The discovery of the double-helix structure of DNA by Watson and Crick (1953), in the same year when Miller's experiment was proposed (Miller 1953),

K. Kawamura (✉)
Department of Human Environmental Studies, Hiroshima Shudo University, Ozuka-Higashi 1-1-1, Asaminami-Ku, Hiroshima 731-3195, Japan
e-mail: kawamura@shudo-u.ac.jp

P. Pontarotti (ed.), *Evolution, Origin of Life, Concepts and Methods*,
https://doi.org/10.1007/978-3-030-30363-1_1

prompted the rapid development of molecular biology, leading to a strong influence on the origin-of-life study. For instance, the RNA world hypothesis was proposed in the stream of the development of molecular biology (Gilbert 1986). In addition, the geology and astronomy regarding the origin-of-life study have been developed in the last half-century.

The RNA world hypothesis visualizes the central question of how the system for genetic information that flows in modern organisms (Cech et al. 1981; Guerrier-Takada et al. 1983; Cech 1986; Joyce 2002; Orgel 2004) could have emerged in the primitive Earth environments (Crick 1970). The information flow involves a principle that the assignment between genotype and phenotype or the connection between information and function is essential characteristics in organisms (Eigen 1971; Eigen and Shuster 1979) (Fig. 1.1, top). Thus, different approaches including chemical evolution experiments, molecular biological approaches, theoretical approaches, and geological evidence would have been combined to find out the realistic feature of RNA by the principle of assignment between genotype and phenotype. According to the RNA world hypothesis, RNA molecules without any complicated translation system had simply maintained the assignment between genotype and

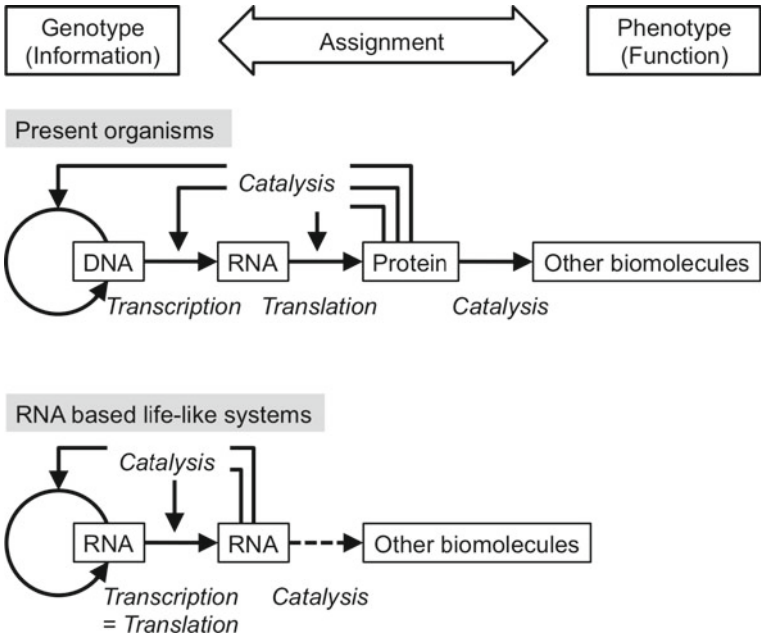


Fig. 1.1 Assignment between genotype and phenotype molecules in the present organisms and the RNA-based life-like system. Present organisms: The information preserved in DNA is used for the amino acid sequences of proteins via transcription and translation. Enzymatic reactions with proteins indirectly assign other biomolecules. RNA-based life-like systems: The information preserved in RNA is used for the production of functional RNA molecules by replication. Ribozymes may indirectly assign other biomolecules

phenotype molecules (genetic information and function) (Nemoto and Husimi 1995; Nemoto et al 1997). The assignment method of genotype and phenotype molecules in modern organisms is that the machinery is involved in each cell. According to a simplified view of information flow in the organism, the genetic information of modern organisms is preserved as DNA sequences. RNA molecules copy the DNA sequences and translate them into amino acid sequences on ribozyme (Fig. 1.1). Furthermore, the biomolecules at the downstream position of the information flow from DNA are indirectly assigned as a one-to-one correspondence by the enzymatic actions of proteins. Thus, it is regarded that the sum of the reactions in organisms is assigned by the DNA sequence directly and indirectly.

The assignment method between the genotype and phenotype molecules in the RNA-based life-like systems should be the simplest because RNA molecules play both the functions of genotype and phenotype molecules (Fig. 1.1, bottom). This is the central principle of the RNA world hypothesis. The difference between modern organisms and the plausible RNA-based life-like systems is quite large. Therefore, the uncertainty on how the RNA-based life-like system could have evolved to the modern method including the complicated assignment mechanism between DNA sequence and amino acid sequence of proteins is currently an important issue. The gap between the assignment methods between genotype and phenotype in the RNA-based life-like system and that in modern organisms involves the emergence of tRNA, rRNA, and aminoacyl-tRNA synthetase.

The RNA world hypothesis indicates that RNA or RNA-like molecules should have played a central role in the emergence of the most primitive life-like systems from a simple chemical network to a living system (Kauffman 1986, 2007; Copley et al. 2007; Wu and Higgs 2009; Vaidya et al. 2012; Vasas et al. 2012; Nghe et al. 2015; Higgs and Lehman 2015; Kim and Higgs 2016). As a most probable system of the RNA world hypothesis, one can suppose that a life-like chemical network consisting of RNA and other molecules should have emerged toward the most primitive life-like systems, of which RNA molecules play roles as both the genotype and phenotype molecules in the system (Fig. 1.1, bottom).

Conclusively, the method of modern organisms for assignment between genotype and phenotype molecules is very complicated. Thus, such a system hardly appears by a single or a small number of events of chemical evolution under the primitive Earth. The appearance of the principle of the assignment method between genotype and phenotype is very important for the origin-of-life study. Thus, the RNA world and following steps of chemical evolution to modern organisms became the main target of the origin-of-life study. In this chapter, the perspective views of the RNA world (Kawamura 2012a, b, 2016), our hypothesis on the emergence of life-like systems from two genes (Kawamura 2016), and the experimental approaches for chemical evolution under the Hadean Earth environments, (Kawamura 2011, 2017) are summarized. This would provide insight into a real feature of the RNA-based life-like system.

1.2 Drawbacks Regarding the RNA World Hypothesis

1.2.1 RNA-Based Life-like System Involving Both Information and Metabolism Machineries

The RNA world hypothesis is a strong hypothesis to approach the origin-of-life problem. At the same time, this hypothesis involves several drawbacks (Kawamura 2012a). The concept of the RNA world hypothesis is based on the fact that the function of RNA molecules comprises both preserving genetic information and enzymatic functions. It is considered that the RNA-dependent RNA polymerase ribozyme (RP ribozyme) is a key ribozyme (Johnston et al. 2001; Zaher and Unrau 2007; Horning and Joyce 2016). Concurrently, different kinds of functional RNA molecules should have been present for the construction of the RNA-based life-like systems. However, it is unclear what kinds of functions were necessary to make a life-like system.

RP ribozymes possessing high fidelity are normally focused upon (Johnston et al. 2001; Zaher and Unrau 2007; Horning and Joyce 2016). However, the system consisting of the ideal RP ribozyme is not regarded to perform further chemical evolution to higher stages because the mutation of RNA molecules does not occur due to the ideal replication (Fig. 1.2). This view about a plausible primitive RP ribozyme with mutation is important since the mutation should have emerged automatically in the primitive RP ribozyme consisting of actual RNA molecules. The mutation is an essential function for constructing life-like systems on the basis of the two-gene hypothesis. It is reasonable that the primitive RP should have possessed very low fidelity as compared to reproduction in modern organisms. Thus, the RNA molecules should have

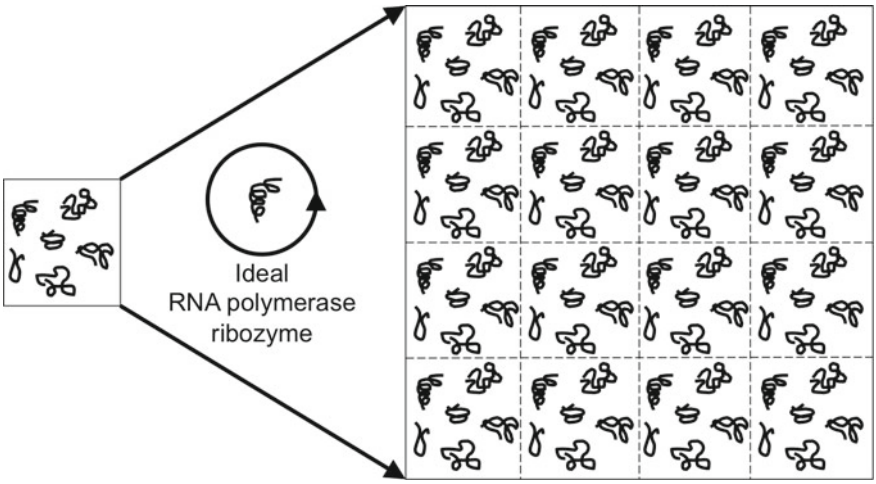


Fig. 1.2 An ideal RNA polymerase ribozyme. An ideal RNA ribozyme only produces the perfect copy of the population of RNA molecules. This does not result in the mutation of the RNA molecules. Mutation is an essential function for the CCSI, which is described in the later section

resulted in a variety of mutations during the replication of RNA molecules with the RP ribozyme. Conclusively, although the RP ribozyme should have been essential for the emergence of the RNA-based life-like systems, it is not yet clear what kinds of functional RNA molecules were necessary. Thus, the next step after the emergence of RP ribozyme should be verified to the life-like system from such simple chemical networks.

1.2.2 Compartmentalization for the RNA-Based Life-like System

The RNA-based life-like systems can be categorized into two types according to compartmentalization (Fig. 1.3). The first type, the RNA molecules are considered minimum elements of a life-like system. The life-like system consists of several RNA molecules displaying biological functions including genetic information and catalytic functions (Fig. 1.3, left). Eigen's hypercycle model is considered a model of this type of life-like system, where the members of RNA molecules act like living organisms composing quasi-species (Eigen 1971). Naturally, the importance

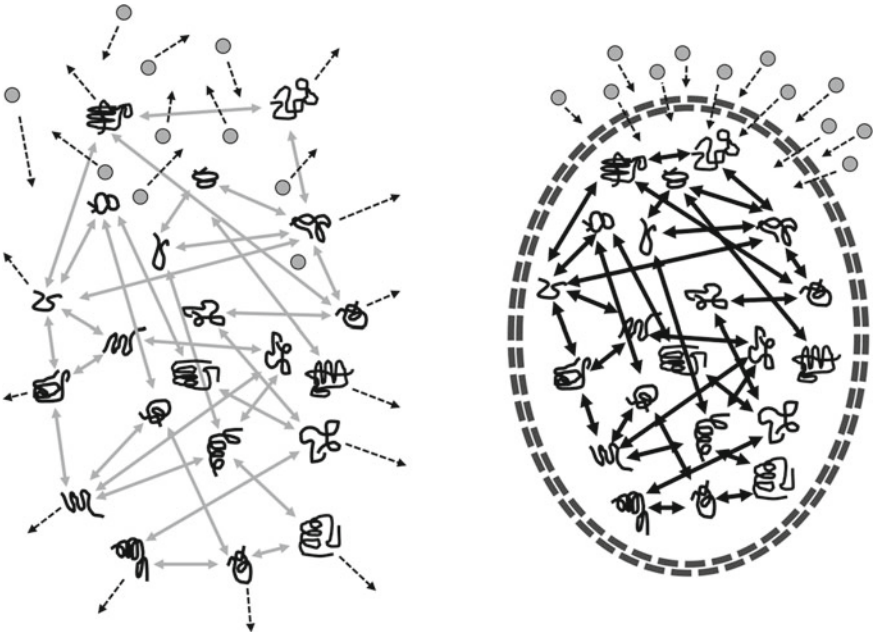


Fig. 1.3 Importance and advantages of compartment for the RNA-based life-like system. Left: a system without compartment, right: a system with compartment. Molecules hardly interact with each other without compartment and dissipate to distant areas hardly interact

of a compartment was also evaluated based on the hypercycle model (Eigen and Shuster 1979). The hypercycle type system is considered advantageous for evolution in higher stages connecting to protein-like molecules (Nemoto and Husimi 1995; Nemoto et al 1997). The hypercycle model was proposed many years before the RNA world hypothesis was proposed. Naturally, other types of biomolecules would be helpful in such systems. Thus, as a second type of life-like system, RNA molecules and other biomolecules are packed in the compartments and are categorized (Fig. 1.3, right). The importance of compartment was evaluated from both the theoretical and experimental points of view (Hogeweg and Takeuchi 2003; Szostak et al. 2016; Joyce and Szostak 2018). These two systems may be regarded as similar when we describe the importance of compartments with respect to the RNA-based life-like systems (Kawamura 2012b, 2016). If there were no compartments, the biomolecules would hardly encounter other molecules under the primitive Earth environments. On the contrary, the molecules in the compartment easily encounter other molecules. Thus, the compartment should have been very powerful to keep the biomolecules as a unit in the system, resulting in reactions among the molecules.

1.2.3 Importance of Additional Functions to the Self-replication of RNA

The features of the RNA-based life-like systems and the connection pathways from the RNA-based life-like system to the modern system are unclear. To solve these problems, we focused on what were the essential functions and how many functions were necessary for the construction of an RNA-based life-like system.

There is no doubt that the characteristic potential of RNA molecules should have been important to form a most primitive life-like system. However, as mentioned, the RNA-based-life-like systems in the presence of only the ideal RP ribozyme would not have resulted in further chemical evolution in most primitive organisms. Conclusively, the emergence of the RP ribozyme is important, but not enough for the continuous chemical evolution of the RNA systems. The whole process to the spontaneous formation of RP from nucleotide monomers under prebiotic conditions is not yet clear. Presumably, primitive RP ribozymes would have been produced spontaneously as the model RP ribozymes can be constructed by using the in vitro selection engineering method of RNA (Johnston et al. 2001; Zaher and Unrau 2007; Horning and Joyce 2016). Although the in vitro selection methods involve the analogical nature of Darwinian evolution, this needs to be molecular biological materials under the careful control of the artificial reaction conditions. The hypercycle-based theories for the emergence of life-like systems would answer to a part of the processes from a simple replication by the presence of the RP ribozyme. Some experiments based on molecular biological techniques for the evolution of chemical networks have been successfully carried out using simple members of chemical networks. These have shown the characteristics of the evolution of chemical networks (Kaufman 1986,

2007; Hogeweg and Takeuchi 2003). However, it should be noted that a lack of evidence that these experiments are normally constructed by the energy supply of activated building blocks. For the actual emergence process of life-like systems, the process of how energy supplying methods to the metabolism system was incorporated into the system should be clarified. The activation groups or the condensation reagents should have played a role in supplying energy for the phosphodiester bond formation.

The relationship between the energy supply to the life-like system and the information flow including the replication process is not easily investigated from the theoretical side of the origin-of-life problem (Copley et al. 2007; Saladino et al. 2012; Neveu et al. 2013). Our recent hypothesis that the connection between replication and metabolism transforms the life-like system from a simple chemical network of building blocks to a hierarchical level of chemicals (Kawamura 2016). This hypothesis possesses the potential to clarify the emergence of RNA-based life-like systems.

1.3 Two-Gene Hypothesis

The two-gene hypothesis on the origin of life was proposed from the observation of biosystems at different hierarchical levels (Kawamura 2002b, 2003c, 2007). Here, the two-gene hypothesis of the origin of life will be briefly described. This principle is proposed from a different viewpoint than the principles regarding the hypercycle and the RNA world. We proposed the importance of comparative analysis of the characteristics of biosystems at different hierarchical organism levels including eukaryotes, prokaryotes, social insects, ecosystems, species, civilizations, viruses, and viroids. These analyses are useful to identify the essential properties to see if a biosystem is regarded as alive. Thus, the analyses also gave us a hint of the stepwise emergence of a higher level of biosystems from a lower level to beyond the hierarchical levels. The analysis is briefly summarized as follows.

Biosystems are considered the system including organisms, the organism itself, and/or the system that is organism like, where the biosystems consist of its building blocks. The biosystems can be classified into roughly two types according to the following characteristics. This classification is based on whether the system possesses or not the central controlling system for information (CCSI) and the central controlling machinery for inflow/outflow, formation/degradation of energy, material, and information from environments (CMIO). Naturally, eukaryotes and prokaryotes including single-celled and multicellular organisms are classified into this type. This is because they possess inherent CCSI and CMIO at the hierarchical level, at which the system is regarded alive (Fig. 1.4). On the other hand, for instance, the ecosystem may be different from this type of biosystems with regard to the characteristics of CCSI and CMIO. The presence of CCSI and CMIO in civilizations as a biosystem supports the importance of these biosystems.

Generally, CCSI involves the function of replication (or amplification) of information, the function of modification of information, the function of incorporation

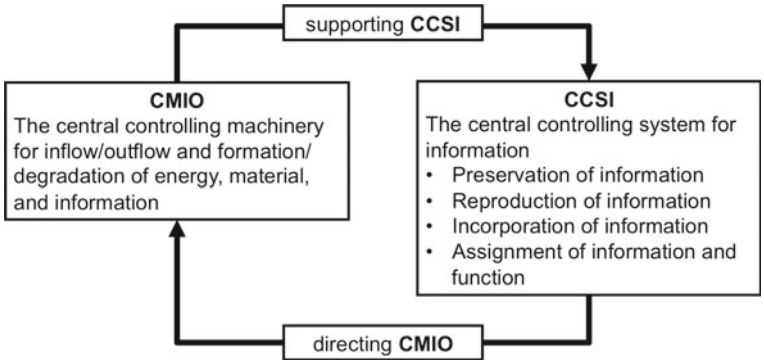


Fig. 1.4 Relationship between the CCSI and CMIO. CCSI is the central controlling system for information, and CMIO is the central controlling machinery for inflow/outflow and formation/degradation of energy, material, and information from environments. CCSI involves four kinds of functions, preservation of information, reproduction of information, incorporation of information, assignment of information and function. It is assumed that the RNA-based life-like system should have involved spontaneously these functions. The figure was modified based on our previous publication (Kawamura 2016)

(or addition) of information to the system, and the function of assignment between information and actual functions in the system (Table 1.1). According to the RNA world hypothesis and the experimental evidence of chemical evolution, it can be assumed that these functions were maintained by RNA molecules. It is likely that if the RP ribozyme was present, this would automatically involve these functions regarding information because the genotype and phenotype are identical. In short, an RP ribozyme enhances the formation of RNA molecules so the RNA population can be amplified by repetitive replication with the RP ribozyme. In addition, the primitive RP ribozyme would readily undergo mutations during replication due to the low fidelity, facilitating new information incorporated automatically into the system

Table 1.1 Characteristics of the CCSI for the RNA-based life-like system

Keywords	Description for the plausible system
Preservation of information	The double-helix structure of RNA keeps the stable Watson–Crick base pairing
Reproduction of information	Replication of double-helix RNA with the primitive RNA polymerase ribozyme
Incorporation of information	Mutation of RNA sequences occurs during the replication of double-helix structure RNA with the low fidelity of replication of RNA
Assignment between genotype and phenotype	RNA molecules are replicative. The RNA molecules possess biochemical functions including enzymatic functions

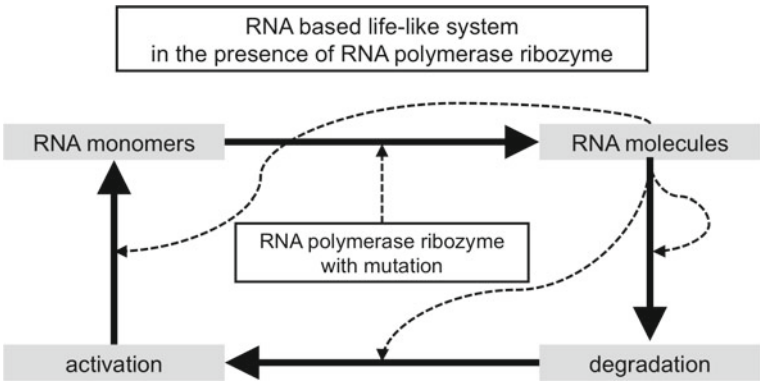


Fig. 1.5 Simplified metabolic pathway for the continuous chemical evolution of RNA. RNA molecules are reproduced and amplified with the mutations in the presence of RNA polymerase ribozyme (RP ribozyme). A ribozyme that enhances the pathway affecting the formation of RNA molecules could have formed by a random mutation from the RNA population in the presence of RP ribozyme in the CCSI. This improves the efficiency of the system. How the efficiency feeds back for enhancing the quality of the system is illustrated in Fig. 1.13. The figure was modified based on our previous publication (Kawamura 2016)

during replication. Furthermore, the genotype and phenotype in RNA molecules are the same, so that incorporation and assignment are automatically accomplished.

In contrast, the actual activity of CCSI should be supported by CMIO, which is connected to the imaginable cyclic chemical network from formation to deformation of the RNA population. Once a new ribozyme, which can enhance the metabolic pathway connecting with the RNA replication, appears by random mutation of RNA molecules by CCSI, this improves the activity of CCSI by feedback from CMIO (Fig. 1.5). The enhancement of CCSI gives feedback to the activity of CMIO, and the enhancement of CMIO gives feedback to the activity of CCSI. On account of this relationship between CCSI and CMIO, we proposed a hypothesis that the connection between CCSI and CMIO would have provided a starting point for the emergence of biosystems at the higher hierarchical level beyond the building blocks. In the case of the most primitive life-like systems, individual RNA molecules correspond to the building block and the life-like system consisting of RNA molecules, indicative of a system involving CCSI and CMIO. The connection between CCSI and CMIO would start with the event with the lowest efficient pathway in the CMIO being enhanced by the ribozyme that emerged by random mutation in the RNA population of CCSI. If the slowest path (lowest efficient path) is accelerated beyond the second slowest path, this gives feedback to the CCSI (Fig. 1.6). Thus, the second slowest path becomes the slowest path. Subsequently, repetitive step-by-step accelerations of this circular pathway by gene mutations in CCSI could have improved the whole metabolic circular network. This principle seems to be consistent with the fact that modern enzymatic reaction rates are controlled at a narrow range of reaction rates for both regular enzymes and thermophilic enzymes (Fig. 1.7). Thus, the initial acceleration

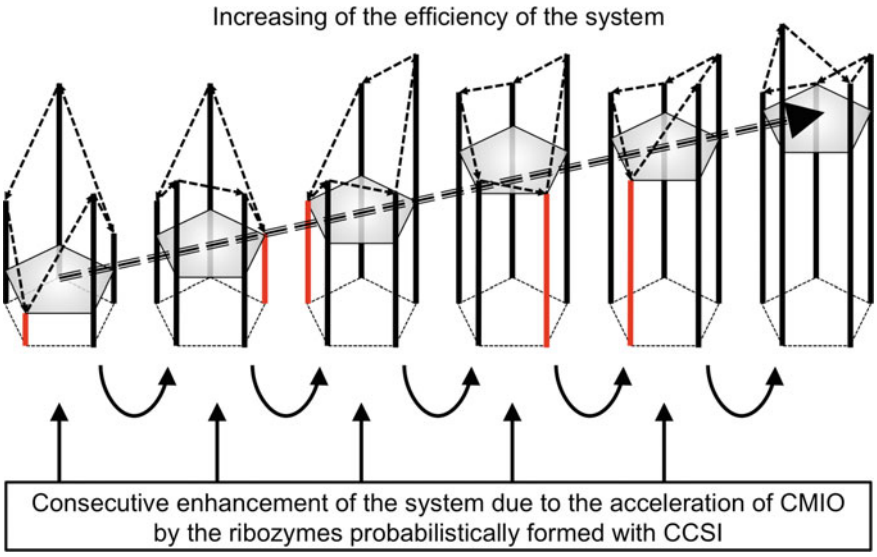


Fig. 1.6 Mechanism for improvement of the life-like system. If a ribozyme, which enhances the slowest path (shown as the red bar at the leftmost) among the circular processes connecting to the CCSI, formed by a random mutation in the CCSI, the efficiency of the system is improved. At the next step, the second slowest path (shown as the red bar at second leftmost) among the circular processes connecting to the CCSI, formed by a random mutation in the CCSI, the efficiency of the system is again improved. Repetitive step-by-step enhancement in the circular processes could have improved the whole efficiency of the system. The figure was modified based on our previous publication (Kawamura 2016)

by a ribozyme, which was formed by random mutation of an RP ribozyme, served as the starting point for the connection between CCSI and CMIO. Hence, this principle was named as the two-gene hypothesis for the origin of life. The study on the relation between the two-gene hypothesis and other works regarding the origin-of-life theory, including the hypercycle, would be important for its evaluation.

According to the two-gene hypothesis, the role of the compartment for the RNA-based life-like systems is important because the compartment enhances the connection between CCSI and CMIO. The life-like systems that possess both the CCSI and CMIO seem to also possess the boundary system that is controlled by CCSI and CMIO. Thus, the two-gene hypothesis may imply that a life-like system would have possessed the boundary system before it acquired the connection between CCSI and CMIO.

Here, we added that the principle of the two-gene hypothesis should be applicable to the emergence of higher levels of historical events in biological evolution, such as the emergence of prokaryotes, multicellular organisms, and societies of organisms.

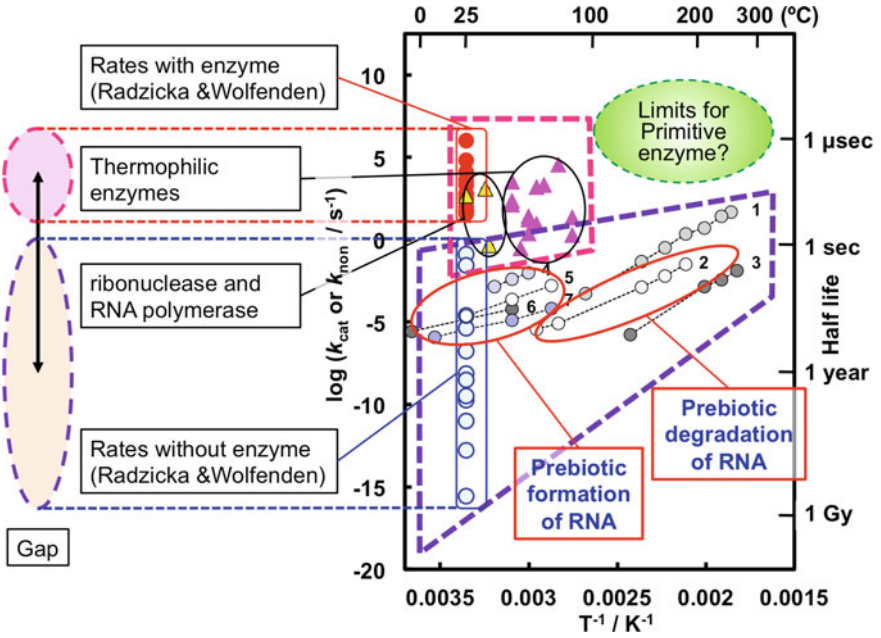


Fig. 1.7 Comparison of the reaction rate with enzymes and without enzymes regarding prebiotic reactions. The horizontal axis indicates inverse values of temperature (T^{-1}), the vertical axis left indicates logarithmic values of reaction rates, and the vertical axis right indicates the half-life time of the reactions. The large difference between the enzymatic and non-enzymatic reaction rates indicates the proficiency of enzymes (Radzicka and Wolfenden 1995). This relationship was attempted to extend to thermophilic enzymes (Kawamura 2004, 2010). The enzymatic rates for both the enzymes worked at 25 and 50–90 °C are positioned to a similar region of the rate constant of the graph. The presence of the gap between the reaction rates with and without enzyme is a required condition for the enzyme. The space around 300 °C assumed from the extrapolation of the rate constants at lower temperatures may indicate a possible condition for the emergence of enzymes. The figure was modified based on our previous publications (Kawamura 2004, 2010). Closed circles in red and open circles in blue at 25 °C: the rate constants for enzymatic and non-enzymatic reactions (Radzicka and Wolfenden 1995 and references therein); closed triangles in yellow at 25–37 °C for the rate constants for ribonucleases (Kawamura 2004 and references therein); closed triangles in pink at 50–90 °C for the rate constants for several types of thermophilic enzymes (Kawamura 2004, 2010; and references therein); circles numbered 1–7 with dot lines at 65–315 °C: 1: ATP hydrolysis, 2: $C^{3'}$pG cleavage, 3: racemization of alanine, 4: 4-mer oligoguanylate formation by the template-directed reaction, 5: cyclization of d(pGCGCG)rC, 6: 4-mer oligocytidylate formation by the clay-catalyzed reaction, 7: 3-mer formation by the Pb^{2+}-catalyzed reaction (Kawamura 2004, 2010; and references therein). Top-right corner (green circle) would indicate the limit temperature and enzymatic reaction rate regarding the origin of life

1.4 Experimental Techniques Simulating the Hadean Earth Environments

1.4.1 Compatibility with the Hadean Earth Environments

These prebiotic simulation experiments are valuable to consider the realistic pathways for the formation of RNA-based life-like systems (Kawamura 2012b). However, the question of whether the simulation experiments for the chemical evolution of RNA molecules are compatible with the Hadean Earth environments could be deduced from investigations in the fields of astronomy, geophysics, and geochemistry. However, the realistic feature of the Hadean Earth environments still involves unknown characteristics including very fundamental conditions, such as temperature and pressure, the presence of oceans, the acidity, and the chemical composition of the atmosphere (Gough 1981; Kasting 1993; Kawamura 2004; Maruyama et al. 2013). These conditions are dependent on the age of the Earth between 4.6 and 3.8 Gya. These 0.8 billion years, between 4.6 and 3.8 Gya, are regarded as a very long time for chemical evolution so that it necessitates the identification of events that happened within this long period. For instance, there are the oldest pieces of evidence discovered, including the liquid water at around 4.3 Gya (Mojzsis et al. 2001; Harrison 2009) and an organism at around 3.8 Gya (Mojzsis et al. 1996; van Zuilen et al. 2002; Nutman et al. 2016).

The primitive Earth environments should have been more extreme environments as compared with the present Earth. For instance, this assumption supports the hypothesis that the last universal common ancestor (LUCA) for all the present organisms should have possessed similar characteristic of hyperthermophilic organisms, although the characteristic of LUCA is still disputed (Corliss et al. 1981; Pace 1991; Galtier et al. 1999; Schwartman and Lineweaver 2004; Akanuma et al 2013). LUCA may have been a survivor from the Late Heavy Bombardment (Sleep 2018). At the same time, the RNA molecules had been considered to be unstable against the extreme Hadean Earth environments (White 1984; Larralde et al. 1995; Kawamura 1998, 1999, 2000). However, sufficient data are not yet obtained to evaluate whether the RNA world hypothesis is compatible with such extreme conditions (White 1984; Larralde et al. 1995) before our group started investigations of reaction behaviors of biomolecules including RNA and its moieties under hydrothermal conditions (Kawamura et al. 1997, Kawamura 1998, 1999, 2000). There was one problem, which was that efficient research tools were not present for investigation of these reaction behaviors in aqueous solutions at high temperature and pressure. The investigations of aqueous phase reactions at high temperatures were normally carried out by using the conventional batch reactors (Larralde et al. 1995). The conventional batch methods were not useful for monitoring rapid reactions and behaviors of biomolecules at high temperature and pressure.

Also, different investigations regarding the chemical evolution of RNA including the prebiotic formation of RNA and its moieties under the Hadean Earth environments, the conventional RNA formation using the activated nucleotides, in vitro

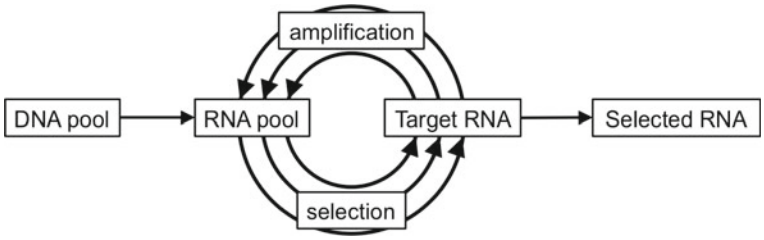


Fig. 1.8 In vitro selection for artificial functional RNA molecules. The method involves selection and amplification with the mutation. This is analogical to the Darwinian evolution. If this type mechanism consisting of totally prebiotic materials worked under the Hadean Earth environments, it is strongly assumed that the population of RNA molecules had evolved to a life-like system

selection of functional RNA molecules by the modern molecular biology techniques were indeed carried out at very mild conditions. For instance, the in vitro selection techniques for functional RNA molecules are considered as an analogy of Darwinian evolution of RNA molecules (Ellington and Szostak 1990; Tuerk and Gold 1990) (Fig. 1.8). Another example of the successful prebiotic RNA formation model, such as the template-directed formation of RNA and the clay-catalyzed RNA formation, is supposed to simulate the RNA formation at very mild conditions near the modern Earth conditions (Ferris and Ertem 1992; Kawamura and Ferris 1994). Thus, it is necessary to evaluate whether these pieces of evidence would have been compatible with the Hadean Earth environments.

1.4.2 Hydrothermal Flow Reactor Systems

The discovery of thermophilic organisms and the phylogenetic tree analysis suggested that the ancient organisms could have possessed the characteristics of thermophilic organisms and LUCA should have possessed characteristics of the hyperthermophilic organisms. However, one may imagine that RNA molecules are not so stable under such extreme conditions, which would be incompatible with the Hadean Earth environments (Kawamura 2011, 2012a, 2016). The extreme environments of primitive Earth should involve a variety of environments, such as a wide range of temperature and pressure, acidity of the ocean, and strength of irradiation energy from the sun.

Imai's group has independently designed with different concepts an experimental setup with hydrothermal environments similar to the origin of life (Imai et al. 1999a). Independently, we have successfully developed hydrothermal flow reactor systems from 1997, which enable monitoring chemical reactions at high temperatures up to 400 °C at a pressure up to 30 MPa within a very short time scale at 0.002–200 s (Kawamura 1998, 1999, 2000). For instance, we started to develop a monitoring method for hydrothermal reactions within a short time scale, which can be used for

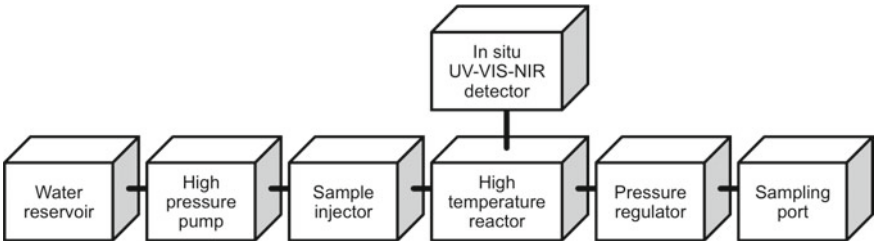


Fig. 1.9 Components of the hydrothermal flow reactor systems. Our groups and others developed a series of hydrothermal flow reactor systems (Kawamura 1998, 1999, 2000; Imai et al. 1999a). The flow system consists of a water reservoir, a high-pressure pump, a sample injector, a high-temperature reactor, a pressure regulator, and a sampling port. An in situ UV-VIS-NIR detector can be attached to the system

kinetic measurements at the millisecond to second time scale at high temperatures up to 400 °C (Fig. 1.9). An experimental setup simulating the circular dynamics of hydrothermal flow in the submarine vent systems was then developed by Imai and coworkers (Imai et al. 1999a). Both the concepts of flow reactors have been useful as research tools for the origin-of-life studies. The importance of the experimental improvement was realized to apply the hydrothermal flow reactors for the chemical evolution of biomolecules. By the proposal of these methods, different and/or modified methods had been developed (Islam et al. 2003; Cleaves et al. 2009). These hydrothermal reactor systems are based on the concept that the flow reactor systems consist of narrow tubing reactors instead of the traditional static reactor vessels. Our group developed different types of hydrothermal flow reactor systems, which are designed for mainly kinetic measurements and in situ absorption spectrophotometric measurements (Kawamura 2002a, 2011, 2017; Kawamura et al. 2010). In addition, the influence of solid-state mineral particles can be observed using the hydrothermal flow reactors (Kawamura et al. 2011, 2016, 2017). In addition, the flow reactor systems for simulating the circulation in hydrothermal vent systems sometimes include a relatively large volume of the reactor vessel (Imai et al. 1999a).

1.4.3 Other Methods on Extreme Conditions

The experimental conditions on the aqueous phase reactions are important for the evaluation of the chemical evolution of biomolecules. However, the term “extreme” does not only indicate high temperature and pressure, but also the primitive extreme Earth environments such as strong acidity, irradiation, and meteorite impacts. Studies have also attempted to simulate some of these environments. High-energy sources such as cosmic rays and meteorite impacts would be useful for the formation of simple organic molecules rather than the oligomer formations of RNA and peptides. It would be reasonable that these high-energy sources likely degraded these molecules.

Conversely, the exposure of different energy levels of electromagnetic radiation, such as X-ray, γ-ray, and UV light, is also regarded as an extreme condition for chemical evolution. The accumulation of these techniques would provide insight into the details of chemical evolution under the Hadean Earth environments (Blank et al. 2001; Boillot et al. 2002; Gontareva et al. 2009; Furukawa et al. 2009; Otake et al. 2011; Kaiser et al. 2013; Martins et al. 2013; Nagafuchi et al. 2013). These extreme conditions would provide high-energy conditions to form simple biomolecules. Also, the dehydration reactions including oligomer formation sometimes proceed more efficiently rather than those in an aqueous solution. For instance, it was elucidated that short oligopeptides could have formed efficiently with meteorite impacts, which were performed using a specialized experimental setup (Furukawa et al. 2009; Otake et al. 2011). These facts support that instrumentation is an important approach for origin-of-life studies.

1.5 Chemical Evolution of Proteins and RNA Under the Hadean Earth Environments

1.5.1 Protein-like Molecules

Here, we briefly consider proteins from the viewpoint of the RNA world hypothesis (Kawamura 2016). The term "proteins" involves the molecules formed by organisms. To clarify this situation, we proposed a definition that proteins from the viewpoint of chemical evolution to the most primitive life-like system are biomolecules, of which the corresponding DNA sequences dictate the amino acid sequences. Thus, the long peptides formed under the primitive Earth environments are not regarded as proteins unless informational molecules, such as DNA assigned the amino acid sequences of these molecules. Proteins do not possess the self-replication function although there are some abiotic examples that special sequences of peptides are possible to replicate (Lee et al. 1997; Yao et al. 1998). If the protein-like molecules could have possessed weak replication characteristics to form copies of the protein-like molecules, which was hypothetically present before the modern organisms, this hypothesis would be exposed to the same question: How such a primitive protein-based system would have evolved to the modern system (Ikehara 2005, 2009; Maury 2009)? Thus, an argument that the difficulty for solving the connective pathway between the RNA-based life-like systems to the modern systems is evidence to deny the RNA world hypothesis is not correct. The hypothetical protein-like-molecule-based life-like system should have possessed an assignment method between genotype and phenotype at least if it was present before the modern system. Protein or protein-like molecules are considered as key molecules during the chemical evolution from the RNA-based life-like system to the most primitive organism.

The fact that amino acids, peptides, and proteins (Kawamura and Yukioka 2001; Kawamura et al. 2005b, 2009) are more stable as compared with nucleosides,

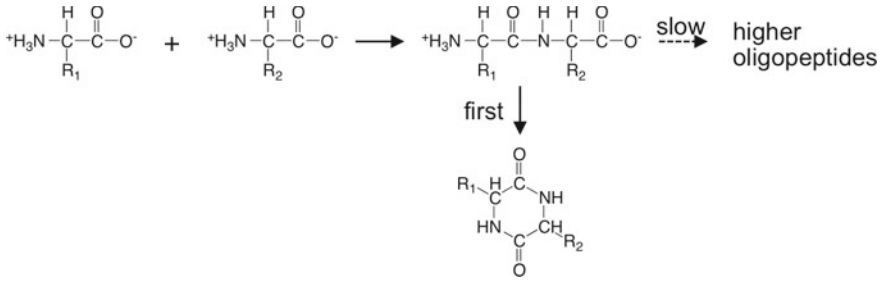


Fig. 1.10 Difficulty of oligopeptide formation due to diketopiperazines. Diketopiperazines are to stop further elongation of oligopeptides so normally the elongation of oligopeptides stops at the 2-mer formations. This results in low efficiency of oligopeptide formations

nucleotides, oligonucleotides, and functional RNA (Kawamura 2000, 2001, 2003a, b; Kawamura et al. 1997, 1999, 2004, 2005a; El-Murr et al. 2012) has been demonstrated using hydrothermal flow reactors. For the spontaneous formation of oligopeptides, it was shown that peptides could have formed under such extreme Earth conditions. However, the peptide formation found by the simulation experiments was not efficient, as the yields of oligopeptides remain 0.1–1% (Imai et al. 1999a, b; Islam et al. 2003; Cleaves et al. 2009). This is mainly due to the formation of diketopiperazine, which is a stable cyclic dimer of amino acids to stop the further elongation of oligopeptides (Fig. 1.10) (Kawamura et al. 2005b). Our studies using hydrothermal flow reactors demonstrated successful examples to bypass this pathway using tetramer, larger oligopeptides (Kawamura et al. 2005b), or using glutamic and aspartic acid (Kawamura and Shimahashi 2008) as starting oligopeptides. This reaction is also accelerated by carbonate minerals and clays, which are assumed to be present on the Hadean Earth. The process provides higher yields of oligopeptides, such as 30% of the total oligopeptides at a temperature 275 °C (Kawamura et al. 2011, 2018).

1.5.2 RNA

If the RNA world hypothesis is true, RNA or RNA-like molecules should have been accumulated under the Hadean Earth environments. Investigations on the prebiotic formation of RNA and its moieties have been carried out since around the 1960s before the time when the RNA world hypothesis was proposed (Oró 1961; Sanchez et al. 1966; Ferris et al. 1968; Crick 1968; Orgel and Crick 1993). The consecutive abiotic formation of RNA starting from the formation of nucleoside bases and ribose, nucleosides, nucleotides, and the oligomer formation of RNA has been evaluated by the experiments simulating the prebiotic Earth environments (Fig. 1.11) (Ponnamperuma and Mack 1965; Waehneldt and Fox 1967; Fuller et al. 1972a, b; Lorhmann and Orgel 1973; Orgel and Lohrmann 1974; Sawai 1976; Lohrmann 1977; Sawai et al. 1981; Inoue and Orgel 1982, 1983; Joyce et al. 1984; Ferris and Ertem 1992;

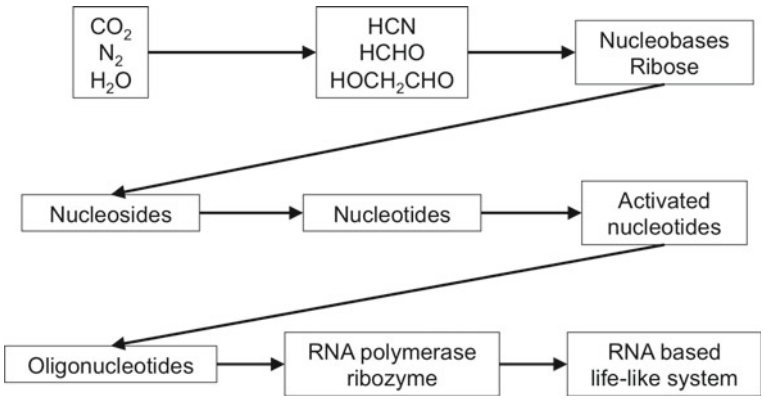


Fig. 1.11 Simplified chemical evolution from inorganic molecules to the RNA-based life-like system. Although some of these steps have been well investigated, the connection pathways among these steps are not identified

Kawamura and Ferris 1994; Ertem and Ferris 1996; Ferris et al. 1996; Ferris 2002; Orgel 2004). The stepwise chemical evolution of RNA is summarized in our previous publication (Kawamura and Maurel 2017). Based on extensive experimental studies regarding the chemical evolution of RNA, possible pathways for these steps from simple molecules to oligonucleotides have been gradually identified. However, the proposed pathways involve a question of whether these pathways were likely under the Hadean Earth environments (Kawamura and Maurel 2017).

For instance, we pointed out that some of the pathways such as the formation of ribose, the formation of nucleotides, and the replication of RNA are intricate. First, the foremose reaction is known from previous work as a ribose formation pathway. However, the foremose reaction under different conditions normally produces only small amounts of ribose (Shapiro 1988). Besides, the typical conditions for the foremose reaction comprise alkaline environments, which are not suitable for other steps in the chemical evolution of RNA molecules. The amount of RNA molecules involved in the present organisms is considerably high. Thus, efficient formation of ribose would have been necessary during the prebiotic chemical processes under primitive Earth environments. The presence of borate minerals enhances the selective formation of ribose (Ricardo et al. 2004).

Second, the formation of nucleosides from ribose and bases and nucleotides from nucleosides are also difficult processes (Fuller et al. 1972a, b; Yamagata et al. 1992; Reimann and Zubay 1999; Costanzo et al. 2007). These reactions include the dehydration processes during N-glycoside bond formation for nucleoside formation as well as phosphodiester bond formation for nucleotide formation. The dehydration reaction is challenging in aqueous solutions from a thermodynamic viewpoint. Thus, some specific dehydration processes, such as dry conditions, are frequently assumed to avoid this difficulty. As described in a later section, a dry–wet cycle on the Hadean Earth would have hardly occurred. It is assumed that the Hadean Earth surface was

probably covered with the ocean as the continent growth occurred after 4 Gya (Buick et al. 1995; Nutman et al. 2001; Harrison 2009) and the ocean was highly toxic (Maruyama et al. 2013; Santosha et al. 2017).

Third, the possible prebiotic conditions for replication of RNA are not yet identified although the prebiotic replication of RNA is the main rationale supporting the RNA world hypothesis (Hill et al. 1993; Kawamura 2012a). It is known that the prebiotic formation of RNA oligomers could have proceeded with activated nucleotide monomers under different conditions. Especially, clay mineral catalysts (Ferris and Ertem 1992; Kawamura and Ferris 1994; Ertem and Ferris 1996; Ferris et al. 1996) and metal ion catalysts (Sawai 1976; Sawai et al. 1981, 1989) are effective for the formation of oligonucleotides with 30–50-mers in length; such long RNA molecules are regarded as sufficient for displaying biological functions. Conversely, the prebiotic replication of RNA molecules has been studied from the 1960s to 1980s. These results showed that the guanosine 5′-phosphorimidazolide or guanosine 5′-phospho-2-methylimidazolide forms oligoguanylate with up to 40-nucleotide units in the presence of a polycytidylic acid template with the Watson–Crick-type complementary base pairing (Inoue and Orgel 1982, 1983; Joyce et al. 1984). This is called template-directed formation of oligonucleotides. However, the different combinations of Watson–Crick-type base pairing, such as the activated nucleotide monomer with adenine, cytidine, and uracil, do not work in the presence of complementary polynucleic acid templates (Hill et al. 1993; Kawamura 2012a). This is probably due to the π–π stacking interaction between nucleotide bases which does not work efficiently for the other cases of the combination of activated nucleotide monomers with the polynucleotide templates, resulting in an unstable double-helical structure (Kanavarioti et al. 1993; Kawamura and Ferris 1994; Kawamura and Umehara 2001; Kawamura and Maeda 2008). Different activated nucleotides promote the formation of oligonucleotides at elevated temperatures (Costanzo et al. 2009), and the hydration–dehydration cycle of nucleotides (Da Silva et al. 2015) provides efficiently long oligonucleotides without template polynucleotides. Efficient conditions for the replication of oligonucleotides in aqueous solutions remain an important issue in the RNA world hypothesis.

Spontaneous formation of RNA from nucleotide monomers under the simulated Hadean Earth conditions has been extensively studied (Sawai 1976; Lohrmann 1977; Sawai et al. 1981; Inoue and Orgel 1982, 1983; Joyce et al. 1984; Ferris and Ertem 1992; Kawamura and Ferris 1994; Costanzo et al. 2009; Da Silva et al. 2015). The suitable conditions for the steps from inorganic materials to functional RNA molecules in Fig. 1.11 are isolated and do not overlap, so there are gaps in the proposed chain under the Earth conditions (Stüeken et al. 2013; Kawamura and Maurel 2017). For instance, the oligomer formation from the activated nucleotide monomers in the presence of catalysts, such as clay minerals, metal ions, and a complementary polynucleotide template, proceeds normally under mild conditions or up to 37 °C (Fakhrai et al. 1984) (Fig. 1.12). As we pointed out the fact that the accumulation of such unstable molecules simply is not only determined by the stability of molecules, but also determined by both the relative rates of formation and decomposition (Kawamura 2012a, b, 2016). The accumulation would be possible within a

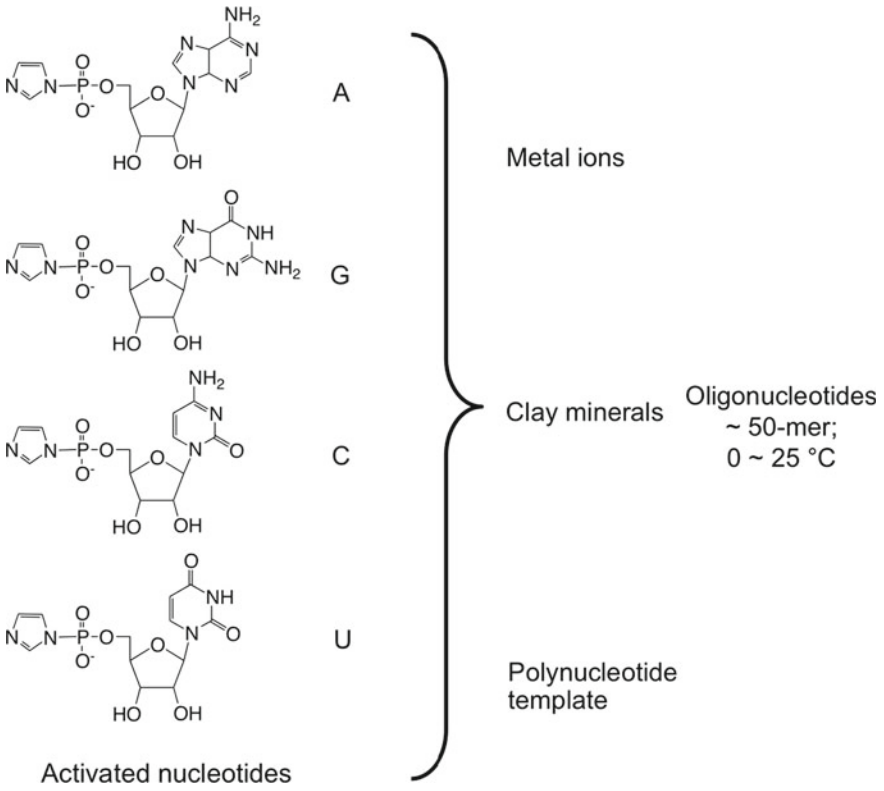


Fig. 1.12 Plausible prebiotic formation of oligonucleotides. The oligopeptide formation proceeds efficiently in the presence of metal ion catalysts, clay minerals catalysts, or polynucleotide template

very short time scale if the formation rate is faster than the degradation rate (Fig. 1.13). We showed the reason that the prebiotic formation does not proceed efficiently at higher temperatures by kinetic analysis (Kawamura and Umehara 2001; Kawamura and Maeda 2008). Also, the maximum temperature where the accumulation of RNA could have occurred was estimated as shown in the next section. The analysis implies that the rate of prebiotic formation of RNA from the activated nucleotide monomers, which is assumed as a prebiotic monomer (Lohrmann and Orgel 1973; Lohrmann 1977), becomes competitive to the degradation of RNA at temperatures around 300 °C (Kawamura 2004, 2010, 2012b; Kawamura and Maeda 2008).

However, the interactions of biomolecules, such as hydrogen bonding and hydrophobic interaction, become weak at high temperatures, posing another shortcoming in the RNA world hypothesis. For instance, the double-helical structures of DNA and RNA are normally melted at temperatures below 100 °C (Schildkraut and Lifson 1965; Wetmur and Davidson 1968). We have measured the stability of the double-helical structure at temperatures over 200 °C using the hydrothermal flow

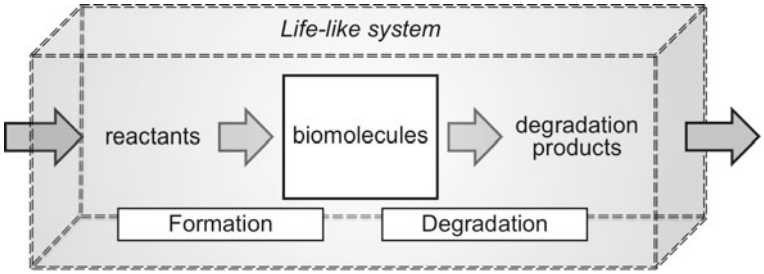


Fig. 1.13 Importance of relative rates for the formation and degradation of biomolecules for the accumulation of biomolecules. The accumulation of biomolecules is determined by both the rates for the formation and degradation of the biomolecules. Also, the accumulation is also determined by the inflow and outflow of the biomolecules to the system. The figure was modified based on our previous publication (Kawamura 2012a)

reactor system (Kawamura 2005; Kawamura and Nagayoshi 2007). The experiment clearly showed that these are biologically important, but weak interactions are not effective at high temperatures. Naturally, this situation is the same for protein-like molecules as the protein-like molecules hardly form three-dimensional structures at high temperatures (Kawamura et al. 2010). Additionally, some organisms grow or survive at temperatures over 100 °C (Cowan 2004; Pikuta et al. 2007; Harrison et al 2013). This fact reflects that the interactions of biomolecules are active and the condition that the formation rate should be faster than the degradation rate to accumulate the target molecules is properly controlled in the organisms by enzymes at higher temperatures over 100 °C.

Finally, it should be noted that the experimental approaches regarding the co-chemical evolution of RNA and protein-like molecules are not sufficiently progressed and the results are fragmentary (Landweber 1999; Kawamura et al. 2004, 2005a), although these are important. This may be due to the fact that the conditions for the chemical evolution of RNA and protein-like molecules are fairly different and the sample analysis of the mixed system of RNA and protein-like molecules becomes difficult. In addition, these experiments include long-term trial and error.

1.5.3 High-Temperature Origin of Life

The experimental measurements of stability, the biologically important interactions of RNA, and protein-like molecules deduced a maximum temperature for life. According to the experimental verifications of stability, prebiotic formation, and roles of biologically important interactions under hydrothermal environments, the temperature dependence for the chemical evolution is not simply understood. Here, we briefly introduce our estimation of the limit temperature for the emergence of enzymes (Kawamura 2004, 2010, 2012b). First, the accumulation of RNA and

protein-like molecules should have occurred if the formation rate is much faster than that of degradation at even high temperatures, as mentioned. Second, it has been noted that the primitive enzymes controlled these reactions in the life-like systems (Kawamura 2004, 2010, 2012b, 2016); naturally, primitive enzymes include primitive ribozymes. According to these two concepts, we compared the reaction rates for the formation and degradation of RNA molecules at high temperatures (Kawamura and Umehara 2001; Kawamura and Maeda 2008). Additionally, we compared the enzymatic reaction rates and corresponding non-enzymatic reaction rates at high temperatures. Conclusively, the emergence of primitive enzymes including ribozymes might have been possible at an extremely high temperature of approximately 300 °C (Fig. 1.7) (Kawamura 2010).

A general trend has been observed that the enzymatic reaction rates are much greater than those corresponding to the background rates without enzymes at normal temperatures (Radzicka and Wolfenden 1995) (Fig. 1.7, circles at 25 °C), and the enzymatic reaction rates are controlled in the narrow window of the reaction rate. The large difference between the enzymatic and non-enzymatic reaction rates allows the proficiency of enzymes of rapid reaction rates and selectivity. We analyzed the difference between the enzymatic and non-enzymatic reaction rates for thermophilic enzymes including ribonuclease and RNA polymerase (Fig. 1.7, circles and triangles at 50–90 °C), where non-enzymatic reaction rates were obtained from our corresponding kinetic data that examined using the hydrothermal studies and prebiotic formation of RNA.

This analysis showed that the gap between the reaction rates with and without enzymes is quite large, not only at the low temperatures but also at temperatures for thermophiles. The presence of the gap between the reaction rates with and without the enzyme is a required condition for the enzyme. The space at approximately 300 °C assumed from the extrapolation of the rate constants at lower temperatures may reflect a possible condition for the emergence of enzymes. This is also consistent with the highest temperatures where the formation and degradation rates are competitive, which were deduced from our hydrothermal experiments (Kawaura and Umehara 2001; Kawamura and Maeda 2008). The formation of RNA and its building blocks under such conditions and the measurements of biologically important interactions are not yet verified. Thus, hydrothermal flow reactor systems will be useful for such experiments.

1.6 Conclusions

Although the RNA world hypothesis presents several drawbacks, efforts to overcome some of these drawbacks have been attempted using research tools developed in different laboratories. The features of the RNA-based life-like system are unclear, although it is believed that the RP ribozyme is essential for the emergence of life. The two-gene hypothesis of the origin of life implies that an additional ribozyme, which accelerates the metabolic pathway, would have been necessary to bootstrap

the simple chemical network including RP ribozyme to a realistic life-like system consisting of a number of metabolic reactions. Furthermore, we highlighted that the improvement of research tools for the origin-of-life studies is an important key for further investigation of the RNA world hypothesis and the characteristics of RNA-based life-like systems.

Acknowledgements This study was supported by the Hiroshima Shudo University grant at 2017 and 2018, Aoba Foundation for the Promotion of Engineering 2012, the JSPS KAKENHI Grant JP15H01069 in 2015–2017, and the JSPS KAKENHI Grant JP15K12144 in 2015–2017, and the Bilateral Joint Research Projects/Seminars between the Japan Society for the Promotion of Science (JSPS) and the Centre National de la Recherche Scientifique (CNRS) in 2015–2017.

References

Akanuma S, Nakajima Y, Yokoboria S, Kimura M, Nemoto N, Mase T, Miyazono K, Tanokura M, Yamagishia A (2013) Experimental evidence for the thermophilicity of ancestral life. Proc Natl Acad Sci USA 110:11067–11072

Blank JG, Miller GH, Ahrens MJ, Winans RE (2001) Experimental shock chemistry of aqueous amino acid solutions and the cometary delivery of prebiotic compounds. Orig Life Evol Biosphe 31:15–51

Boillot F, Chabin A, Buré C, Venet M, Belsky A, Bertrand-Urbaniak M, Delmas A, Brack A, Barbier B (2002) The perseus exobiology mission on MIR: behaviour of amino acids and peptides in Earth orbit. Orig Life Evol Biosph 32:359–385

Buick R, Thornett JR, Mcnaughton NJ, Smith JB, Barley ME, Savage M (1995) Record of emergent continental-crust similar-to-3.5 billion years ago in the Pilbara craton of Australia. Nature 375:574–577

Cech TR (1986) A model for the RNA-catalyzed replication of RNA. Proc Natl Acad Sci USA 83:4360–4363

Cech TR, Zaung AJ, Grabowski PJ (1981) In vitro splicing of the ribosomal RNA precousor of Tetrahymena: involvement of a guanosine nucleotide in the excision of the intervening sequence. Cell 27:487–496

Cleaves HJ, Aubrey AD, Bada JL (2009) An evaluation of the critical parameters for abiotic peptide synthesis in submarine hydrothermal systems. Orig Life Evol Biosph 39:109–126

Corliss JB, Baross JA, Hoffman SE (1981) An hypothesis concerning the relationship between submarine hot springs and the origin of life on Earth. Oceanol Acta 4:59–69

Costanzo G, Saladino R, Crestini C, Ciciriello F, Di Mauro E (2007) Nucleoside phosphorylation by phosphate minerals. J Biol Chem 282:16729–16735

Costanzo G, Pino S, Ciciriello F, Di Mauro E (2009) Generation of long RNA chains in water. J Biol Chem 284:33206–33216

Copley SD, Smith E, Morowitz HJ (2007) The origin of the RNA world: Co-evolution of genes and metabolism. Bioorg Chem 35:430–443

Cowan DA (2004) The upper temperature for life—where do we draw the line? Trends Microbiol 12(2):58–60. https://doi.org/10.1016/j.tim.2003.12.001

Crick F (1970) Central dogma of molecular biology. Nature 227:561–563

Crick FHC (1968) The origin of the genetic code. J Mol Biol 38:367–379

Da Silva L, Maurel M-C, Deamer D (2015) Salt-promoted synthesis of RNA-like molecules in simulated hydrothermal conditions. J Mol Evol 80:86–97

Eigen M (1971) Self-organization of matter and the evolution of biological macromolecules. Naturwissenschaften 65:7–41

Eigen M, Shuster O (1979) The hypercycle. A principle of natural self-organization. Part B: the abstract hypercycle. Naturwissenschaften 58:465–523
Ellington AD, Szostak JW (1990) In vitro selection of RNA molecules that bind specific ligands. Nature 346:818–822
El-Murr N, Maurel M-C, Rihova M, Vergne J, Hervé G, Kato M, Kawamura K (2012) Behavior of a hammerhead ribozyme in aqueous solution at medium to high temperatures. Naturwissenschaften 99:731–738
Ertem G, Ferris JP (1996) Synthesis of RNA oligomers on heterogeneous templates. Nature 379:238–240
Fakhrai H, Inoue T, Orgel LE (1984) Temperature-dependence of the template-directed synthesis of oligoguanylates. Tetrahedron 40(1):39–45
Ferris JP (2002) Montmorillonite catalysis of 30–50 mer oligonucleotides: laboratory demonstration of potential steps in the origin of the RNA world. Orig Life Evol Biosph 32:311–332
Ferris JP, Ertem G (1992) Oligomerization of ribonucleotides on montmorillonite: reaction of the 5′-phosphorimidazolide of adenosine. Science 257:1387–1389
Ferris JP, Sanchez RA, Orgel LE (1968) Studies in prebiotic synthesis: III. Synthesis of pyrimidines from cyanoacetylene and cyanate. J Mol Biol 33:693–704
Ferris JP, Hill AR, Liu JR, Orgel LE (1996) Synthesis of long prebiotic oligomers on mineral surfaces. Nature 381:59–61
Fuller WD, Sanchez RA, Orgel LE (1972a) Studies in prebiotic synthesis VI. Synthesis of purine nucleosides. J Mol Biol 67:25–33
Fuller WD, Sanchez RA, Orgel LE (1972b) Studies in prebiotic synthesis VII. Solid-state synthesis of purine nucleosides. J Mol Evol 1:249–257
Furukawa Y, Sekine T, Oba M, Kakegawa T, Nakazawa H (2009) Biomolecule formation by oceanic impacts on early Earth. Nat Geosci 2:62–66
Galtier N, Tourasse N, Gouy M (1999) A nonhyperthermophilic common ancestor to extant life forms. Science 283:220–221
Gilbert W (1986) Origin of life: the RNA world. Nature 319:618
Gontareva NB, Kuzicheva EA, Shelegedin VN (2009) Synthesis and characterization of peptides after high-energy impact on the icy matrix: Preliminary step for further UV-induced formation. Planet Space Sci 57:441–445
Gough DO (1981) Solar interior structure and luminosity variations. Solar Phys 74:21–34
Guerrier-Takada C, Gardiner K, Marsh T, Pace N, Altman S (1983) The RNA moiety of ribonuclease P is the catalytic sybuit of the enzyme. Cell 35:849–857
Harrison TM (2009) The Hadean Crust: Evidence from >4 Ga Zircons. Ann rev Earth Planet Sci 37:479–505
Harrison JP, Gheeraert N, Tsigelnitskiy D, Cockell CS (2013) The limits for life under multiple extremes. Trends Microbiol 21(4):204–212
Higgs PG, Lehman N (2015) The RNA World: molecular cooperation at the origins of life. Nat Rev Gene 16:7–17. https://doi.org/10.1038/nrg3841
Hill AR Jr, Orgel LE, Wu T (1993) The limits of template-directed synthesis with nucleoside-5′-phosphoro(2-methyl) imidazolides. Orig Life Evol Biosph 23:285–290
Hogeweg P, Takeuchi N (2003) Multilevel selection in models of prebiotic evolution: compartments and spatial self-organization. Orig Life Evol Biosph 33:375–403
Horning DP, Joyce GF (2016) Amplification of RNA by an RNA polymerase ribozyme. Proc Natl Acad Sci USA 113:9786–9791
Ikehara K (2005) Possible steps to the emergence of life: the [GADV]-protein world hypothesis. Chem Rec 5:107–118
Ikehara K (2009) Pseudo-replication of [GADV]-proteins and origin of life. Int J Mol Sci 10:1525–1537. https://doi.org/10.3390/ijms10041525
Imai E, Honda H, Hatori K, Brack A, Matsuno K (1999a) Elongation of oligopeptides in a simulated submarine hydrothermal system. Science 283:831–833

Imai E, Honda H, Hatori K, Matsuno K (1999b) Autocatalytic synthesis of oligoglycine in a simulated submarine hydrothermal system. Orig Life Evol Biosph 29:249–259
Inoue T, Orgel LE (1982) Oligomerization of (guanosine 5′-phosphor)-2-methyl-imidazolide on poly(C), an RNA polymerase model. J Mol Biol 162:201–217
Inoue T, Orgel LE (1983) A nonenzymatic RNA polymerase model. Science 219:859–862
Islam MN, Kaneko T, Kobayashi K (2003) Reaction of amino acids in a Supercritical water-flow reactor simulating submarine hydrothermal systems. Bull Chem Soc Jpn 76:1171–1178
Johnston WK, Unrau PJ, Lawrence MS, Glasner ME, Bartel DP (2001) RNA-catalyzed RNA polymerization: accurate and general RNA-templated primer extension. Science 292:1319–1325
Joyce GF (2002) The antiquity of RNA based evolution. Nature 418:214–221
Joyce GF, Szostak JW (2018) Protocells and RNA self-replication. Cold Spring Harbor Per Biol 10(9):a034801. https://doi.org/10.1101/cshperspect.a034801
Joyce GF, Inoue T, Orgel LE (1984) Non-enzymatic template-directed synthesis on RNA random copolymers, poly(C, U) templates. J Mol Evol 176:278–306
Kaiser RI, Stockton AM, Kim YS, Jensen EC, Mathies RA (2013) On the formation of dipeptides in interstellar model ices. Astrophys J 765:111. https://doi.org/10.1088/0004-637X/765/2/111
Kanavarioti A, Bernasconi CF, Alberas DJ, Baird EE (1993) Kinetic dissection of individual steps on the poly(c)-directed oligoguanylate synthesis from guanosine 5′-monophosphate 2-methylimidazolide. J Am Chem Soc 115(19):8537–8546
Kasting JF (1993) Earth's early atmosphere. Science 259:920–926
Kauffman S (1986) Autocatalytic sets of proteins. J Theor Biol 119:1–24
Kauffman S (2007) Question 1: origin of life and the living state. Orig Life Evol Biosph 37:315–322
Kawamura K (1998) Kinetic analysis of hydrothermal reactions by flow tube reactor—hydrolysis of adenosine 5′-triphosphate at 398–573 K. Nippon Kagaku Kaishi 255–262
Kawamura K (1999) Monitoring of hydrothermal reactions in 3 ms using fused-silica capillary tubing. Chem Lett 28:125–126
Kawamura K (2000) Monitoring hydrothermal reactions on the millisecond time scale using a micro-tube flow reactor and kinetics of ATP hydrolysis for the RNA world hypothesis. Bull Chem Soc Jpn 73:1805–1811
Kawamura K (2001) Hydrolytic stability of ribose phosphodiester bonds within several oligonucleotides at high temperatures using a real-time monitoring method for hydrothermal reactions. Chem Lett 30(11):1120–1121
Kawamura K (2002a) In situ UV-VIS detection of hydrothermal reactions using fused-silica capillary tubing within 0.08–3.2 s at high temperatures. Anal Sci 18:715–716
Kawamura K (2002b) The origin of life from the life of subjectivity. In: Palyi G, Zucchi C, Caglioti L (eds) Fundamentals of life. Elsevier, Paris, pp 563–574
Kawamura K (2003a) Kinetics and activation parameter analyses of hydrolysis and interconversion of 2′,5′- and 3′,5′-linked dinucleoside monophosphate at extremely high temperatures. Biochim Biophys Acta 1620:199–210
Kawamura K (2003b) Kinetic analysis of cleavage of ribose phosphodiester bond within guanine and cytosine rich oligonucleotides and dinucleotides at 65–200 & °C and its implications on the chemical evolution of RNA. Bull Chem Soc Jpn 76:153–162
Kawamura K (2003c) The relative importance of genes, subjectivity, and self-organization for the origin and evolution of life. In: Levit GS, Popov IY, Hossfeld U, Breidbach O (eds) In the shadow of Darwinism: alternative evolutionary theories in the 20th century. Fineday-press, St-Petersburg, pp 218–239
Kawamura K (2004) Behavior of RNA under hydrothermal conditions and the origins of life. Inter J Astrobiol 3:301–309
Kawamura K (2005) A new probe for the indirect measurement of the conformation and interaction of biopolymers at extremely high temperatures using a capillary flow hydrothermal reactor system for UV-visible spectrophotometry. Anal Chim Acta 543(1–2):236–241
Kawamura K (2007) Civilization as a biosystem examined by the comparative analysis of biosystems. BioSystems 90(1):139–150

Kawamura K (2010) Temperature limit for the emergence of life-like system deduced from the prebiotic chemical kinetics under the hydrothermal conditions. In: Fellermann H, Dörr M, Hanczyc MM, Lauren LL, Maurer S, Merkle D, Monnard P-A, Støy K, Rasmussen S (eds) Proceedings of the twelfth international conference on the simulation and synthesis of living systems, pp 37–44

Kawamura K (2011) Development of micro-flow hydrothermal monitoring systems and their applications to the origin of life study on earth. Anal Sci 27(7):675–683

Kawamura K (2012a) Drawbacks of the ancient RNA-based life-like system under primitive earth conditions. Biochimie 94(7):1441–1450

Kawamura K (2012b) Reality of the emergence of life-like systems from simple prebiotic polymers on primitive earth. In: Seckbach J, Gordon R (eds) Genesis—in the beginning: precursors of life, chemical models and early biological evolution. Springer, Dordrecht, pp 123–144

Kawamura K (2016) A hypothesis: life initiated from two genes, as deduced from the RNA world hypothesis and the characteristics of life-like systems. Life 6(3):29

Kawamura K (2017) Hydrothermal microflow technology as a research tool for origin-of-life studies in extreme Earth environments. Life 7(4):37

Kawamura K, Ferris JP (1994) Kinetics and mechanistic analysis of dinucleotide and oligonucleotide formation from the 5′-phosphorimidazolide of adenosine on Na^+-montmorillonite. J Am Chem Soc 116:7564–7572

Kawamura K, Umehara M (2001) Kinetic analysis of the temperature dependence of the template-directed formation of oligoguanylate from the 5′-phosphorimidazolide of guanosine on a poly(C) template with Zn^{2+}. Bull Chem Soc Jpn 74(5):927–935

Kawamura K, Yukioka M (2001) Kinetics of the racemization of amino acids at 225–275 °C using a real-time monitoring method of hydrothermal reactions. Thermochim Acta 375:9–16

Kawamura K, Maeda J (2008) Kinetics and activation parameter analysis for the prebiotic oligocytidylate formation on Na^+-montmorillonite at 0–100 °C. J Phys Chem A 112:8015–8023

Kawamura K, Shimahashi M (2008) One-step formation of oligopeptide-like molecules from Glu and Asp in hydrothermal environments. Naturwissenschaften 95(5):449–454

Kawamura K, Nagayoshi H (2007) Behavior of DNA under hydrothermal conditions with $MgCl_2$ additive using an in situ UV-visible spectrophotometer. Thermochim Acta 466:63–68

Kawamura K, Maurel M-C (2017) Walking over 4 Gya: chemical evolution from photochemistry to mineral and organic chemistries leading to an RNA world. Orig Life Evol Biopsh 47:281–296

Kawamura K, Yosida A, Matumoto O (1997) Kinetic investigations for the hydrolysis of adenosine 5′-triphosphate at elevated temperatures: Prospects for the chemical evolution of RNA. Viva Origino 25(3):177–197

Kawamura K, Kameyama N, Matumoto O (1999) Kinetics of hydrolysis of ribonucleotide polymers in aqueous solution at elevated temperatures: implications of chemical evolution of RNA and primitive ribonuclease. Viva Origino 27(2):107–118

Kawamura K, Kuranoue K, Nagahama M (2004) Prebiotic inhibitory activity of protein-like molecules to the template-directed formation of oligoguanylate from guanosine 5′-monophosphate 2-methylimidazolide on a polycytidylic acid template. Bull Chem Soc Jpn 77(7):1367–1375

Kawamura K, Nagahama M, Kuranoue K (2005a) Chemical evolution of RNA under hydrothermal conditions and the role of thermal copolymers of amino acids for the prebiotic degradation and formation of RNA. Adv Space Res 35(9):1626–1633

Kawamura K, Nishi T, Sakiyama T (2005b) Consecutive elongation of alanine oligopeptides at the second time range under hydrothermal condition using a micro flow reactor system. J Am Chem Soc 127(2):522–523

Kawamura K, Nagayoshi H, Yao T (2009) Stability of ribonuclease a under hydrothermal conditions in relation to the origin-of-life hypothesis: verification with the hydrothermal micro-flow reactor system. Res Chem Intermed 35:879–891

Kawamura K, Nagayoshi H, Yao T (2010) In situ analysis of proteins at high temperatures mediated by capillary-flow hydrothermal UV-Vis spectrophotometer with a water-soluble chromogenic reagent. Anal Chim Acta 667:88–95

Kawamura K, Takeya H, Kushibe T, Koizumi Y (2011) Mineral-enhanced hydrothermal oligopeptide formation at the second time scale. Astrobiology 11(5):461–469
Kawamura K, Yasuda T, Hatanaka T, Hamahiga K, Matsuda N, Ueshima M Nakai K (2016) Oxidation of aliphatic alcohols and benzyl alcohol by H_2O_2 under the hydrothermal conditions in the presence of solid-state catalysts using batch and flow reactors. Chem Eng J 285:49–56
Kawamura K, Yasuda T, Hatanaka T, Hamahiga K, Matsuda N, Ueshima M Nakai K (2017) In situ UV-VIS spectrophotometry within the second time scale as a research tool for solid-state catalyst and liquid-phase reactions at high temperatures: Its application to the formation of HMF from glucose and cellulose. Chem Eng J 307:1066–1075
Kawamura K, Konagaya N, Maruoka Y (2018) Enhancement and inhibitory activities of minerals for alanine oligopeptide elongation under hydrothermal conditions. Astrobiology 18(11):1403–1413
Kim YE, Higgs PG (2016) Co-operation between polymerases and nucleotide synthetases in the RNA World. Plos Comp Biol 12(11):e1005161. https://doi.org/10.1371/journal.pcbi.1005161
Landweber LF (1999) Testing ancient RNA–protein interactions. Proc Natl Acad Sci USA 96:11067–11068
Larralde R, Robertson MP, Miller SL (1995) Rates of decomposition of ribose and other sugars: implications for chemical evolution. Proc Natl Acad Sci USA 92:8158–8160
Lee DH, Severin K, Yokobayashi Y, Ghadiri MR (1997) Emergence of symbiosis in peptide self-replication through a hypercyclic network. Nature 390:591–594
Lohrmann R (1977) Formation of nucleoside 5′-phosphorimidates under potentially prebiological conditions. J Mol Evol 10:137–154
Lohrmann R, Orgel LE (1973) Prebiotic activation processes. Nature 244:418–420
Martins A, Price MC, Goldman N, Sephton MA, Burchell MJ (2013) Shock synthesis of amino acids from impacting cometary and icy planet surface analogues. Nat Geosci 6:1045–1049
Maruyama S, Ikoma M, Genda H, Hirose K, Yokoyama T, Santosh M (2013) The naked planet earth: most essential pre-requisite for the origin and evolution of life. Geosci Front 4:141–165
Maury CPJ (2009) Self-propagating β-sheet polypeptide structures as prebiotic informational molecular entities: the amyloid world. Orig Life Evol Biosph 39:141–150
Miller SL (1953) A production of amino acids under possible primitive Earth conditions. Science 117:528–529
Mojzsis SJ, Arrhenius G, McKeegan KD, Harrison TM, Nutman AP, Friend CRL (1996) Evidence for life on Earth before 3,800 million years ago. Nature 384:55–59
Mojzsis SJ, Harrison TM, Pidgeon RT (2001) Oxygen-isotope evidence from ancient zircons for liquid water at the Earth's surface 4,300 Myr ago. Nature 409:178–181
Nagafuchi K, Nagira A, Akiyama H, Sasaki M, Kawamura K (2013) Oligopeptide production from alanine monomer by pulsed corona discharge plasma in ambient and supercritical argon. Chem Eng Sci 1(3):41–45. https://doi.org/10.12691/ces-1-3-2
Nemoto N, Husimi Y (1995) A model of the virus-type strategy in the early stage of encoded molecular evolution. J Theor Biol 176:67–77
Nemoto N, Miyamoto-Sato E, Husimi Y, Yanagawa H (1997) In vitro virus: bonding of mRNA bearing puromycine at the 30-terminal end to the C-terminal end of its encoded protein on the ribosome in vitro. FEBS Lett 414:405–408
Neveu M, Kim H-J, Benner SA (2013) The "Strong" RNA world hypothesis: fifty years old. Astrobiology 13(4):391–403
Nghe P, Hordijk W, Kauffman SA, Walker SI, Schmidt FJ, Kemble H, Yeates JAM, Lehman N (2015) Prebiotic network evolution: six key parameters. Mol BioSys 11:3206–3217. https://doi.org/10.1039/c5mb00593k
Nutman AP, Friend CRL, Bennett VC (2001) Review of the oldest (4400–3600 Ma) geological and mineralogical record: glimpses of the beginning. Episode 24:93–101
Nutman AP, Bennet VC, Friend CRL, Van Kranendonk M, Chivas AR (2016) Rapid emergence of life shown by discovery of 3,700-million-year-old microbial structures. Nature 537:535–538
Oparin AI (1924) Proiskhozhdenie zhizni. Moscow Izd. Moskovskii Rabochii, Russia

Orgel LE (2004) Prebiotic chemistry and the origin of the RNA world. Crit Rev Biochem Mol Biol 39:99–123
Orgel LE, Crick FH (1993) Anticipating an RNA world. Some past speculations on the origin of life: where are they today? FEBS J 7:238–239
Orgel LE, Lohrmann R (1974) Prebiotic chemistry and nucleic acid replication. Acc Chem Res 7:368–377
Oró J (1961) Mechanism of synthesis of adenine from hydrogen cyanide under possible primitive earth conditions. Nature 191:1193–1194
Otake T, Taniguchi T, Furukawa Y, Kawamura F, Nakazawa H, Kakegawa T (2011) Stability of amino acids and their oligomerization under high-pressure conditions: implications for prebiotic chemistry. Astrobiolgy 11:799–813
Pace NR (1991) Origin of life—facing up to the physical setting. Cell 65:531–533
Pasteur L (1861) Mémeoire sur les corpuscules organisés qui existent dans l'atmospère. Examen de la adoctrine des générations spontenées (tran: in Japanese (1970), Yamaguchi S (translated)). Iwanami Bunko, Tokyo
Pikuta EV, Hoover RB, Tang J (2007) Microbial extremophiles at the limits of life. Cri Rev Microbiol 33(3):183–209. https://doi.org/10.1080/10408410701451948
Ponnamperuma C, Mack R (1965) Nucleotide synthesis under possible primitive earth conditions. Science 148:1221–1223
Radzicka A, Wolfenden R (1995) A proficient enzyme. Science 267:90–93
Reimann E, Zubay G (1999) Nucleoside phosphorylation: a feasible step in the prebiotic pathway to RNA. Orig Life Evol Biosph 29:229–247
Ricardo A, Carrigan MA, Olcott AN, Benner SA (2004) Borate mineral stabilize ribose. Science 303:196
Saladino R, Botta G, Pino S, Costanzoc G, Di Mauro E (2012) Genetics first or metabolism first? The formamide clue. Chem Soc Rev 41:5526–5565
Sanchez RA, Ferris JP, Orgel LE (1966) Cyanoacetylene on prebiotic synthesis. Science 154:784–785
Santosha M, Arai T, Maruyama S (2017) Hadean earth and primordial continents: The cradle of prebiotic life. Geosci Front 8:309–327
Sawai H (1976) Catalysis of internucleotide bond formation by divalent metal ions. J Am Chem Soc 98:7037–7039
Sawai H, Kuroda K, Hojo H (1989) Uranyl ion as a highly effective catalyst for internucleotide bond formation. Bull Chem Soc Jpn 62:2018–2023
Sawai H, Shibata T, Ohno M (1981) Preparation of oligoadenylares with 20-50 linkage using Pb^{2+} ion catalyst. Tetrahedron 37:481–485
Schildkraut C, Lifson S (1965) Dependence of the melting temperature of DNA on salt concentration. Biopolymers 3(2):195–208
Schwartman DW, Lineweaver CH (2004) The hyperthermophilic origin of life revisited. Biochem Soc Trans 32:168–171
Shapiro R (1988) Prebiotic ribose synthesis: a critical analysis. Orig Life Evol Biosph 18:71–85
Sleep NH (2018) Geological and geochemical constraints on the origin and evolution of life. Astrobiol 18:1199–1219
Stüeken EE, Anderson RE, Bowman JS, Brazelton WJ, Colangelo-Lillis J, Goldman AD, Som SM, Baross JA (2013) Did life originate from a global chemical reactor? Geobiol 11:101–126
Szostak N, Wasik S, Blazewicz J (2016) Hypercycle. Plos Comp Biol 12(4):e1004853. https://doi.org/10.1371/journal.pcbi.1004853
Tuerk C, Gold L (1990) Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. Science 249:505–510
Vaidya N, Manapat ML, Chen IA, Xulvi-Brunet R, Hayden EJ, Lehman N (2012) Spontaneous network formation among cooperative RNA replicators. Nature 491:72–78
van Zuilen MA, Lepland A, Arrhenius G (2002) Reassessing the evidence for the earliest traces of life. Nature 418:627–630

Vasas V, Fernando C, Santos M, Kauffman S, Szathmáry E (2012) Evolution before genes. Biol Dire 7:1. https://doi.org/10.1186/1745-6150-7-1
Waehneldt TV, Fox SW (1967) Phosphorylation of nucleosides with poly-phosphoric acid. Biochim Biophys Acta 134:1–8
Watson JD, Crick FHC (1953) Molecular structure of nucleic acids. Nature 171:737–738
Wetmur JG, Davidson N (1968) Kinetics of renaturation of DNA. J Mol Biol 31:349–370
White RH (1984) Hydrolytic stability of biomolecules at high temperatures and its implication for life at 250 °C. Nature 310:430–432
Wu M, Higgs PG (2009) Origin of self-replicating biopolymers: autocatalytic feedback can jump-start the RNA world. J Mol Biol 69:541–554
Yamagata Y, Watanabe H, Namba T (1992) Volcanic production of polyphosphates and its relevance to prebiotic evolution. Nature 352:516–519
Yao S, Ghosh I, Zutshi R, Chmielewski J (1998) Selective amplification by auto- and cross-catalysis in a replicating peptide system. Nature 396:447–450
Zaher HS, Unrau PJ (2007) Selection of an improved RNA polymerase ribozyme with superior extension and fidelity. RNA 13:1017–1026

Chapter 2
Formation of Nucleosides and Nucleotides in Chemical Evolution

Hideo Hashizume, Benny K. G. Theng, Sjerry van der Gaast and Kazuko Fujii

Abstract Nucleosides and nucleotides are important biomolecules. Following Gilbert's (Nature 319:618, 1986) proposal of an "RNA world," various processes for the formation of nucleosides (from nucleobases and ribose) and the polymerization of nucleotides have been suggested. Problems associated with the formation of RNA have also been pointed out. The constituents of RNA are nucleobases, ribose, and phosphate. Ribose has five conformational isomers or conformers, each of which can react with a nucleobase. In life, however, only the β-furanose form of ribose is used. Curiously, when a nucleobase reacts with ribose in an aqueous solution, only a small amount of nucleoside with a β-ribofuranose component is detectable in the total products. Thus, the RNA world hypothesis has reached a deadlock. Here, we summarize the important points in the synthesis of nucleobases and ribose. We also describe the selective formation of nucleosides and touch on the one-pot synthesis of nucleotides.

2.1 Introduction

Deoxyribonucleic acid (DNA) and ribonucleic acid (RNA) are the most important biopolymers. For this reason, the formation of DNA and RNA under prebiotic conditions is central to questions regarding the origin of life on Earth. In living organisms, DNA serves as a storehouse of genetic information, while RNA transcribes this information in catalyzing protein synthesis. On the other hand, proteins have various functions such as enzyme and antibody. Some proteins are involved in the restoration of DNA and cell division, indicating the close relationship between DNA and protein. The question arises which polymer came first in the primordial and prebiotic

H. Hashizume (✉) · K. Fujii
National Institute for Materials Science, Namiki, Tsukuba, Japan
e-mail: HASHIZUME.Hideo@nims.go.jp

B. K. G. Theng
Manaaki Whenua–Landcare Research, Palmerston North, New Zealand

S. van der Gaast
Royal Netherlands Institute for Sea Research, Den Burg, The Netherlands

P. Pontarotti (ed.), *Evolution, Origin of Life, Concepts and Methods*,
https://doi.org/10.1007/978-3-030-30363-1_2

Earth. Earlier on, Rich (1962) postulated that polynucleotide chains could catalyze the polymerization of complementary nucleotide residues to yield a double-stranded DNA-like molecule. The subsequent discovery by Cech (1986) of an RNA showing catalytic activity provides support for the RNA world hypothesis (Gilbert 1986) that RNA formation preceded that of DNA and proteins.

RNA is composed of nucleobases (adenine, guanine, cytosine, uracil), ribose, and phosphate. When a nucleobase reacts with ribose, a nucleoside is formed. Condensation of a nucleoside with phosphate yields a nucleotide which, in turn, can polymerize to form RNA. By comparison with peptide formation from amino acids, however, the synthesis of RNA from its components (nucleobase, ribose, phosphate) is problematic (e.g., Robertson and Joyce 2012; Fiore and Strazewski 2016). Here, we describe the synthesis of nucleobases and ribose under possible prebiotic conditions on the early Earth and in space, leading to the formation of nucleosides and nucleotides.

2.2 Formation of Nucleobases

The nucleobases, adenine, guanine, cytosine, uracil and thymine (Fig. 2.1), can be synthesized from simple organic molecules under relatively mild experimental conditions. Uracil occurs in RNA, and thymine is part of the DNA structure. Adenine and guanine are purines, while cytosine, uracil, and thymine are pyrimidines.

Purine bases can form from hydrogen cyanide and pyrimidine bases from cyanoacetylene (Fig. 2.2). For example, adenine is a pentamer of hydrogen cyanide **1** (Sanchez et al. 1966a) as indicated in Fig. 2.2a. Guanine can also form by a similar process. Starting with 4-aminoimidazole-5-carboxamid **2**, guanine is formed by the addition of cyanogen (dicyan) and H_2O (Sanchez et al. 1966b) as shown in Fig. 2.2b. Robertson and Miller (1995) synthesized cytosine by heating cyanoacetaldehyde **3** and urea **4** at 100 °C (Fig. 2.2c). Cytosine can also be obtained from

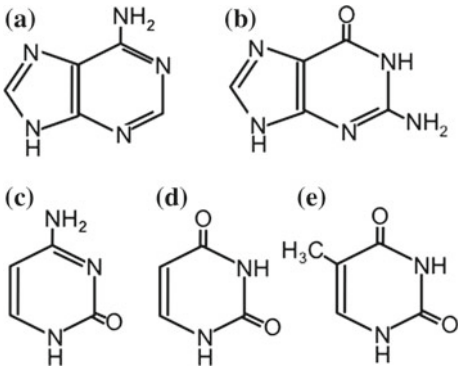


Fig. 2.1 Structural scheme of nucleobases: **a** adenine; **b** guanine; **c** cytosine; **d** uracil; **e** thymine

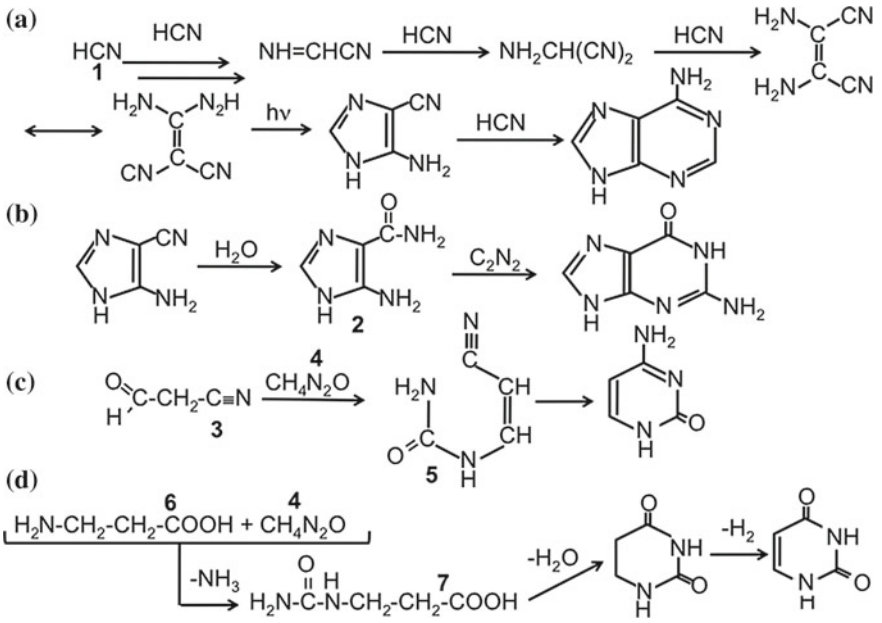


Fig. 2.2 Possible prebiotic pathways for the synthesis of adenine (**a**), guanine (**b**), cytosine (**c**), and uracil (**d**). After Sanchez et al. (1966a, b), Robertson and Miller (1995) and Chittenden and Schwartz (1976)

trans-cyanovinylurea **5**. Figure 2.2d shows that uracil can be synthesized from β-alanine **6** and urea **4** via β-ureidopropionic acid **7**, through dehydration and photodehydrogenation (Chittenden and Schwartz 1976).

The environment of primitive Earth was not always mild. During the Late Heavy Bombardment period, from 4 to 3.85 Ga, a huge number of meteorites impacted on Earth, leading to the possible formation of biomolecular pre-cursors. In simulating such conditions, Ferus et al. (2015) used a high-power laser to promote the dielectric collapse of the plasma produced by a meteorite impact. When liquid formamide was irradiated by this means, the molecule decomposed to yield highly reactive CN· and NH· radicals which reacted with formamide to produce adenine, guanine, cytosine, and uracil in the presence of clay. Cytosine was not detected in the presence or absence of a meteorite (chondrite), while irradiation of 2,3-diaminomaleonitrile also produced the four nucleobases.

Astrobiological research has indicated that the planets, and such satellites as Titan and Europa, are covered with ice that might contain simple organic molecules. Biomolecules could form when the ice is irradiated with UV light or high-energy cosmic rays. Similarly, meteorites and comets might hold some biomolecules. The formation of nucleobases under cold conditions has been investigated by Menor-Salván and Marin-Yaseli (2013). By irradiating urea in an acetylene atmosphere with UV light at −21 to 5 °C, they were able to detect guanine, uric acid, uracil, and

cytosine in the reaction products. Earlier, Miyakawa et al. (2000) reported the formation of guanine by irradiating a N_2–CO–H_2O gas mixture with a high-temperature plasma.

2.3 Formation and Conformation of Ribose

It is well known that in the presence of metal catalysts, sugars can arise from formaldehyde by means of the "formose" reaction. Thus, condensation of formaldehyde gives rise to glycolaldehyde which then reacts with formaldehyde (by the aldol reaction) to form glyceraldehyde. Isomerization of glyceraldehyde yields dihydroxyacetone which on reaction with glycolaldehyde gives rise to ribulose and then to ribose by isomerization.

Gabel and Ponnamperuma (1967) have reported the formation of monosaccharides from formaldehyde in the presence of alumina and clay minerals (kaolinite and illite) under hydrothermal conditions. Likewise, Saladino et al. (2010) were able to synthesize amino sugar derivatives, adenine, and cytosine from formamide, formaldehyde, and water in the presence of $Mg(OH)_2$, $Pb(NO_3)$ and montmorillonite. The formation of amino sugar derivatives was highly dependent on the volume of water used.

Earlier, Ricardo et al. (2004) reported the formation of pentoses, such as ribose, arabinose, xylose, and lyxose from glycolaldehyde in the presence of borate (Fig. 2.3). In the presence of $Ca(OH)_2$, glycolaldehyde **8** forms glyceraldehyde **9** which can react with borate to yield a diglyceraldehyde–borate complex, while enolization of glyceraldehyde is suppressed. Combination of this complex with the enediolate of glyceraldehyde (ethenediol) **10** gives rise to a pentose **11** and a furanose-type ribose–borate complex. Similarly, Furukawa and Kakegawa (2017) have experimentally shown that borate has a strong affinity for ribose and can stabilize ribose by complexation during the formose reaction.

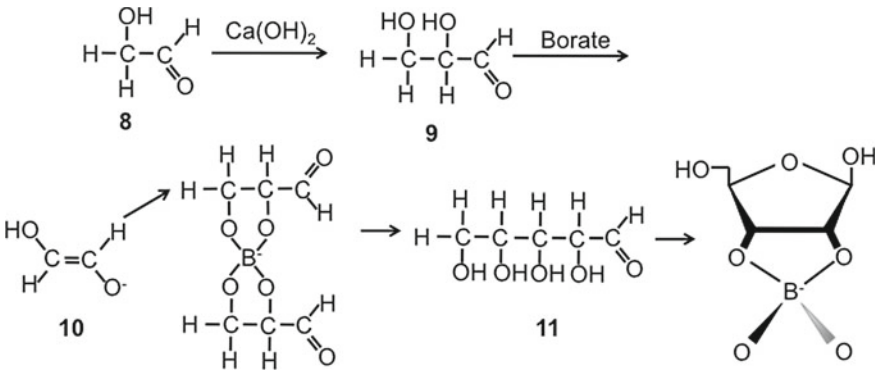


Fig. 2.3 Formation of ribose in the presence of borate. After Ricardo et al. (2004)

The presence of glycolaldehyde in the galaxy opens the possibility that this compound was brought to the early Earth through a meteorite. Civiš et al. (2016) have investigated the formation of sugars at high temperatures simulating a meteorite impact on Earth. These workers irradiated frozen paraformaldehyde and anatase with a high-density energy plasma at a temperature of ca. 4500 K. Several sugars were detected including arabinose, ribose, and xylose.

Research into the molecular structure of ribose has focused on the D-enantiomer since L-ribose is not found in nature. The five conformational isomers of D-ribose are α- and β-pyranose, α- and β-furanose, and the open-chain structure (Fig. 2.4). The relative abundance in water of the five conformers varies with temperature and pH (Cortes et al. 1991). At 36 °C and pH 7.0, the ratio of α-pyranose: β-pyranose: α-furanose: β-furanose: open-chain structure is approximately 20:58:7.0:13.5:1.5. The concentration of open-chain ribose is generally so low as to be negligible. In reacting adenine with ribose in water, we might therefore expect to obtain ~10% of β-furanose adenosine. However, much less of this compound is commonly detected in the reaction products (Fuller et al. 1972).

The stabilizing effect of cations, silicate, and borate with respect to the conformational isomers of ribose and prebiotic organic compounds, in general, has received a great deal of attention (Ricardo et al. 2004; Maurel and Leclerc 2016; Furukawa and Kakegawa 2017).

Using solid-state ^{13}C NMR spectroscopy, Akouche et al. (2016) have found that different cations in solution can influence the ratio of α- and β-pyranose, and α- and β-furanose in the presence of amorphous silica (Table 2.1). Interestingly, β-furanose was not detected in the absence of any metal cations and amorphous silica.

Several investigators (Benner et al. 2012; Furukawa et al. 2013, 2015) have reported that borate plays an important role in the formation of the β-furanose form of ribose. Earlier, Amaral et al. (2008) used ^{1}H, ^{11}B, and ^{13}C NMR spectroscopy

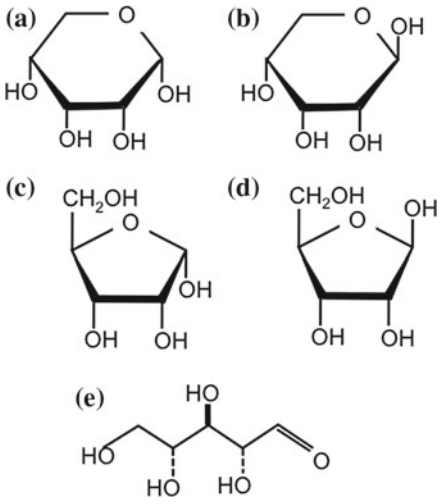


Fig. 2.4 Different structural forms (conformers) of D-ribose: **a** α-pyranose; **b** β-pyranose; **c** α-furanose; **d** β-furanose; and **e** open-chain form

Table 2.1 Ratio of ribose isomers (%) in the presence of cations and amorphous silica (Akouche et al. 2016)

Addition	α-pyr	β-pyr	Total pyr	α-fur	β-fur	Total fur
Ribose in solution	20	62	82	6	12	18
Solid ribose	n.d.	n.d	100	–	–	0
SiO_2	29	47	76	15	9	24
$Mg + SiO_2$	18	48	66	18	16	34
$Ca + SiO_2$	23	26	49	29	22	51
$Zn + SiO_2$	27	9	36	14	50	64

pyr pyranose, *fur* furanose, *n.d.* not detected

to analyze the water stability of β-ribofuranose in the presence of borate, carbonate, bicarbonate, and sulfate of alkaline or alkaline earth cations. They were able to detect β-ribofuranose in the presence of borate. On the other hand, carbonate or bicarbonate did not affect either the formation or the stability of β-ribofuranose. However, when borate was added to the solution of carbonate or bicarbonate, β-ribofuranose reappeared. They also observed that the ratio of α- to β-furanose varied with the nature of the cation (Na^+, K^+, Mg^{2+}, Ca^2, and Sr^{2+}) in the borate salt.

Similarly, Lambert et al. (2004) reported that silicate could form a stable complex with ribose. At pH ~7, the ratio of α-pyranose: β-pyranose: α-furanose: β-furanose was 21.5:58.5:6.5:13.5. At a pH of about 12, however, the presence of α- and β-furanose was not detected by ^{13}C NMR spectroscopy, and only the pyranose forms of ribose existed in solution. In the presence of silicate, on the other hand, α- and β-furanose could form even at pH 12. By forming a complex with silicate, ribofuranose was stabilized against conversion to the corresponding pyranose form.

Šišak et al. (2010) have determined the crystal structure of D-ribose using X-ray diffraction and solid-state ^{13}C NMR spectroscopy. Near the melting point of ribose (90 °C), the ratio of α-pyranose:β-pyranose:α-furanose:β-furanose was 46:25:12:17. On cooling, the structure was consistent with the formation of a glassy state, but at 4 °C, the crystal structure reappeared.

Hu et al. (2013) used FT-IR, F-IR, THz, and Raman spectroscopy to analyze the structure of the complex between ribose and Cs^+ that formed by drying a solution of ribose with CsCl at 80 °C. In this instance, D-ribose has the pyranose type structure in line with the finding by Šišak et al. (2010).

The conformation of D-ribose in water solution has been assessed by Quesada-Moreno et al. (2013) using computer simulation. The ratios of α-pyranose:β-pyranose:α-furanose:β-furanose: open-chain structure, derived from three theoretical models, were 71:28.8:0.1:0.1:0; 71.6:28.3:0.1:0:0 and 69.2:30.6:0.1:0.1:0, respectively. The results suggest that the α- and β-pyranose forms are dominant, while the furanose conformer would not survive in an aqueous environment. The results of experimental vibrational studies using infrared, Raman, and vibrational circular dichroism spectroscopies were consistent with those obtained by computer simulation. In the case of solid ribose, computer simulation failed to give any results.

The rate of decomposition of ribose at different temperatures and pH was investigated by Larralde et al. (1995). At 100, 25, and 0 °C and pH 7, the half-life of ribose was 73 min, 300 days, and 44 years, respectively. The sensitivity of ribose to temperature at neutral pH raises questions about its availability as a prebiotic reagent.

We may conclude that in both aqueous solution and the solid phase, the pyranose form of ribose is dominant over the furanose conformer, and the β-furanose structure is preferred to its alpha counterpart (Cortes et al. 1991). The β-furanose conformer is also enriched when a ribose solution is dried in the presence of borate and some metal cations.

2.4 Formation of Nucleosides and Nucleotides

Nucleosides and nucleotides may be synthesized in one of two ways. In the stepwise or building block approach, a nucleobase combines with ribose to form a nucleoside which then reacts with phosphate to yield a nucleotide. Alternatively, the various components are assembled in one pot and allowed to react.

Fuller et al. (1972) made an early attempt at synthesizing nucleosides by reacting a purine (adenine, guanine, hypoxanthine) with D-ribose at different temperatures (30–100 °C) and pH (2–11) in water. They were able to obtain β-inosine, β-adenosine, and β-guanosine, but yields were small (<10%). More recently, Nam et al. (2018) reported the synthesis of purine and pyrimidine nucleosides by electrospraying aqueous microdroplets of D-ribose, phosphoric acid, and a nucleobase together with Mg^{2+} (as a catalyst) into a mass spectrometer. Mass spectrometry indicated the formation of uridine (2.5%), adenosine (2.5%), cytidine (0.7%), and inosine (1.7%) during a flight time of ca. 50 μs. The nucleoside products, however, were not recovered.

Akouche et al. (2017) have reported the formation of adenosine monophosphate (AMP) by adding amorphous silica to a solution of monophosphate, D-ribose, and adenine and drying at 70 °C. Phosphoribosyl pyrophosphate, formed from phosphate and D-ribose, reacted with adenine to produce AMP.

Kim and Kim (2019) were able to synthesize nucleotides by reacting ribose-1,2-cyclic phosphate with nucleobases and an aqueous solution of $CaCl_2$ in an Eppendorf tube at 75–125 °C for 3–96 h until all the water evaporated. The yield of cytidine-2′-phosphate from the reaction of ribose-1,2-cyclic phosphate with cytosine was about 7% but much less when uracil and guanine served as the nucleobase.

Earlier, Powner et al. (2009) described the formation of pyrimidine ribonucleotides from cyanamide **12**, cyanoacetylene **13**, glycoladehyde **8**, glyceraldehyde **9,** and phosphate. The reaction of amino oxazole **14** with glyceraldehyde **9** gave rise to pentose amino oxazoline **15** and arabinose derivatives. They also detected the formation of β-ribocytidine-2′3′-cyclic phosphate **16** by phosphorylation of anhydroarabinonucleoside (Fig. 2.5).

Xu et al. (2017) mentioned the importance of thiolysis in the formation of pyrimidine β-ribonucleoside. Thus, anhydroarabinonucleoside, formed by reacting amino oxazoline **14** and glyceraldehyde **9** with cyanoacetylene **13**, could be thiolysed to

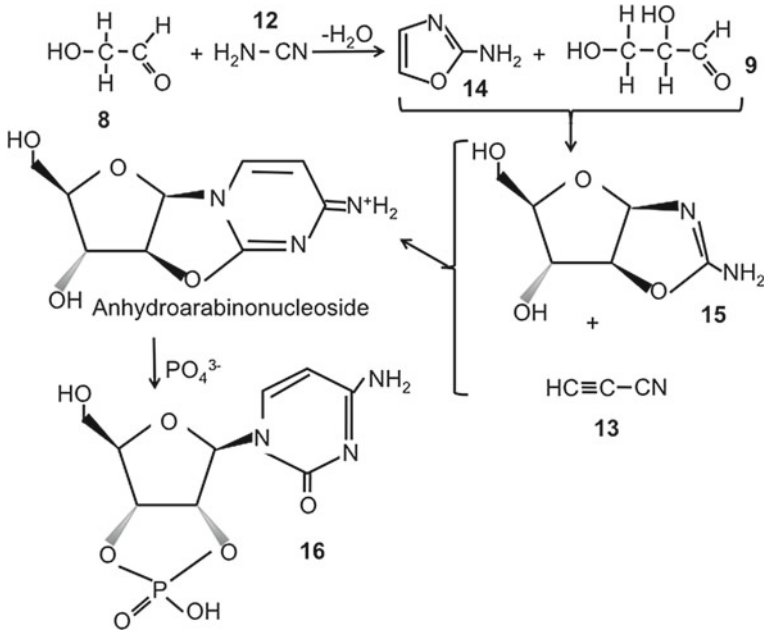


Fig. 2.5 Possible prebiotic formation of β-ribocytidine-2′3′-cyclic phosphate. After Powner et al. (2009)

yield α-2-thioribocytidine **17**. Irradiation of this compound with UV light produced β-2-thioribocytidine **18** which, on hydrolysis, yielded β-ribocytidine. When β-2-thioribocytidine was reacted with phosphate, they obtained β-ribocytidine-2′3′-cyclic phosphate **19** (Fig. 2.6).

An example of the one-pot synthesis of purine nucleosides is the formation of aminopyrimidine **21** from the reaction of guanidine with aminomalononitrile and its conversion to formamidopyrimidine **20** (Fig. 2.7). When formamidopyrimidine combines with open-chain ribose **22**, α- and β-riboadenosine are formed. The yield of β-furanose adenosine is 20% and that of α-pyranose adenosine is about 18% (Becker et al. 2016). Subsequently, Becker et al. (2018) proposed a model of nucleoside formation under conditions of repeated wetting and drying. When a pond or lake containing simple organic molecules is dried, such molecules as hydroxyimino malononitrile and amidine derivatives are concentrated as the water evaporates. Pyrimidine derivatives, such as formamidopyrimidine, are formed at around 100 °C. This process opens the way for the formation of pyrimidine and purine nucleosides as shown in Fig. 2.7.

Irradiation of liquid formamide with high-energy proton beams, in the presence of powdered meteorites, produced adenosine, cytidine, thymidine, and uridine besides amino acids, carboxylic acids, nucleobases, and sugars (Saladino et al. 2015). In the presence of montmorillonite, sulfide minerals, and magnetite, the yield of prebiotic

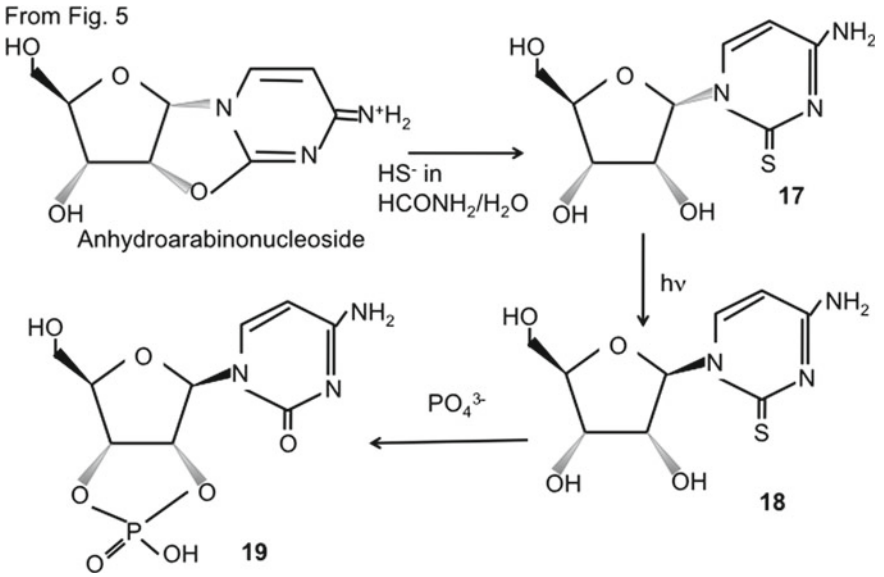


Fig. 2.6 Formation of β-ribocytidine-2′3′-cyclic phosphate by thiolysis of anhydroarabinonucleoside. After Xu et al. (2017)

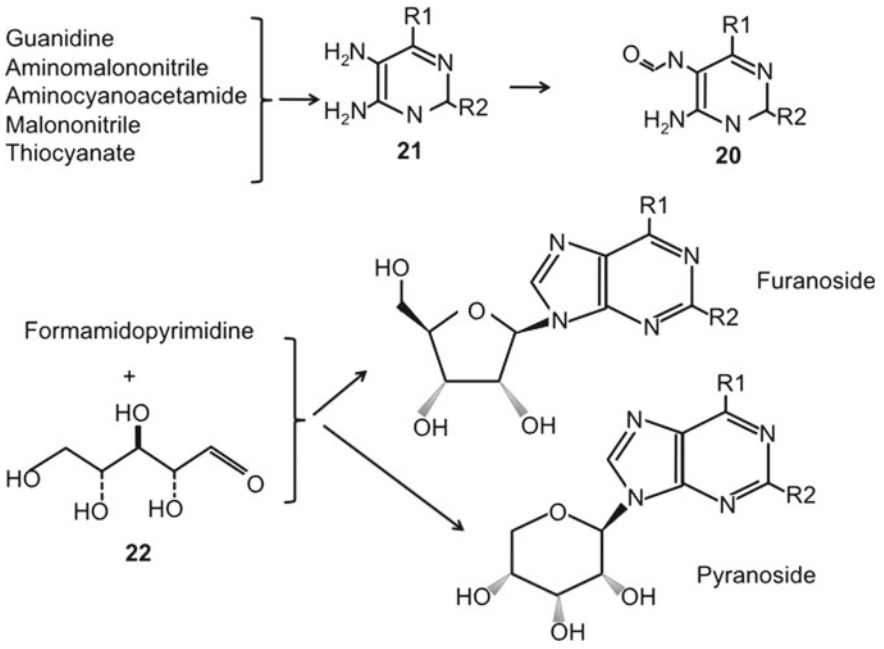


Fig. 2.7 One-pot synthesis of nucleosides (Becker et al. 2016)

compounds was lower than that obtained with meteorites as catalysts. Likewise, Saladino et al. (2017) have reported that a chondrite meteorite can promote the one-pot synthesis of ribo- and 2′-deoxy-ribonucleosides from sugars and purine nucleobases through proton irradiation, mimicking conditions on early Earth fluxed by slow solar wind protons. Besides being based on a known radical mechanism, the N-glycosidic bond formation is stereo- and regio-selective.

Wang et al. (2017) used density functional theory to probe the formation of nucleoside by reacting 2,4,6-triaminopyrimidine (TAP) with all five conformers of D-ribose. Interestingly, the open-chain ribose was the most feasible to react with TAP during which process the ribose structure changed to the furanose form. Water molecules played an important "H-bridging" role during the formation of α-ribofuranoside.

2.5 Concluding Remarks

We have provided a brief review of the formation of nucleosides and nucleotides, including that of nucleobases together with the factors controlling the stability of the conformational isomers (conformers) of ribose. The formation of RNA from its constituent parts (building blocks), however, is problematic (Orgel 2004). In this respect, a major difficulty relates to the occurrence of different structural isomers of ribose and the inherent instability of ribose. Ribose can be stabilized and protected from decomposition, by forming complexes with various anions, cations, and minerals. The use of borate has received particular attention (Benner et al. 2012). If a solution of ribose can be enriched with respect to the β-furanose form, the synthesis of β-furanose nucleosides from nucleobases and ribose would be facilitated. This idea has prompted investigations into the one-pot synthesis of nucleosides. The overall reaction, however, is very slow, involving many steps (Xu et al. 2017).

The formation of nucleosides and nucleotides in the presence of metals and minerals has also been the subject of much research. Besides improving product yields, solid catalysts may be capable of controlling the conformation of ribose. The literature on the role of clay minerals in mediating and promoting prebiotic organic reactions, including nucleotides, has been summarized by Maurel and Leclerc (2016) and Theng (2018).

References

Akouche M, Jaber M, Maurel M-C, Lambert J-F, Georgelin T (2017) Phosphoribosyl pyrophosphate: a molecular vestige of the origin of life on minerals. Angew Chem Int Ed 56:7920–7923. https://doi.org/10.1002/anie201702633

Akouche M, Jaber M, Zins E-M, Maurel M-C, Lambert J-F, Georgelin T (2016) Thermal behavior of D-ribose adsorbed on silica: effect of inorganic salt coadsorption and significance for prebiotic chemistry. Chem Eur J 22:15834–15846

Amaral AF, Marques MM, da Silva JAL, da Silva JJRF (2008) Interactions of D-ribose with polyatomic anions, and alkaline and alkaline-earth cations: possible clues to environmental synthesis conditions in the pre-RNA world. New J Chem 32:2043–2049

Becker S, Thoma I, Deutsch A, Gehrke T, Mayer P, Zipse H, Carell T (2016) A high-yielding strictly regioselective prebiotic purine nucleoside formation pathway. Science 352:833–836

Becker S, Schneider C, Okamura H, Crisp A, Amatov T, Dejmek A, Carell T (2018) Wet-dry cycles enable the parallel origin of canonical and non-canonical nucleosides by continuous synthesis. Nat. Com 9:163. https://doi.org/10.1038/s41467-017-02639-1

Benner SA, Kim H-J, Carrigan MA (2012) Asphalt, water and the prebiotic synthesis of ribose, ribonucleoside and RNA. Acc Chem Res 45:2025–2034

Cech TR (1986) A model for the RNA-catalyzed replication of RNA. Proc Natl Acad Sci USA 83:4360–4363

Chittenden GJF, Schwartz A (1976) Possible pathway for prebiotic uracil synthesis by photodehydrogenation. Nature 263:350–351

Civiš S, Szabla R, Szyja BM, Smykowski D, Ivanek O, Knížek A, Kubelik P, Šponer J, Ferus M, Šponer JE (2016) TiO_2-calayzed synthesis of sugars from formaldehyde in extraterrestrial impacts on the early Earth. Sci Rep 6:23199. https://doi.org/10.1038/srep23199

Cortes SJ, Mega TL, Van Etten RL (1991) The ^{18}O isotope shift in ^{13}C nuclear magnetic resonance spectroscopy. 14. kinetics of oxygen exchange at the anomeric carbon of D-ribose and D-2-deoxyribose. J Org Chem 56:943–947

Ferus M, Nesvorný D, Šponer J, Kubelík P, Michalcíková R, Shestivská V, Šponer JE, Civiš S (2015) High-energy chemistry of formamide: a unified mechanism of nulcoebase formation. PNAS 112:657–662. https://doi.org/10.1073/pnas.1412072111

Fiore M, Strazewski P (2016) Bringing prebiotic nucleosides and nucleotides down to Earth. Angew Chem Int Ed 55:13930–13933. https://doi.org/10.1002/anie.201606232

Fuller WD, Sanchez RA, Orgel LE (1972) Studies in prebiotic synthesis VI. Synthesis of purine nucleosides. J Mol Biol 67:25–33

Furukawa Y, Kakegawa T (2017) Borate and the origin of RNA: a model for the precursors to life. Elements 13:261–265

Furukawa Y, Horiuchi M, Kakegawa T (2013) Selective stabilization of ribose by borate. Origins Life Evol Biosph 43:353–361

Furukawa Y, Kim HJ, Hutter D, Benner SA (2015) Abiotic regioselective phosphorylation of adenosine with borate in formamide. Astrobiology 15:259–267

Gabel NW, Ponnamperuma C (1967) Model of origin of monosaccharides. Nature 216:453–455

Gilbert W (1986) The RNA world. Nature 319:618

Hu H, Xue J, Wen X, Li W, Zhang C, Yang L, Xu Y, Zhao G, Bu X, Liu K, Chen J, Wu J (2013) Sugar-metal ion interactions: the complicated coordination structures of Cesium ion with D-ribose and myo-inositol. Inorg Chem 52:13132–13145

Kim H-J, Kim J (2019) A prebiotic synthesis of canonical pyrimidine and purine ribonucleotides. Astrobiology. https://doi.org/10.1089/ast.2018.1935

Lambert JB, Lu G, Singer SR, Kolb VM (2004) Silicate complexes of sugars in aqueous solution. J Am Chem Soc 126:9611–9625

Larralde R, Robertson MP, Miller SL (1995) Rates of decomposition of ribose and other sugars: implications for chemical evolution. Proc Natl Acad Sci USA 92:8158–8160

Maurel M-C, Leclerc F (2016) From foundation stones to life: Concepts and results. Elements 12:407–412

Menor-Salván C, Marín-Yaseli MR (2013) A new route for the prebiotic synthesis of nucleobases and hydantoins in water/ice solutions involving the photochemistry of acetylene. Chem Eur J 19:6488–6497

Miyawaki S, Murasawa K, Kobayashi K, Sawaoka AB (2000) Abiotic synthesis of guanine with high-temperature plasma. Origin Life Evol Biosph 30:557–566

Nam I, Nam HG, Zare RN (2018) Abiotic synthesis of purine and pyrimidine ribonucleosides in aqueous microdroplets. PNAS 115:36–40. https://doi.org/10.1073/pnas.1718559115

Orgel LE (2004) Prebiotic chemistry and the origin of the RNA world. Crit Rev Biochem Mol Biol 39:99–123

Powner MW, Gerland B, Sutherland JD (2009) Synthesis of activated pyrimidine ribonucleotides in prebiotically plausible conditions. Nature 459:239–242

Quesada-Moreno MM, Azofra LM, Avilés-Moreno JR, Alkorta I, Elguero J, López-González JJ (2013) Conformational preference and chiroptical response of carbohydrates D-ribose and 2-deoxy-D-ribose in aqueous and solid phases. J Phys Chem B 117:14599–14614

Ricardo A, Carrigan MA, Olcott AN, Benner SA (2004) Borate minerals stabilize ribose. Science 303:196

Rich A (1962) On the problems of evolution and biochemical information transfer. In: Kasha M, Pullman B (eds) Horizons in biochemistry. Academic Press, New York, pp 103–126

Robertson MP, Miller SL (1995) An efficient prebiotic synthesis of cytosine and uracil. Nature 375:772–774

Robertson MP, Joyce GF (2012) The origins of the RNA world. Cold Spring Harb Perspect Biol 4:a003608. https://doi.org/10.1101/cshperspect.a003608

Saladino R, Bizzarri BM, Botta L, Šponer J, Šponer JE, Georgelin T, Jaber M, Rigaud B, Kapralov M, Timoshenko GN, Rozanov A, Krasavin E, Timperio AM, Mauro ED (2017) Proton irradiation: a key to the challenge of N-glycosidic bond formation in a prebiotic context. Sci Rep 7:14709. https://doi.org/10.1038/s41598-017-15392-8

Saladino R, Carota E, Botta G, Kapralov M, Timoshenko GN, Rozanov AY, Krasavin E, Mauro ED (2015) Meteorite-catalyzed syntheses of nucleosides and of otherprebiotic compounds from formamide under proton irradiation. PNAS 112:E2746–E2755. https://doi.org/10.1073/pnas.1422225112

Saladino R, Neri V, Crestini C (2010) Role of clays in the preobiotic synthesis of sugar derivatives from formamide. Phil Mag 90:2329–2337

Sanchez RA, Ferris JP, Orgel LE (1966a) Conditions for purine synthesis: did prebiotic synthesis occur at low temperature? Science 153:72–73

Sanchez RA, Ferris JP, Orgel LE (1966b) Cyanoacetylene in prebiotic synthesis. Science 154:784–785

Šišak D, McCusker LB, Zandomeneghi G, Meier BH, Bläser D, Boese R, Schweizaer WB, Gilmour R, Dunitz JD (2010) The crystal structure of D-ribose- At last! Angew Chem Int Ed 49:4503–4505

Theng BKG (2018) Clay mineral catalysis of organic reactions. CRC Press, Boca Raton (FL)

Wang W, Huang F, Sun C, Liu J, Sheng X, Chen D (2017) A theoretical insight into the formation mechanisms of C/N-ribonucleosides with pyrimidine and ribose. Phys Chem Chem Phys 19:10413–10426

Xu J, Tsanakopoulou M, Magnani CJ, Szabla R, Šponer JE, Šponer J, Góra RW, Sutherland JD (2017) A prebiotically plausible synthesis of pyrimidine β-ribonucleosides and their phosphate derivatives involving photoanomerization. Nat Chem 9:303–309. https://doi.org/10.1038/NCHEM.2664

Chapter 3
The First Universal Common Ancestor (FUCA) as the Earliest Ancestor of LUCA's (Last UCA) Lineage

Francisco Prosdocimi, Marco V. José and Sávio Torres de Farias

Abstract The existence of a common ancestor to all living organisms in Earth is a necessary corollary of Darwin idea of common ancestry. The last universal common ancestor (LUCA) has been normally considered as the ancestor of cellular organisms that originated the three domains of life: Bacteria, Archaea and Eukarya. Recent studies about the nature of LUCA indicate that this first organism should present hundreds of genes and a complex metabolism. Trying to bring another of Darwin ideas into the origins of life discussion, we went back into the prebiotic chemistry trying to understand how LUCA could be originated under gradualist assumptions. Along this line of reasoning, it became clear to us that the definition of another ancestral should be of particular relevance to the understanding about the emergence of biological systems. Together with the view of biology as a language for chemical translation, on which proteins are encoded into nucleic acids polymers, we glimpse a point in the deep past on which this translation mechanism could have taken place. Thus, we propose the emergence of this process shared by all biological systems as a point of interest and propose the existence of this pre-cellular entity named FUCA, as the first universal common ancestor. FUCA was born in the very instant on which RNA-world replicators started to be capable to catalyze the bonding of amino acids into oligopeptides. FUCA has been considered mature when the translation system apparatus has been assembled together with the establishment of a primeval, possibly error-prone genetic code. This is FUCA, the earliest ancestor of LUCA's lineage.

F. Prosdocimi (✉)
Laboratório de Biologia Teórica e de Sistemas, Instituto de Bioquímica Médica Leopoldo de Meis, Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brazil
e-mail: prosdocimi@bioqmed.ufrj.br

M. V. José
Theoretical Biology Group, Instituto de Investigaciones Biomédicas, Universidad Nacional Autónoma de México, Ciudad Universitaria, 04510 Mexico City, Mexico

S. T. de Farias (✉)
Laboratório de Genética Evolutiva Paulo Leminsk, Departamento de Biologia Molecular, Universidade Federal da Paraíba, João Pessoa, Paraíba, Brazil
e-mail: stfarias@yahoo.com.br

P. Pontarotti (ed.), *Evolution, Origin of Life, Concepts and Methods*,
https://doi.org/10.1007/978-3-030-30363-1_3

3.1 Historical Background

In the mid of the nineteenth century, Charles Darwin proposed a point of unification for all living beings through the idea of common descent (Darwin 1859). Concomitantly with the Darwin's postulates, another strong idea was under development: the cell theory. The explanation power of cell theory established the paradigm that cells were the most basic units of life. Together, these two powerful theories were united into the idea of a single organism that should represent the universal ancestor of all living beings. Named after the last universal common ancestor (LUCA), this organism would necessarily be composed of a structured cellular unit.

The comprehension about the nature of LUCA was strikingly important and allowed researchers to better understand the constitution of the first cells that most likely emerged in Earth about 3.8 billion years ago. LUCA is currently seen as a cellular organism that presents a lipid membrane and a complex metabolism composed of hundreds of gene families and dozens of biochemical pathways (Penny and Poole 1999; Delaye et al. 2005; Weiss et al. 2016). Although comparative genomics allows us to have a glimpse about the molecular nature of this important cellular ancestral, it is clear that some simpler biological system must have preceded and gave rise to LUCA.

About one century after Darwin, an original and prolific field of biological research started to develop shortly after the discovery of DNA double-stranded structure by Rosalind Franklin, James Watson, and Francis Crick (1953). The development of molecular biology modified our comprehension of the biological world and allowed scientists to perform experiments using genes and proteins inside and outside the cell, bringing a deeper understanding about how biological processes operated in the biochemical level.

In the very same year that Watson and Crick published their paper, a young American researcher named Stanley Miller was inspired by the ideas of Aleksandr Oparin and JBS Haldane in the 1920s. They had demonstrated that the assembly of basic chemical molecules under specific physicochemical conditions could produce amino acids, the building blocks of proteins. Miller's experiment updated the simulation of primeval Earth conditions and confirmed the production of amino acids using only water (H_2O), methane (CH_4), ammonia (NH_3), and hydrogen (H_2) as input (Miller 1953). Thus, a point of view about the initial organization of the biological systems started as the studies of prebiotic chemistry provided an innovative way to think about the origins of life.

Then, by the early years of the 1980s, the discovery of catalytic properties of RNA molecules introduced another element through which the initial organization of biological systems could be understood (Kruger et al. 1982; Guerrier-Takada et al. 1983). For the first time, these findings made possible the proposition of hypotheses capable to describe biological entities that did not need cellular structures. This idea culminated with the proposition of a molecular-based *RNA world* on which self-replicative and catalytic molecules of nucleic acids could interact and be target of natural selection, pushing forward the path into the origins of life (Gilbert 1986).

3.1.1 The Lineage of LUCA

It has been shortly before the proposal of the RNA world theory, in the late 1970s, that the American microbiologist Carl Woese started to produce sequences for a specific RNA molecule known as 16S rRNA, a constituent of the small ribosomal subunit (Woese and Fox 1977a). Containing about ~1500 nucleotides, Woese found specific oligonucleotide signatures in the sequences of the rRNAs that were capable to differentiate molecules coming from either Bacteria or Eukarya organisms.

Woese's interest in the 16S rRNA was associated with the fact that ribosomes were abundant in cells and could be easily separated and radiolabeled by the molecular techniques available. However, when he sequenced for the first time the rRNA from a methanogen organism named *Methanobacterium thermoautotrophicum*, he was unable to find those expected signatures … But what should this result mean: did he found an organism that was neither Bacteria nor Eukarya?

By the end of 1976, Woese and George Fox were capable to sequence other methanogen organisms that could not be classified in these two groups based on their 16S rRNA sequences, confirming their previous results. On the other hand, this new group presented a different type of oligonucleotide signature that was observed to be common among them. Therefore, the 16S rRNA molecule revealed to present a bipartite diversity in the prokaryotic world. And it then became clear that a completely unknown group of organisms existed: a new, different, and unexpected clade of unicellular organisms.

Due to the fact that most members of this new group of prokaryotes lived in extreme environment, Woese and Fox believed them to be the closest relatives of LUCA, the most ancient form of life that lived in Earth. In the early days of our planet, it was known that the atmosphere was completely different, most likely based on a reductive environment, hot and possibly very dry. They discovered that the 16S rRNA of a bacteria living in high-salt environments (an halophile) clustered also into this new group of organisms. The results suggested that most members of the third group inhabited extreme, prebiotic-like environments. This led the authors to name the new clade as the Archaebacteria (or simply Archaea) by the use of the Greek word *archae-*, meaning ancient. Thus, a new urkingdom of microbes was defined and it has probably participated closely to the heritage of LUCA.

However, further research on Archaebacteria conducted by Woese himself and collaborators changed the picture initially proposed (Woese et al. 1990). Actually, the Archaea could be found not only in extreme but also in typical mesophile environments. Studying other phenotypes of this group, it became clear that they were actually a missing link between Bacteria and Eukarya. It has been found that Archaea presented characteristics shared with Bacteria, such as (i) prokaryotic ribosomes; (ii) circular chromosomes; and (iii) lack of membrane-enclosed organelles; but also, others shared with Eukaryotes, such as a (i′) DNA associated to histones; (ii′) several types of RNA polymerases; and (iii′) the use of methionine as the initiator amino acid from protein synthesis; among others.

These observations now seemed to indicate that the early cellular ancestor (LUCA) was probably a population of ancestral organisms that presented more similarities to Bacteria than to Archaea. Ever since, the origin of life and the origin of LUCA were coincidently placed in the tree of life as the middle point of a trichotomy in-between Bacteria, Archaea, and Eukarya domains. More recently, however, LUCA is been considered as the ancestor of only Bacteria and Archaea, while Eukarya are being considered as a group that evolved from a lineage inside the archaean superphylum Asgard (Williams et al. 2012; Raymann et al. 2015; Zaremba-Niedzwiedzka et al. 2017).

The most recent genomic study on the deep nature of LUCA's genome has found that this ancestral organism should present at least 355 gene families, being thermophilic, anaerobic, capable to fix CO_2 and N, and possibly H_2-dependent (Weiss et al. 2016).

3.1.2 LUCA and the Viruses

The questions about the deep lineage of LUCA are complex to approach, once non-cellular living organisms do not exist in current days. The sole non-cellular biological systems that exist today are viruses. Viruses are frequently considered as non-living organisms because they are not free-living entities, as they need to highjack cells to be able to manifest their metabolism and reproduce. Besides, viruses have been excluded from deep trees of life by the simple reason that they do not have ribosomes; and without 16S rRNA molecules they cannot figure in those trees.

However, nowadays there is an emerging view that virus-like biological systems may have played important roles in pre-cellular living systems (Forterre 2006). Besides, the evolutionary history of viruses seems to be polyphyletic as there is evidence that some groups were formed by the further simplification of cellular organisms. Thus, virus should be understood much more as *a strategy of life* rather than a monophyletic group that share the same evolutionary origin (Nasir and Caetano-Anolles 2015). Recent researches with giant mimivirus seem to indicate polyphyly, although there is still much controversy in those grounds (Harish et al. 2016; Forterre and Gaia 2016). By the way, it is difficult to use the word "life" and "living" to represent viruses and it seems reasonable to consider that virus-like biological systems may have existed long before lipid cellular barriers were coopted to be the basis of life in Earth.

3.2 Life Versus Biological Systems

The suggestion of an alternative view to the origins of life requires that we take a special attention to the concepts we are working on. The concept of life, as commonly understood in biology, is normally considered as cellular life. Although there is

constant matter and new appreciations and proposals about the concept, we cannot argue against a so-established concept. On the borders of our understanding, however, there has always been this question about the nature of viruses. Current considerations under virology often consider that viruses can only be said to be alive when they are practicing their metabolism inside a cell. Out of cells, viruses are considered to be non-living crystals of ribonucleoproteins.

Although viruses do not present cells, they do present something that is strikingly important to any biological system: a well-established genetic code. That code is actually the key to provide a chemical translation and guarantee that their nucleic acid information will produce proteins that will allow their metabolism and reproduction. Viruses may not be cellular, viruses may not be living, but viruses do speak the language of biology. They do present complex proteins encoded in their genomes and they do have a plan of existence chemically written in their nucleic acids (being DNA or RNA).

Though they do not have ways to execute their plans by their own means, it is possible to consider a world made of non-cellular, virus-like entities that might deliver their codes into a translation system operating outside them. Thus, it is not difficult to imagine pre-cellular virus-like particles capable to exist and to attach into other systems that might translate their information and allow their reproduction.

Even Carl Woese knew that pre-cellular entities were needed to explain the origin of life in Earth before LUCA. He defined the concept of progenotes as being protocells that probably presented (i) error-prone genetic codes, (ii) high mutation rates, and (iii) high exchange of genetic material (Woese and Fox 1977b). These entities were necessary to accumulate genetic material from different biochemical pathways to be able to form larger genomes.

Thus, we propose the usage of the term *Biological system* as an alternative for life. This decision has proven to be fertile and clarifying in many circumstances, as we shall see. However, to make a better use of this term, we need to clearly define it. For us, *a biological system is such a system on which its molecular nature is centered in the controlled interaction between polymers of nucleotides (nucleic acids) and polymers of amino acids (peptides and/or proteins)*. In general, the biological system can be considered a system based in a process of chemical translation, on which information stored in some chemical polymer can be translated in another chemical polymer.

This view puts the process of protein synthesis in the center of what we consider biology to be. All biology is based on a process of chemical encoding, a system of chemical language translation on which a polymer become another by following specific grammatical rules. Thus, biological systems maturated in the very moment on which those rules emerged.

3.2.1 Origins Versus Emergence

Another important conceptual issue that must be better understood into this new view of what are biological systems is the difference between the terms *origin* and *emergence*. The term origin often refers to the first rise of something that has never

been before. It suggests the occurrence of a very singular, special event that, for the first time, brings something from non-existence to existence. It is a jump, a rising, an appearance.

On the other hand, the concept of emergence should be understood as a more continuous path on which something comes to be. It is not a drastic appearance as the concept of origin denotes, but a more subtle process on which a system can be built slowly and persistently over time.

Also, inheriting the concept of emergence from complex systems theory, we aim to signify a process on which "the whole is greater than the sum of the parts." This is clearly what happened at the emergence of biological systems, on which a polymer of nucleic acids starts to produce peptides initially by nearly random attachments until the emergence of the genetic code, throughout the history of this early process that we named as FUCA.

3.3 Biological Systems are Chemical Translators

Although it is possible to wonder the existence of non-cellular biological systems, it is almost impossible to think in a biological entity harboring a metabolism without the translation apparatus for protein synthesis. Having (i) proteins encoded in a three-letter genetic code made of nucleic acids and a (ii) metabolism primarily coordinated by proteins and enzymes seem to be features that unequivocally define a biological system.

Both RNA-world advocates and researchers that propose the origin of life by hypercycles (the theory suggesting that biology started with a protein-based protometabolism happening without codification mechanisms) have difficulties to go forward within their theories by adding the other macromolecule into them.

When we consider RNA-world ribozymes capable of both self-replication and catalysis (Higgs and Lehman 2015), it is difficult to see how a ribozyme-based protometabolism could further become proteic. It seems more parsimonious to think of a biological system that already starts together, with peptides interacting with proto-RNAs or RNA-like polymers. Proteins are the most important molecules to coordinate the cellular metabolism, interacting with compounds, and catalyzing chemical reactions, even if the importance of RNAs is also enormous. Even nowadays, the importance of RNAs to the cellular metabolism has been underestimated and new results from the ENCODE project suggest that three-quarters of the human genomic DNA should be transcribed at some moment (Djebali et al. 2012).

The molecule of DNA, however, seems to have a very limited importance at the emergence of biological systems. Actually, it is somewhat consensual that DNA has been added to the system much time later (Forterre 2002, 2006, 2013). Double helix polymers of the desoxyribonucleic acid that nowadays compose the genomes of most organisms can be seen much more like a *high-security media* to store molecular information that emerged later and allowed an important stabilization for both heredity and the control of metabolism. Although its importance has been inestimable, it

was probably absent at the scenario on which the biological systems emerged. The existence of viruses harboring RNA genomes, even in single strand, can be seen as evidence that DNA is not essential to biological systems. Being a secure media to store molecular information, it has been placed in a safe place in the nucleus of eukaryotic cells, though it seems to have no role in the beginnings.

The emergence of a chemical translation process on which RNA-like molecules convert their information into peptides seems to be the main characteristic shared by any biological system and therefore seem to have evolved from a common origin at the first universal common ancestor (FUCA).

Once we suggest that the most realistic scenario for the emergence of biological systems would happen on the interaction between RNAs and proteins, the importance to explain the origins of the ribosome and the genetic code turns key. However, the ribosomes present today in cells are highly specialized and complex molecular machines and it is necessary to understand their old history of development.

3.3.1 The Molecular Establishment of the Genetic Code

Chemically speaking, the genetic code on which codons encode amino acids is molecularly produced by the action of extremely important enzymes named tRNA-aminoacyl syntethases. There are two different families for these proteins and they are responsible to bind each amino acid to the transfer RNAs presenting not only specific anti-codons but also other conserved binding sites on their cloverleaf structure (Zamudio and José 2018). These enzymes are of paramount importance for the understanding of the origin and evolution of the genetic code. However, in the very early time of biology, we should suppose that the translation system did not functioned very well and codon to peptide encoding should have been performed with no specificity, without a clear system of codification.

In that sense, the birth of biological systems is not at the emergence of the code, but at the rise of some sort of ribozyme that was capable to bind together amino acids. FUCA is, therefore, a process that started with no code, but her point of maturation happened when the code was completely established. Thus, to understand the initial steps in the emergence of FUCA we need to study deeply the anatomy of ribosomes.

3.3.2 The Anatomy of Ribosomes

The ribosomes of any biological system present two subunits. The smaller one binds the messenger RNA while the larger presents three sites for the binding of transfer RNAs bringing specific amino acids to the system.

The exact ribosomal site on which amino acids are bound together is considered as its catalytic center. This catalytic center has been named as peptidyl transferase center, or simply PTC. The PTC is part of the 23S rRNA of bacteria and it is known

to catalyze the binding of two separate amino acids into a dipeptide using an Adenine as the most important catalytic center. In this synthesis by dehydration, a molecule of water is jumped out while the C-terminal of an amino acid binds to the N-terminal of another, starting with the polymerization. Other amino acids can be further added to this dipeptide, raising the number of amino acids that can be bound together and allowing the formation of oligopeptides.

In our view, the appearance of a nucleic acid molecule in the early Earth that was capable to bind together two amino acids, somehow started a process of chemical symbiosis on which the binding of this nucleic acid to the peptides produced by itself allowed the system to both (i) stabilize under a self-referential perspective and (ii) aggregate complexity in layers, probably through the mechanism known as accretion.

3.4 FUCA Is Born at the Proto-PTC

Here, we propose that the biological systems originated in the very moment on which a macromolecule of nucleic acid containing dozens of nucleotides were capable to fold in the 3D space and catalyze the junction of two amino acids into a dipeptide. Thus, FUCA was born when a proto-PTC started to emerge for the first time, allowing already existing self-replicating nucleic acid polymers to produce random di- and oligopeptides. These random peptides produced possibly bound back to the single-strand nucleic acid polymers and allowed a higher stabilization of the system that got more robust and was further bound to other stabilizing molecules.

Chaos theory advocates might say that a strange attractor should have been formed at that point. Astrophysics uses the term accretion to explain how planets and other bodies may be formed by the aggregation of material in overlapping layers, due to the force of gravity. Here, we also use the idea of accretion to explain how the ribosome has been assembled from the proto-PTC into a higher-level, multiple layer system. Although the first peptides were bound randomly, the process itself cannot be seen as random because there should have been very simple and specific types of amino acids existing in the primordial pool of molecules. These early amino acids were most likely Glycine, Serine, Alanine, and others (Miller 1953; Paker et al. 2011). Their bound together attracted other amino acids and have possibly formed a first layer of peptides that bound to the proto-PTC and stabilized its interaction to the proto-tRNAs. After subsequent layers of complexification, these processes would evolve to the creation of the larger ribosomal subunit with its tRNA sites A, P, and E. Many works have already tried to understand how this subunit has been clearly formed; some considering that PTC were at the beginning (Petrov et al. 2015) and others presenting evidences that the ribosome structure started elsewhere (Caetano-Anolles 2015). It is consensual; however, that the process of ribosome assembly took a long time until it could became functional and efficient in its task of amino acid binding.

3.5 The Maturation of FUCA

The maturation of FUCA happened when the basis for what we call today the genetic code has been achieved. With the basis for the code, we aim to describe the three types of ancestral RNAs working as a system for encoding (mRNA), decoding (tRNA), and binding amino acids (rRNA). Recent studies seem to indicate that the same module of 30–70 nucleotides should have operated to produce initially a tRNA-like molecule. Specific duplications and head-to-tail concatenations of this initial polymeric module might have produced both the proto-PTC and the proto-messenger (Farias et al. 2014, 2016). And by the further attachment of specific proto-peptides produced by the own system some polynucleotidic modules formed each time more stable linear molecules, cloverleaf structures or ribozymes.

The first universal common ancestor (FUCA) is, therefore, an ancestor of LUCA's lineage. It was born when self-replicating polymers of RNA-like nucleotides started to bind amino acids, and its maturation happened with the establishment of the genetic code.

We understand the birth of FUCA as a (i) process of chemical symbiosis and as a (ii) revelation about the importance of symbiotic processes to the emergence of the most fundamental biological process. The idea of an ancestral of LUCA's lineage is important for the understanding that FUCA emerged much earlier than LUCA.

The emergence of a proto-PTC has been a contingent moment of enlighten to the creativity of the universe and to the union of macromolecules. At that time, molecules could only collaborate by interacting through binding, but never before by building other collaborator molecules. When these ancient nucleic acids were capable to domesticate the abundant amino acids and interact with them achieving more stabilization than each one by itself, a new age has risen and FUCA has been born.

The molecular nature of the universe has discovered that molecules could collaborate and help themselves while helping others. A moral molecular imperative has become truth. Instead of competing and destructing each other in a RNA-based world on which replicators destroyed each other to get their monomers and build their own copies (Dawkins 1978), some other replicators found that binding amino acid molecules together produced some new form of mutual stabilization. A chemical symbiotic relationship emerged as one polymeric macromolecule has now helped the other in a world of chemical war. Together, their stable complex aggregated other ions and molecules, allowing new layers to be produced and augmenting the interaction and balance between themselves.

This system had specialized to be the very language and grammar of biology, the language of chemical interchange that would further evolved to the formation of the complete ribosome, the genetic code, and the maturation of the molecular translational process.

3.6 Conclusions

We do not aim here to convince the readers about the ideas presented here, as our intention is only to present interesting and inspiring theoretical interpretations about how the biological systems may have emerged. We still know very few about our deep molecular origins and we felt that we need to focus in the tentative to explain how these wonderful, nearly chaotic phenomena that gave rise to the biological systems might have been originated. Our exploration surely aims to be scientific and many propositions presented here are being tested in laboratories all over the world under the scrutiny of modern science. Much more hypotheses can be further tested experimentally and also by the use of theoretical models to rebuild ancestral molecules, and empirically resurrect them. In this current essay, our aim was also to propose fundamentally philosophical and epistemological ideas about the deep origins of biological systems that would further produce life and suggest interesting points for research.

The view of (i) life as cellular and the (ii) proposal of the last universal common ancestor (LUCA) as the branching point on which cellular organisms originated was very important at their time. The very idea of both LUCA and FUCA are clearly to be considered as a corollary of Darwin's core idea regarding the common ancestry among all living organisms. In one of the most important experiments in the history of science, the Darwin contemporary Louis Pasteur has demonstrated that biogenesis could not happen spontaneously and organisms need other organisms to arise. The chain of life is therefore linked together back to LUCA.

There is an agreement among scholars studying the origins of life that LUCA should be considered the ancestral of cellular organisms. However, the most recent proposal about LUCA's genome reveals a highly complex cellular organism with about 355 gene families working together to produce life into a cellular organism.

Being enlightened by Darwinian ideas, it has been our aim to bring gradualism under the theoretical research about the origins of life. This brought us back to consider the origin of life from the point of prebiotic chemistry and we propose here that biological systems started to emerge when RNA-like molecules started to bind amino acids together. This is the place on time when FUCA has been born. The history of FUCA development, however, has probably taken a long time. We consider her to be maturated at the moment on which the system known nowadays as translation has been completely developed; together with a functional genetic code responsible to translate the information present in a nucleic acid into a peptide under an organized form.

Acknowledgements We would like to thank FAPERJ (CNE E-26/202.780/2018) for funding FP. MVJ was financially supported by PAPIIT-IN224015; UNAM; México.

Conflicts of Interest The authors declare no conflict of interest.

References

Caetano-Anollés G (2015) Ancestral insertions and expansions of rRNA do not support an origin of the ribosome in its peptidyl transferase center. J Mol Evol 80(3–4):162–165. https://doi.org/10.1007/s00239-015-9677-9

Darwin CR (1859) On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life, 1st edn. John Murray, London

Dawkins R (1978, ©1976) The selfish gene. Oxford University Press, New York

Delaye L, Becerra A, Lazcano A (2005) The last common ancestor: what's in a name? Orig Life Evol Biosph 35(6):537–554

Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi A, Tanzer A, Lagarde J, Lin W, Schlesinger F, Xue C, Marinov GK, Khatun J, Williams BA, Zaleski C, Rozowsky J, Röder M, Kokocinski F, Abdelhamid RF, Alioto T, Antoshechkin I, Baer MT, Bar NS, Batut P, Bell K, Bell I, Chakrabortty S, Chen X, Chrast J, Curado J, Derrien T, Drenkow J, Dumais E, Dumais J, Duttagupta R, Falconnet E, Fastuca M, Fejes-Toth K, Ferreira P, Foissac S, Fullwood MJ, Gao H, Gonzalez D, Gordon A, Gunawardena H, Howald C, Jha S, Johnson R, Kapranov P, King B, Kingswood C, Luo OJ, Park E, Persaud K, Preall JB, Ribeca P, Risk B, Robyr D, Sammeth M, Schaffer L, See LH, Shahab A, Skancke J, Suzuki AM, Takahashi H, Tilgner H, Trout D, Walters N, Wang H, Wrobel J, Yu Y, Ruan X, Hayashizaki Y, Harrow J, Gerstein M, Hubbard T, Reymond A, Antonarakis SE, Hannon G, Giddings MC, Ruan Y, Wold B, Carninci P, Guigó R, Gingeras TR (2012) Landscape of transcription in human cells. Nature 489(7414):101–108. https://doi.org/10.1038/nature11233 (PubMed PMID: 22955620; PubMed Central PMCID: PMC3684276)

Farias ST, do Rêgo TG, José MV (2014) Evolution of transfer RNA and the origin of the translation system. Front Genet 5:303. https://doi.org/10.3389/fgene.2014.00303 (eCollection 2014)

Farias ST, Rêgo TG, José MV (2016) tRNA core hypothesis for the transition from the RNA world to the ribonucleoprotein world. Life (Basel) 6(2):pii: E15. https://doi.org/10.3390/life6020015

Forterre P (2002) The origin of DNA genomes and DNA replication proteins. Curr Opin Microbiol 5(5):525–532 (Review. PubMed PMID: 12354562)

Forterre P (2006) The origin of viruses and their possible roles in major evolutionary transitions. Virus Res 117(1):5–16 (Epub 14 Feb 2006. Review. PubMed PMID: 16476498)

Forterre P (2013) The great virus comeback. Biol Aujourdhui 207(3):153–168. https://doi.org/10.1051/jbio/2013018 (Epub 13 Dec 2013. Review. French. PubMed PMID: 24330969)

Forterre P, Gaïa M (2016) Giant viruses and the origin of modern eukaryotes. Curr Opin Microbiol 31:44–49. https://doi.org/10.1016/j.mib.2016.02.001

Gilbert W (1986) The RNA world. Nature 319:618

Guerrier-Takada C, Gardiner K, Marsh T, Pace N, Altman S (1983) The RNA moiety of ribonuclease P is the catalytic subunit of the enzyme. Cell 35(3 Pt 2):849–857

Harish A, Abroi A, Gough J, Kurland C (2016) Did viruses evolve as a distinct supergroup from common ancestors of cells? Genome Biol Evol 8(8):2474–2481. https://doi.org/10.1093/gbe/evw175

Higgs PG, Lehman N (2015) The RNA world: molecular cooperation at the origins of life. Nat Rev Genet 16(1):7–17. https://doi.org/10.1038/nrg3841

Kruger K, Grabowski PJ, Zaug AJ, Sands J, Gottschling DE, Cech TR (1982) Self-splicing RNA: autoexcision and autocyclization of the ribosomal RNA intervening sequence of tetrahymena. Cell 31(1):147–157

Miller SL (1953) A production of amino acids under possible primitive earth conditions. Science 117(3046):528–529 (PubMed PMID: 13056598)

Nasir A, Caetano-Anollés G (2015) A phylogenomic data-driven exploration of viral origins and evolution. Sci Adv 1(8):e1500527. https://doi.org/10.1126/sciadv.1500527

Parker ET, Cleaves HJ, Dworkin JP, Glavin DP, Callahan M, Aubrey A, Lazcano A, Bada JL (2011) Primordial synthesis of amines and amino acids in a 1958 Miller H_2S-rich spark discharge experiment. Proc Natl Acad Sci USA 108(14):5526–5531. https://doi.org/10.1073/pnas.1019191108

Penny D, Poole A (1999) The nature of the last universal common ancestor. Curr Opin Genet Dev 9(6):672–677

Petrov AS, Gulen B, Norris AM, Kovacs NA, Bernier CR, Lanier KA, Fox GE, Harvey SC, Wartell RM, Hud NV, Williams LD (2015) History of the ribosome and the origin of translation. Proc Natl Acad Sci USA 112(50):15396–15401. https://doi.org/10.1073/pnas.1509761112

Raymann K, Brochier-Armanet C, Gribaldo S (2015) The two-domain tree of life is linked to a new root for the archaea. Proc Natl Acad Sci USA 112(21):6670–6675. https://doi.org/10.1073/pnas.1420858112 (Epub 11 May 2015. PubMed PMID: 25964353; PubMed Central PMCID: PMC4450401)

Watson JD, Crick FHC (1953) A structure for deoxyribose nucleic acid. Nature 171(4356):737–738

Weiss MC, Sousa FL, Mrnjavac N, Neukirchen S, Roettger M, Nelson-Sathi S, Martin WF (2016) The physiology and habitat of the last universal common ancestor. Nat Microbiol 1(9):16116. https://doi.org/10.1038/nmicrobiol.2016.116

Williams TA, Foster PG, Nye TM, Cox CJ, Embley TM (2012) A congruent phylogenomic signal places eukaryotes within the archaea. Proc Biol Sci 279(1749):4870–4879. https://doi.org/10.1098/rspb.2012.1795 (Epub 24 Oct 2012. PubMed PMID: 23097517; PubMed Central PMCID: PMC3497233)

Woese CR, Fox GE (1977a) Phylogenetic structure of the prokaryotic domain: the primary kingdoms. Proc Natl Acad Sci USA 74(11):5088–5090 (PubMed PMID: 270744; PubMed Central PMCID: PMC432104)

Woese CR, Fox GE (1977b) The concept of cellular evolution. J Mol Evol 10(1):1–6 (PubMed PMID: 903983)

Woese CR, Kandler O, Wheelis ML (1990) Towards a natural system of organisms: proposal for the domains archaea, bacteria, and eucarya. Proc Natl Acad Sci USA 87(12):4576–4579

Zamudio GS, José MV (2018) Identity elements of tRNA as derived from information analysis. Orig Life Evol Biosph 48(1):73–81. https://doi.org/10.1007/s11084-017-9541-6 (Epub 28 Jun 2017. PubMed PMID: 28660466)

Zaremba-Niedzwiedzka K, Caceres EF, Saw JH, Bäckström D, Juzokaite L, Vancaester E, Seitz KW, Anantharaman K, Starnawski P, Kjeldsen KU, Stott MB, Nunoura T, Banfield JF, Schramm A, Baker BJ, Spang A, Ettema TJ (2017) Asgard archaea illuminate the origin of eukaryotic cellular complexity. Nature 541(7637):353–358. https://doi.org/10.1038/nature21031 (Epub 11 Jan 2017. PubMed PMID: 28077874)

Part II
Concepts and Methods

Chapter 4
Repeatability and Predictability in Experimental Evolution

Peter A. Lind

Abstract Independent populations often use the same phenotypic and genetic solutions to adapt to a selective challenge, suggesting that evolution is surprisingly repeatable. This observation has inspired a shift in focus for evolutionary biology towards predictive studies, but progress is impeded by a lack of insight into the causes for repeatability, which prevents tests of forecasting models outside the original biological systems. Experimental evolution with microbes could provide a way to identify the causes of repeated evolution, directly test forecasting ability and develop methodology, but a range of difficulties limits successful prediction. This chapter discusses the limitations on forecasting of experimental evolution, what can and cannot be predicted on different biological levels and why predictions will often fail. Focusing on experimental populations of bacteria, the importance of selection, mutational biases and genotype-to-phenotype maps in determining evolutionary outcomes is discussed, as well as the potential for including these factors in forecasting models. The chapter concludes with a discussion on the desired properties of experimental evolution models suitable for testing forecasting models.

4.1 Introduction

The question of the predictability of evolutionary processes has a long history in evolutionary biology. Predictability is often described in terms of the relative importance of deterministic processes, typically natural selection, and stochastic events, including mutations, fluctuating environments and so-called historical contingencies (Blount et al. 2018; Lässig et al. 2017; Orgogozo 2015). Stephen Jay Gould's classic thought experiment of 'replaying the tape of life' (Gould 1989) to see if the results are the same (Gould betted they would not be) are now increasingly being realized with studies on parallel evolution in both natural and laboratory populations (Barrick and Lenski 2013; Bennett et al. 2009; Blount et al. 2012, 2018; Flowers et al. 2009; Gerstein et al. 2012; Herron and Doebeli 2013; McDonald et al. 2009; Stern

P. A. Lind (✉)
Department of Molecular Biology, Umeå University, SE-901 87 Umeå, Sweden
e-mail: peter.lind@umu.se

P. Pontarotti (ed.), *Evolution, Origin of Life, Concepts and Methods*,
https://doi.org/10.1007/978-3-030-30363-1_4

2013; Tenaillon et al. 2012; Weinreich et al. 2006; Zhen et al. 2012). These studies suggest that parallel evolution is quite common and that some aspects of evolution are surprisingly repeatable in the short term, at least when conditions are identical (Blount et al. 2018; Orgogozo 2015; Stern 2013). What is less clear is if this high repeatability could allow us to make forecasts of evolutionary outcomes in terms of fitness, phenotypes and genetics under similar but non-identical conditions.

If evolutionary processes are often repeatable it might, in theory, be possible to make forecasts about evolutionary outcomes and perhaps also intervene to produce a desired outcome (Lässig et al. 2017). This would not only shift evolutionary biology's focus from retrospective studies, but also open up possibilities for high-value applications. These include important problems in medicine, such as predicting antibiotic resistance, vaccine development and personalized cancer prognosis and treatment, as well as forecasting the ability of organisms to adapt to rapidly changing climates. However, high repeatability does not guarantee predictability. Obviously, any prediction will be probabilistic given the stochastic influence of mutational processes and environmental conditions. But it is also possible that the data required for forecasting are not available and must be measured in detail to be able to predict outcomes. For example, a population genetics model that requires knowledge of standing genetic variation in the population, the rates of new mutations and the fitness effects of all mutations in a particular environment is unlikely to be suitable for forecasting. Thus, evolutionary models can be highly predictive and provide an accurate description of reality without being useful for forecasting. Of course, such models may have other uses such as in helping to understand complex problems or for retrospective modelling. Fitness landscapes with differing ruggedness are useful for understanding adaptation, but realistic fitness landscapes are highly multidimensional and their use in forecasting would require either detailed fitness measurements or a way to a priori predict their shape (de Visser et al. 2018; de Visser and Krug 2014). Recent theoretical work has proposed the existence of hard fitness landscapes where even local fitness peaks cannot be found in a moderate number of steps and where fitter mutants will always be available, which could further limit the use of fitness landscapes in forecasting (Kaznatcheev 2019).

A way forward could be to use machine learning to process large amounts of data about the history of an evolutionary process to forecast the next steps. Here, no explicit model of the underlying evolutionary processes is required, only a degree of repeatability. This is also a weakness as what happens in the 'black box' of, for example, neural networks does not easily produce an understanding of the underlying causes, which is required if an intervention is to be possible or in order to produce predictions for novel situations. Many machine learning methods might also produce discoveries that are irreproducible, as they cannot assess the uncertainty of their predictions.

Most progress in forecasting evolutionary outcomes has been achieved for predominantly asexual processes under strong selection and with high supplies of mutations (Lässig et al. 2017). Examples include fitness models for forecasting of influenza epitopes (Luksza and Lassig 2014; Neher et al. 2014), immune cell receptors (Liao et al. 2013) and cancer (Fischer et al. 2014). These types of models provide

a fertile middle ground between too complex models needing inaccessible information and machine learning approaches using only repeatability. However, they are less well suited to produce predictions outside their specific context or to test fundamental assumptions about evolutionary processes and combining these models with more mechanistic models could prove fruitful. Experimental laboratory evolution with microbes has provided many examples of repeatability of evolutionary processes [reviewed in Blount et al. (2018), Long et al. (2015)] and might provide a way to determine the causes of repeated evolution. The aim of this chapter is to explore the potential to use experimental evolution to directly test forecasting models on different biological levels and identify difficulties and limitations that hinders progressing from repeatability to predictability.

4.2 Factors Limiting Predictability in Experimental Evolution

Experimental evolution has long been used to demonstrate the power of evolution in real time under controlled environmental conditions, but recent developments in DNA sequencing technologies have led to a renaissance that has provided experimental support for the repeatability of evolution (Blount et al. 2018; Lässig et al. 2017; Long et al. 2015). At a first glance, experimental evolution seems perfectly suited to explore evolutionary predictability and also to determine the underlying causes of repeated evolution. Evolution in the laboratory can take place under strictly controlled environmental conditions, population sizes and genetic backgrounds (Van den Bergh et al. 2018). Although studies have included sexual reproduction in both microbes and higher organisms (Kawecki et al. 2012), the focus here will be on studies using bacterial populations starting from a clonal population without standing genetic variation, no horizontal gene transfer or migration and under strong selection for adaptation in large populations. This limitation allows asking the most fundamental questions about the possibilities of forecasting evolution and identify what can be predicted, on what time scales, what we need to know and what is missing? What can we hope to be able to predict in the future and what are the properties of experimental evolution model systems that are suitable for testing forecasting ability?

4.2.1 Historical Contingency and Determinism in Experimental Evolution

The definitive review on determinism and contingency in experimental evolution has recently been published by Blount et al. (2018) and this review can add little to their work. Their thorough examination of 51 experimental evolution studies suggest that in a majority of studies historical contingency matters, at least on the genetic level.

So what is historical contingency and how does it specifically matter in experimental evolution to limit our predictive ability? Blount et al. (2018) discuss this in great detail and concludes that 'Historical contingency is a complicated notion, but it essentially means that history matters'. Thus, in order to predict experimental evolution, we must include history in some way, but simply blaming our failure to predict the outcome on history will not suffice. The underlying causes must be understood to determine if and how they can be incorporated into forecasting models.

4.2.2 Influence of Genetic Background

Differences in history lead to accumulation of genetic differences which means that, in general, evolutionary outcomes of experimental evolution are more likely to differ with increasing phylogenetic distance. That two strains from distantly related species, such as *Escherichia coli* and *Staphylococcos aureus*, would not respond to an adaptive challenge in the same way is therefore not surprising. More surprising is that closely related strains of *E. coli* often do not evolve mutations in the same genes under similar conditions even though experimental evolution for each individual strain is highly repeatable (Wang et al. 2018). Clearly history matters. History in these cases however often involves a long time in a laboratory environment, sometimes also including rounds of chemical or physical mutagenesis (Blattner et al. 1997; Daegelen et al. 2009), leading to a range of adaptive and deleterious mutations fixed in the population. This means that in many cases, the mutations detected after experimental evolution are unique to that particular strain simply because they are reversing or compensating for a mutation that would rarely occur in natural populations. Examples include *argR*, *mrdA*, *ybaL*, *rbsD*, *hslU*, *iclR* in *E. coli* REL606, *hfq*, *rph-pyrE* in *E. coli* MG1655 (www.mutationdb.com) (Wang et al. 2018) and *malQ*, *treB* in *Salmonella enterica* serovar Typhimurium LT2 (Knoppel et al. 2018). This shows that evolution might be completely repeatable and we can understand the underlying causes, but the outcome would not contribute to a useful forecast for another related strain.

4.2.3 Environmental Influence

Even if experimental evolution is often highly repeatable under identical conditions, this does not necessarily allow us to predict evolution in a slightly different context, which is required for development of useful applications. For example, adaptation to different temperatures in the same medium produced largely divergent mutational patterns (Deatherage et al. 2017). Partly, this might be due to different requirements to compensate for deleterious mutations at different growth rates or simply that the relative fitness effects vary between environments so that the probable order of mutations is likely to be different. If such environmental sensitivity is common, it will severely limit direct tests of evolutionary forecasts. For example, if evolutionary

outcomes vary depending on minor differences in growth medium composition, temperature, pH, time between bottlenecks or population size (Bailey et al. 2017), there is little hope to move beyond simple repeatability.

At this point, it seems that successful forecasting of experimental evolution will only be possible when there is a dominant selective pressure that can be recognized beforehand. One obvious example is high-level antibiotic resistance. As all sensitive bacteria die and only resistant mutants survive, the prediction on the phenotypic level is trivial and often the mutational target can also be predicted if it is already well known from other strains or species. Under many other experimental conditions, the dominant selective pressure is much more difficult to identify beforehand. Adaptation to new growth media or to a different temperature can be achieved by a large number of phenotypes, which prevents detailed forecasts beyond that of increased fitness and perhaps types of mutations. There might also be arising ecological complexities. Adaptation in a serial transfer experiment can proceed by multiple adaptive pathways that might also lead to several ecological types coexisting over long timescales. This was found by Herron and Doebeli (2013) in a study where *E. coli* populations diverged into one ecotype specializing on growing fast on glucose and another growing rapidly on acetate after glucose was depleted. Unintended differences in selection can also be introduced by the growth dynamics in serial transfer experiments. If one environmental condition allows the bacteria to reach stationary phase faster than the other, the difference in outcome can reflect not only adaptation to the intended condition, but also longer time for mutants with increased growth during stationary phase to increase in frequency.

4.2.4 Complexity and Diversity in Experimental Evolution

As exemplified above, prediction of for example high-level antibiotic resistance to certain antibiotics such as rifampicin can be successful and often predict a handful of likely mutations. However, the prediction would be identical for all species and is based solely on the molecular phenotype of the rifampicin binding site in the RNA polymerase subunit (Goldstein 2014). There is nothing surprising about the success of our forecast, even if it can be extended between species, as there is no known diversity in possible solutions. A useful forecast would need to combine reliance of known information with an ability to make novel predictions in some way and this requires the existence of diverse adaptive outcomes. A related difficulty is how to show that there are alternative phenotypic or genotypic solutions if only one outcome is ever observed (Lind et al. 2015).

Related to the issue of diversity is the level of complexity of the expected adaptive solutions. For rifampicin resistance, the model system is too simple, but media adaptation involves a plenitude of possible adaptive phenotypes and an even greater diversity of mutational paths. In many cases, adaptive mutations occur in global regulators of gene expression, such as RNA polymerase subunits, sigma factors, RNAases,

ribosomal proteins, topoisomerases or DNA-binding proteins that change the expression of hundreds of genes (Good et al. 2017; Tenaillon et al. 2012; Wang et al. 2018). Therefore, it is often difficult to determine the cause of the fitness increase on the molecular or cellular level and thus to define a particular adaptive phenotype. In many cases, mutations in global regulators are also likely to be compensatory mutations restoring gene expression after previous laboratory-induced mutations where a simple reversal or directly related compensatory mutation is unable to restore cellular function as discussed above. This could explain why core genes was found to evolve rapidly in the long-term evolution experiment with *E. coli* (Maddamsetti et al. 2017) in contrast to what is observed in natural populations.

4.2.5 Relevance to Natural Populations

The strength of the simple controlled environments used in experimental evolution studies is also one of its weaknesses. It is debatable how relevant findings will be to natural populations that experience changing environments and ecological interactions and with structured populations that are orders of magnitudes larger (Bailey and Bataillon 2016). What is clear is that the absolute fitness values obtained in laboratory experiments will mean little in natural populations. However, it might be possible to select for ecologically relevant traits that are likely to be correlated with fitness also in natural contexts, which could provide a link between laboratory models and clinical and environmental strains.

In cases where very strong selection for a new function in a particular gene is possible, the potential for a direct link to natural populations is greater. This includes antibiotic resistance studies where the same mutations found after experimental evolution is detected in nature (Hughes and Andersson 2017) although there can also be large differences in the spectrum of mutations found (Thulin et al. 2015). Fitness effects of resistance mutations have also found to be correlated with natural prevalence (O'Neill et al. 2006; Sommer et al. 2017; Thulin et al. 2015).

4.3 Causes of Repeated Experimental Evolution

Forecasts of experimental evolution can range from the binary prediction of whether an adaptation will occur in a defined time to detailed probabilistic models of defined mutational trajectories. In bacterial model systems without standing genetic variation and horizontal gene transfer, the phenotypic variation that natural selection acts upon is produced solely by mutation. This means that in addition to fitness, the outcome will in many cases also depend on the rate of phenotypic production by mutation (Bailey et al. 2017; Lind et al. 2015, 2019; Stern 2013; Yampolsky and Stoltzfus 2001). Origin-fixation models have long been used to describe these dynamics and provide a rational starting point for determining the relative influence of mutation

and selection in experimental evolution studies (McCandlish and Stoltzfus 2014). It is less clear if they are useful for forecasting. This would require us to either measure the fitness of different adaptive mutants and their mutation rates or predict them a priori. In the first case, these parameters would be often inaccessible and when such data could be obtained, it would not be in the spirit of forecasting. In the second case, predictions of either fitness effects of mutations and rates of mutations remain unsolved problems at the heart of biology. Still, the origin-fixation framework points to the underlying causes of repeatability in experimental evolution and might provide a way forward by dividing the forecasting problem into several sub-problems. These processes work at different biological levels and thus no grand theory or general forecasting model is ever expected to be possible.

Consider a simple case of perfectly repeatable experimental evolution (Fig. 4.1). Each time evolution is replayed the same phenotype evolves and a mutation is found in the same gene. Unless the underlying cause of repeatability is understood no prediction can be made outside this specific context as discussed above. The simplest explanation is that only mutations in this particular gene can produce the phenotype and thus repeatability is explained by a lack of alternatives. In a related organism with similar gene content, a reasonable prediction is that the same mutational solution will be used. The more commonly invoked explanation, at least for more complex adaptive scenarios, is that mutations repeatedly occur in the same gene because it

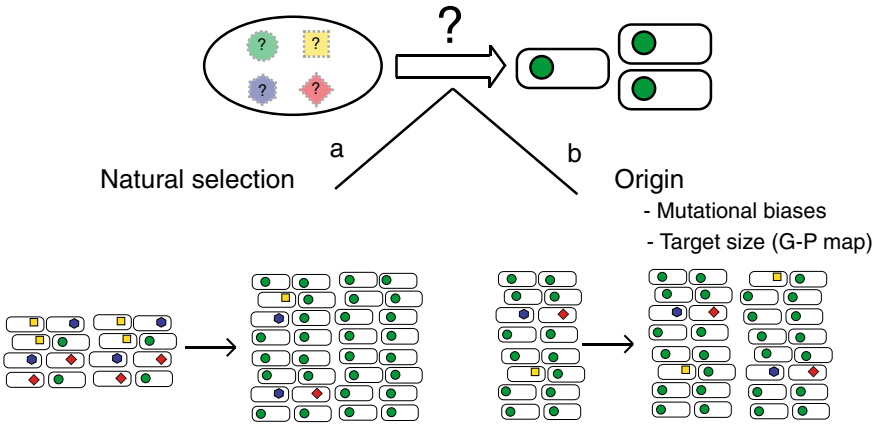


Fig. 4.1 Repeated experimental evolution is caused by the interplay of natural selection and the origin of phenotypic variation by mutation. If there are several adaptive solutions (green, yellow, blue and red), but only one (green) is observed in the evolved population the underlying causes for this is often unclear. **a** Repeated phenotypic and genetic evolution is often interpreted solely in terms of natural selection. Although there are many adaptive mutations, some have higher fitness and as the population grows they outcompete other mutants. **b** There might also be several equally fit adaptive mutants, but they are produced by mutation at different rates so that some phenotypes and genotypes are more frequent. Differences in rates for origin of phenotypic variation can be caused by mutational biases. Alternatively, the genotype-to-phenotype map can bias evolution towards the pathways with the largest target size for adaptive mutations

produces the phenotype with the highest fitness even though there are other adaptive mutations that originate at similar rates (Fig. 4.1a).

An alternative explanation for the repeated evolution is that although mutations in many genes can produce phenotypes of equal adaptive value they occur in a particular gene because it produces phenotypic variation at a higher rate (Fig. 4.1b) (Lind et al. 2015, 2019; Stern 2013; Yampolsky and Stoltzfus 2001). Simply put: they are more frequent. This can be due either to a higher mechanistic mutation rate or a higher capacity to translate genetic variation into phenotypic variation, i.e. target size. These two factors are commonly included together in origin-fixation models, but as will be discussed below the methods for trying to predict them will be fundamentally different. Fitness, mutation rate and mutational target size will in most systems all influence the outcome, but depending on circumstances one may be completely dominant. Each of these factors can also either decrease or increase forecasting potential depending on our ability to understand and predict them. Perfectly deterministic outcomes can be impossible to predict if, for example, it is caused by mutational hot spots that cannot be detected beforehand.

4.4 Potential for Forecasting on Different Biological Levels

Forecasts of experimental evolution can be made at different biological levels (Fig. 4.2) and it is generally assumed that higher-order traits such as fitness will be more easily predicted than specific phenotypes (Lässig et al. 2017). On the genetic levels, a similar assumption is that mutational targets will be increasingly difficult to predict starting with gene networks, operons, genes and down to specific mutations (Lässig et al. 2017). In order to make forecasts on different levels, mapping between genotype and molecular systems in the cell, molecular systems to phenotypes and phenotypes to fitness must be accomplished (Fig. 4.2). The universal conclusion from decades of experimental evolution studies is that fitness increases over time and that it continues to increase for many generations (Good et al. 2017). Depending on the experimental system, the rate of increase and the distribution of fitness effects of beneficial mutations appears to vary considerably (Good et al. 2012; Kassen and Bataillon 2006; MacLean and Buckling 2009; McDonald et al. 2011; Orr 2003; Perfeito et al. 2007; Rokyta et al. 2008). Nonetheless, the rational starting point for making an a priori evolutionary forecast of a new experimental evolution study is fitness.

4.4.1 Prediction of Fitness

In experimental evolution studies of adaptation with bacteria, the majority of mutations rising to high frequencies in the population will increase fitness. Large population sizes eliminate genetic drift and hitchhiking neutral or slightly deleterious

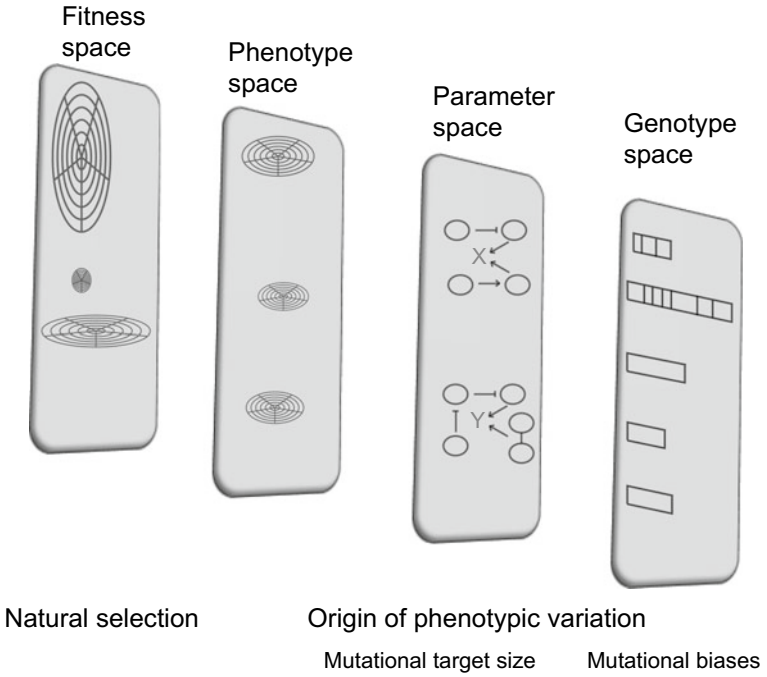


Fig. 4.2 Evolutionary forecasts can be made at different levels, from mutations in specific positions to an increase in fitness. Forecasting on several levels require the ability to connect fitness to specific phenotypes, phenotypes to specific cellular/molecular states in parameter space or connecting genotypic variation to changes in parameter space. The causes of repeated experimental evolution (Fig. 4.1) operate at different levels with natural selection in fitness space, mutational target size in the translation of genotypic variation into phenotypic variation, i.e. parameter space, and mutational biases at the genotype level. Adapted from Lind (2018)

mutations are likely to play a minor part in future adaptation. Much progress has been made in examining the distribution of fitness effects of beneficial mutations both experimentally and through sequence analysis (Eyre-Walker and Keightley 2007; Good et al. 2012; Kassen and Bataillon 2006; Keightley and Eyre-Walker 2010; MacLean and Buckling 2009; McDonald et al. 2011; Orr 2003; Perfeito et al. 2007; Rokyta et al. 2008). The general conclusion is that the large majority of beneficial mutations increase fitness only slightly, with a tail of rare high-fitness mutants, and it has been suggested that the shape is best described by an exponential distribution (Orr 2010) (Fig. 4.3a). But as mutants with higher fitness are more likely to become fixed in the population, the fitness effects of the observed adaptive mutants is much larger than average and adaptation will be dominated by large-effect mutations in most bacterial experimental evolution studies where population sizes are large (Good et al. 2012).

When experimental evolution continues for long enough to allow fixation of several consecutive mutations [or increase in frequency of several mutations as in clonal

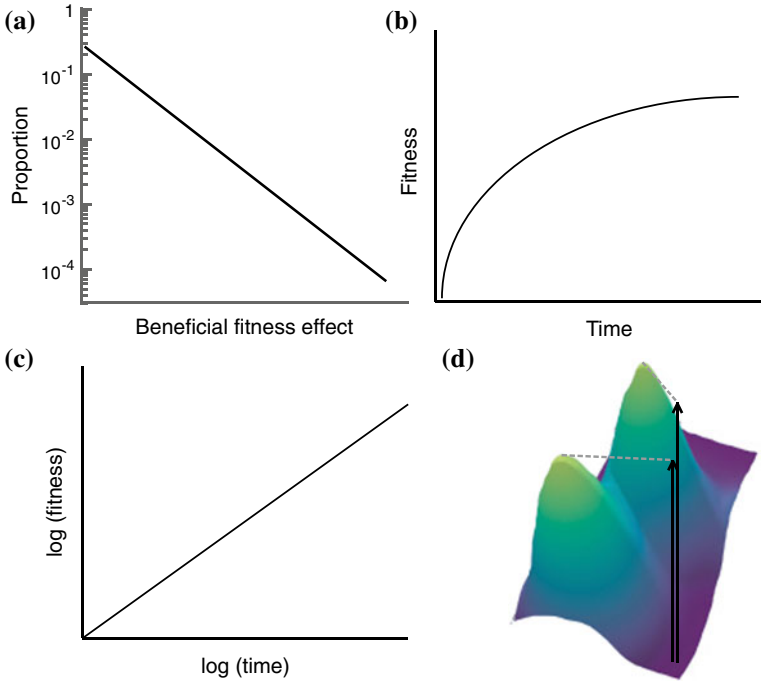


Fig. 4.3 **a** The distribution of fitness effects of beneficial mutations is often assumed to be exponential, based on extreme value theory (Orr 2003). **b** Fitness in experimental evolution increases most rapidly early on and typically continues to increase until the experiment is ended (Good et al. 2017). **c** The increase in fitness over time can been described by a power law where log (fitness) increases linearly with log (time) (Wiser et al. 2013). **d** Even in cases where distinct phenotypic adaptive solutions can be predicted, there is generally no way to predict which solution will be superior and when experimental data are available, it is often unclear to what degree fitness depends on environment and genetic background

interference (Gerrish and Lenski 1998)] the rate of fitness increase can be analysed. For populations adapting to constant environments, such as a new growth medium or increased temperature, the rate is rapid in the beginning and the rate of fitness increase then declines but fitness typically continues to increase until the end of the experiment (Fig. 4.3b). This increase in fitness can be described by a power law where log fitness increases linearly in log time (Wiser et al. 2013) (Fig. 4.3c). Even in cases when there is little parallelism on the genetic level convergence on the level of fitness is common (Blount et al. 2018). Unfortunately, there is currently no possibility to connect forecasts of fitness effects for growth media or temperature adaptation to that of particular phenotypes or particular genes, although mutations in global regulators are common (Fig. 4.2).

Many types of experimental evolution set-ups select for a more specific phenotype that is strongly correlated to fitness in the particular environment, which may limit their generalizability (Van den Bergh et al. 2018). However, by reducing the size of

the relevant phenotype space, they may provide more potential for forecasting not only of fitness effects, but also the mutational targets. Examples include antibiotic resistance, biofilm formation, swarming and evolution of new metabolic activities (Blank et al. 2014; Gullberg et al. 2011; Kovacs and Dragos 2019; Nasvall et al. 2012; Rainey and Travisano 1998; Steenackers et al. 2016; Van den Bergh et al. 2018; van Ditmarsch et al. 2013; Wong et al. 2012). Directly comparing the distributions of beneficial fitness effects for these model systems are difficult due to the difference in fitness assays used. Nevertheless, it is apparent that a limited number of possible phenotypic and genetic solutions to a particular problem mean that no particular distribution is likely to be found. For example, if two different phenotypes can solve an adaptive problem, they will each have a distinct distribution of fitness effects and the mode of the distribution will be completely dependent on the selective challenge. Or using a fitness landscape analogy: if there are only two peaks in the landscape we have no way of telling a priori how high they will be (Fig. 4.3d).

4.4.2 Prediction of Phenotypes

The complexity of adaptation to new growth media or temperature and the commonly observed large-scale changes in gene expression brought about by mutations in global regulators means that there is currently little hope to connect fitness to specific phenotypes or molecular consequences (Fig. 4.2). Therefore, forecasts will often be stuck on the fitness level. Of course, this does not reduce the usefulness of such studies in general, but they are less well suited for testing evolutionary forecasts. Selecting for solutions to more specific adaptive problems comes at a cost of lost generalizability, but also affords the possibility to predict which specific molecular or cellular phenotypes will increase fitness. This opens up possibilities to also predict genetic targets in cases where the genotype-to-phenotype map is at least partly understood or can be predicted in some way.

A priori prediction of adaptive phenotypes relies on previous experimental data and knowledge of cellular and molecular biology as well as the ability to select for a dominant phenotype. For antibiotic resistance, several different general mechanisms are known, including target modification, enzymatic degradation and modification, reduced uptake and increased efflux. It is thus conceivable that is might become possible to predict not only the existence of resistance mechanisms to a novel antibiotic, but also the mutation rate and in which genes mutations would occur and also the molecular changes required. Previous knowledge of fitness effects of mutations in, for example, porins, limiting uptake of an antibiotic, or mutations that activate efflux pumps, may then be used to provide a risk assessment for future resistance that would also need to be complemented by an assessment of resistance by horizontal gene transfer using other methods. In similar ways, specific phenotypes, such as biofilm formation or increased motility, can be directly linked to their structural components, in these cases production of exopolysaccharides or flagella, but also

to the regulatory genes controlling these processes. If the underlying genetic architecture is well understood, this opens the door for parameterization and modelling (Lind et al. 2019; Savageau and Fasani 2009; Valderrama-Gomez et al. 2018).

4.4.3 Origin of Adaptive Mutants

In experimental evolution studies that select for specific phenotypes, several possible molecular solutions are likely to be possible. Although it is often assumed that the solutions commonly found corresponds to the highest fitness phenotypes, it is increasingly recognized that the rate of production of phenotypic variation, the origin process, can also greatly influence the outcome (Lind et al. 2015; McCandlish and Stoltzfus 2014; Yampolsky and Stoltzfus 2001). If we assume that mutations in two genes can produce equally fit phenotypes in the experimental environment, yet only mutations in one gene are ever observed this could have two explanations that are not mutually exclusive. Either there are more mutation sites that can produce the phenotype in one of the genes or the mechanistic mutation rate is higher for one gene (Fig. 4.1b).

4.4.3.1 Mutational Biases

While much effort has been made to elucidate the distribution of fitness effects of new mutations, the importance of the distribution of mutation rates has been less explored (Harpak et al. 2016; Hodgkinson et al. 2009; Johnson and Hellmann 2011). Average mutation rates are available from experimental estimates or sequence analysis for a large number of organisms (Lynch et al. 2016), but just as knowledge of average fitness effects of mutations would help little with predicting the effects of fixed mutations, there is a possibility that a wide distribution of mutation rates could have a major impact on forecasting. If mutation rates at different sites in a genome vary over several orders of magnitude in ways that cannot be predicted, this could completely dominate the outcomes in experimental evolution studies and thwart our attempts to forecast evolution. It is also not known how conserved mutational biases are between strains and species at nucleotide, genic and genomic levels.

There are many ways in which mutation rates vary. The simplest is that different types of mutations have different rates. For example, transitions are more common than transversions in many organisms, which could bias adaptation (Stoltzfus and McCandlish 2017), but given a similar distribution of nucleotides between different genes this is unlikely to bias evolution heavily towards a particular gene or region. The rate of gene duplications has been shown to be very high, up to 10^{-3} per cell per generation in *S. typhimurium* LT2 (Reams et al. 2014), and there appears to be a range of several orders of magnitude for duplication rates at different genomic regions, often with specific recombinational hot spots with long sequence homologies making a significant contribution (Reams and Roth 2015). Once a duplication has occurred,

higher copy number amplifications are produced at even higher rates, given the newly formed long identical sequence homologies, but these are also inherently unstable and are typically lost at high rates when not under selection (Andersson and Hughes 2009; Reams and Roth 2015). Deletions and inversions between long homologies also occur at high rates while deletions that are formed by non-homologous end-joining are orders of magnitude less common (Koskiniemi and Andersson 2009; Lovett et al. 2002). Mobile genetic elements can produce not only insertional inactivation at high rates (Sousa et al. 2013), but can also contribute to deletions and duplications at very high rates when they are present in several copies given their long identical sequence homologies provide substrates for recombination. For example, deletion of *rbs* in *E. coli* occur at high rates (~5×10^{-5} per cell generation) due to a IS150 element upstream of the operon that provides sequence homology for recombinational deletion when a new copy is introduced into the *rbs* operon (Cooper et al. 2001). Given the mobility of these types of sequences their location, copy number and transposition activity differs to a large degree even between closely related strains, which could introduce strong strain-specific variations in mutation rates and therefore also in the rates of production of phenotypic variation and the paths taken by evolution.

Mutational hot spots are also caused by replication slippage leading high mutation rates at homonucleotide tracts, quasipalindromes or deletions between short sequence homologies that are in close proximity (Levinson and Gutman 1987; Lovett 2004; Viswanathan et al. 2000). In addition to these mutational processes that are linked to the local properties of DNA, the molecular processes of the cells could make a significant contribution to mutational biases. Examples of this include transcription–replication collisions that were found to elevate mutations by orders of magnitude at some positions, strand-specific biases in transcription and replication and the possible influence of DNA-binding proteins (Hudson et al. 2003; Lind and Andersson 2008; Sankar et al. 2016).

The spectra of mutations and rates across genomes are also heavily influenced by environmental conditions (Krasovec et al. 2017; Maharjan and Ferenci 2017; Shewaramani et al. 2017) and is expected to vary between different species depending on the complement of DNA repair proteins (Lind and Andersson 2008). Given all these potential biases in mutation rates it is not surprising that no computational approaches to predicting mutation rates for specific nucleotides, genes of genomic regions have been described even though it is acknowledged that biases can contribute to adaptation and the distribution of allele differences observed in a population (Harpak et al. 2016; Stoltzfus and McCandlish 2017). High-throughput sequencing in combination with machine learning is likely to partially fill this gap, but it is not clear to what degree such models will be transferable between different environments or different species.

4.4.3.2 Mutational Target Size

Why would some genes have a higher capacity to translate genotypic variation into phenotypic variation? One simple explanation is that loss-of-function mutations in a

gene are generally assumed to be much more common than gain-of-function mutations. This explains why loss-of-function mutations dominate in many experimental evolution studies (Long et al. 2015). Inactivating mutations does not necessarily mean simply losing something that is not useful in a particular environment, such as flagella or virulence traits, but they can also lead to changes in gene expression and thereby activation of other genes. If an adaptive phenotype can be reached by complete loss-of-function of a gene product or increase in function of another, the expectation would be that the former would have a much larger mutational target than the latter simply on account of loss-of-function mutations being more common than gain-of-function mutations. Gene products have a variety of different functions, including catalysis, transport, structural roles and signalling, but they are all regulated in some way, either through local regulation, for example by a transcription factor, but also through global regulatory systems or a general connection with the cellular physiology or growth rate. The genetic architecture underlying adaptive traits are often complex, requiring the concerted action of many gene products. In such cases, increase or decrease in function of a single gene product is unlikely to have a significant effect on fitness. A simple example would be a bacterial operon where several genes are transcribed as a single unit that encodes proteins needed for growth on a particular carbon source. Only an increase in expression of the entire operon would suffice and increasing the activity or amount of a single gene product will only matter if it is limiting for growth. Molecular networks are however usually much more complex than single operons and shifting the output often requires mutations in key regulator genes. This probably accounts for the high proportion of adaptive mutations found in global regulatory genes such as RNA polymerase subunits, sigma factors and stringent response signalling in experimental evolution studies (Blount et al. 2018; Good et al. 2017; Long et al. 2015). Unfortunately, understanding the molecular/cellular cause of the fitness increase in such cases remains challenging due to its complexity, i.e. mapping of large changes in parameter space to that of specific phenotypes and fitness (Fig. 4.2).

With an ever-increasing number of experimental and bioinformatics studies, there is a growing potential for mechanistic modelling of genotype-to-phenotype maps. Knowledge of the interaction between gene products and how they produce phenotypic variation allow parameterization and could provide methods for contributing to evolutionary forecasting (Lind et al. 2019; Rainey et al. 2017). The *Pseudomonas fluorescens* SBW25 wrinkly spreader model system represents a well understood (Lind et al. 2015, 2019; McDonald et al. 2009; Rainey and Travisano 1998; Spiers et al. 2002) and intermediate complexity model system that provides an opportunity to explore the relative influences of selection, mutational biases and target size (Lind et al. 2019). Static incubation of the wild type *P. fluorescens* SBW25 selects for mutants that colonize the air–liquid interface in order to gain access to oxygen that is rapidly depleted from the growth medium (Fig. 4.4a). Although other adaptive mutants are possible, the ecotypes with highest fitness all colonize the air–liquid interface by increased cell-cell adhesion and/or adhesion to the tube wall and have distinct colony morphologies on agar plates (Ferguson et al. 2013; Lind et al. 2017b).

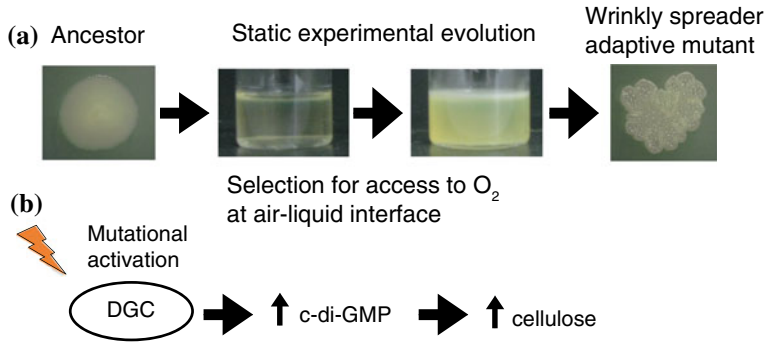


Fig. 4.4 *Pseudomonas fluorescens* SBW25 wrinkly spreader system. **a** The wild type strain has smooth colony morphology on agar plate and when inoculated into static microcosms there is strong selection for colonization of the air–liquid interface where the oxygen concentration is high. Adaptive mutants have increased cell-to-cell or cell-to-wall contact mediated by expression of exopolysaccharides, modification of lipopolysaccharides or cell chaining caused by incomplete cell division. The most successful of these types is the wrinkly spreader that overproduces cellulose and appears wrinkly on agar. **b** The wrinkly spreader phenotype evolves by multiple pathways, all involving mutational activation of diguanylate cyclases that overproduce the second messenger c-di-GMP, which is a conserved signal for biofilm formation, resulting in increased cellulose production. Phenotypically similar mutants, sometimes with mutations in the same genes have been found in clinical isolates of the important pathogen *Pseudomonas aeruginosa* suggesting these types are relevant in natural populations (Malone 2015). Adapted from Lind (2018)

The highest fitness phenotype is the wrinkly spreader that overproduces a cellulosic polymer (Spiers et al. 2003) that functions as glue and that gives it a wrinkly appearance on agar plate. Mutations causing the phenotype have been found to predominantly occur in three loci (McDonald et al. 2009), all encoding a diguanylate cyclase that produce the second messenger c-di-GMP, which is a conserved signal for biofilm formation. Mutations cause activation of c-di-GMP production resulting in overproduction of cellulose (Fig. 4.4b). The majority of mutations are believed to be disabling mutations that disrupt negative regulation by loss of an intermolecular interaction (Lind et al. 2019; McDonald et al. 2009).

Detailed knowledge of the genotype-to-phenotype map underpinning the WS phenotype allowed development of mathematical models (Fig. 4.5a) that can make predictions about the relative rates that the three pathways with differently wired regulation produce wrinkly spreaders (Fig. 4.5b) as well as the genetic targets within each network (Lind et al. 2019). Models were then compared with experimental data (Fig. 4.5c), where mutation was decoupled from selection using a reporter construct, to show that genetic targets could be predicted, including previously unknown mutations (Lind et al. 2019). However, predicted relative mutation rates to the WS phenotype for each pathway did not agree with models due to a mutational hot spot in the *awsX* gene, again demonstrating the importance of mutational biases (Lind et al. 2019). This information was included in a revised model leading to improved predictions (Fig. 4.5d).

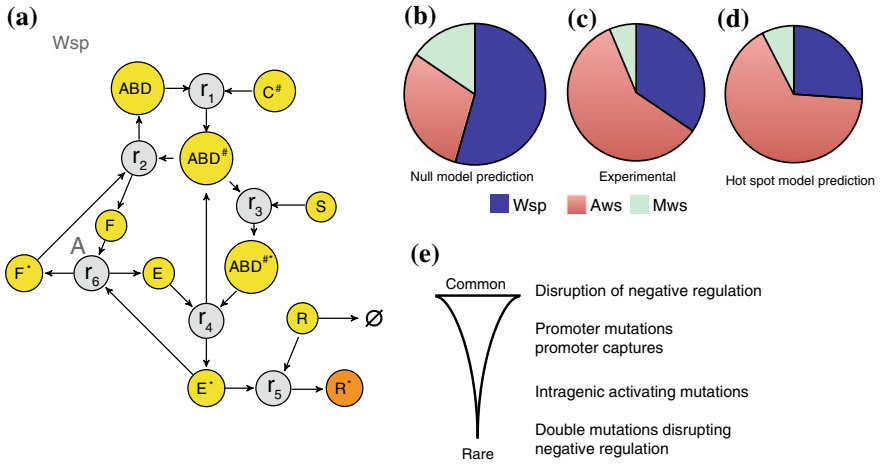


Fig. 4.5 Genotype-to-phenotype map biases evolution of wrinkly spreaders because different diguanylate cyclases have different regulatory networks. **a** Modelling of the reaction networks for the three main pathways to wrinkly spreaders has been described (here only the Wsp network is shown) (Lind et al. 2019). Yellow circles represent proteins (with different modifications for example * is a phosphorylation) and grey circles represent reaction rates. Arrows show the direction in terms of products and reactants. The orange circle R* represent the active form of the diguanylate cyclase and increase of this component after mutational changes to reaction rates was assumed to lead to a wrinkly spreader. For example, mutations that reduce r_6 or r_2 or increase r_5 can all produce a wrinkly spreader. WspF (represented by F and its phosphorylated from F*) can produce wrinkly spreaders by changes in r_2 (reduced demethylation of WspA by loss of enzymatic activity or loss of interaction with WspABD) and r_6 (loss of phosphorylation site or loss of interaction with WspE) (see Fig. 4.6a). **b** Assuming that disabling reactions are more common than enabling reactions and modelling of each network (Wsp, Aws, Mws) produce a null prediction for the relative rates that each pathway will be used. **c** Experimental data on mutation rates to wrinkly spreaders obtained without the biasing influence of natural selection by use of a reporter construct. **d** When a mutational hot spot in one component of the Aws pathway was discovered this information was incorporated into models, which increased predictive ability. **e** In addition to the three main pathways that account for >98% of wrinkly spreaders there are at least 13 additional pathways (Lind et al. 2015). These are used at different rates depending on the types of mutations required to activate diguanylate cyclase activity

There are many examples where regulatory mutations contribute to adaptation in experimental evolution studies focusing on specific phenotype. For example hyperswarming in *Pseudomonas aeruginosa* was repeatedly caused by mutations in the flagellar regulator gene *fleN*, which locks the bacterium in a multiflagellated state (van Ditmarsch et al. 2013), and adaptation to ciprofloxacin by repeated inactivating mutations in the transcription factor *nfxB* that results in overactivation of the efflux pump MexCD-OprJ (Wong et al. 2012).

Disruption of negative regulation provides a likely explanation for why evolution repeatedly uses the three main pathways in the wrinkly spreader system due to the large target sizes. To explore if there are alternative pathways, the three main pathways were deleted and experimental evolution repeated, which allowed identification of

13 alternative pathways to the wrinkly spreader phenotype of approximately equal fitness (Lind et al. 2015). The most common alternative pathway was also subject to negative regulation, but with a small mutation target and resulting in a lower fitness wrinkly spreader. Less frequent among these rare pathways were promoter mutations or promoter captures upstream of diguanylate cyclases, while intragenic activating mutations in diguanylate cyclases and double mutations in negative regulators were even less common. These results lead to a suggested hierarchy of genetic routes based on mutational targets size in cases where a phenotype can be evolved by gene activation (Lind et al. 2015) (Fig. 4.5e). Promoter mutations and promoter captures have been found in other experimental evolution studies as well, including the activation of a citrate transporter in the Lenski's long-term evolution experiment (Blank et al. 2014; Blount et al. 2012; Lamrabet et al. 2019; Lind et al. 2010b). Increased transcription could also be achieved by mutations that create de novo promoters, disruption of upstream transcriptional terminators or promoter gain by insertion of transposable elements with outgoing promoters (Yona et al. 2018). While gene amplification have not observed to be a viable mutational route to wrinkly spreaders, it is expected to be a more common solution when the initial duplication confers a significant benefit (Blank et al. 2014; Lang and Desai 2014; Lind et al. 2010b; Nasvall et al. 2012; Sun et al. 2009).

4.4.4 Prediction of Mutational Effects

The most detailed level of genetic prediction is that of specific mutations. At this level, experimental evolution appears to be less repeatable, as expected, but there are exceptions presumably caused by mutational hot spots or very small mutational targets, i.e. strong functional constraints (Ferguson et al. 2013; Lind et al. 2017b, 2019; Tenaillon et al. 2012). The modelling of genotype-to-phenotype maps provide a way to predict targets of mutations on the level of genes and also in the functional effects of the gene products required to produce the phenotype (Lind et al. 2019). By focusing on a set of reaction networks, such models can predict the functional change required, which can be either an increase or a decrease of a particular reaction (Fig. 4.5a). Moving from reaction rates to predicted mutations thus requires understanding molecular function and structure and how mutations change the function of individual gene products.

Prediction of mutational effects by computational methods could provide a way forward. These are based on a range of different parameters including modelling of thermodynamic effects of mutations, methods based on evolutionary conservation and hybrid methods incorporating mixed data using machine learning (Bromberg and Rost 2007; Capriotti et al. 2013; Celniker et al. 2013; Choi and Chan 2015; Dehouck et al. 2011; Ng and Henikoff 2003). While thermodynamic stability is most certainly an important selective constraint over longer time scales, many proteins appear to tolerate single mutations without major structural changes (Bloom et al. 2006) arguing against the suitability of these methods in predicting mutational effects.

Experimental measures of thermodynamic stability are dominated by relatively small proteins, often enzymes, without interactions with other molecules than its small molecule substrate (Kumar et al. 2006), which could mean that results are less useful for proteins complexes or proteins with multiple dynamic interactions. It has been suggested that rather than being destabilizing, most mutations with large effects on fitness is likely to disrupt intermolecular of interdomain interactions (Lind et al. 2017a). In the case of monomeric enzymes, these mutations will be close to the active site (Lundin et al. 2017) and mutational changes in the core of the protein will typically have large effects also on thermodynamic stability, possibly explaining the sometimes observed correlation between folding stability and mutational effects.

For other proteins, the situation becomes more complex. For the transcriptional activator AraC, different parts of the protein function in DNA binding, dimerization, interaction with RNA polymerase and binding of arabinose (Lind et al. 2017a) (Fig. 4.6a). For the negative regulator WspF, a common target for wrinkly spreader mutations in SBW25 described above, predicted interactions include both proteins WspA and WspE as well as a phosphorylation site and an enzymatic methylesterase site (Lind 2018) (Fig. 4.6b). In these cases large-effect mutations are not found primarily in the core of the proteins but rather concentrated around the functional sites where they are expected to disrupt interactions, including surface exposed regions (Fig. 4.6a, b). Prediction of the thermodynamic effects of mutations on intermolecular interactions is possible when interactions have been experimentally determined, but it requires separation of the effects on overall folding and the stability of the binding interface (Otwinowski 2018). However, in many cases transient and dynamic interactions are not known. Prediction of the size of interaction surfaces between proteins could possibly be used to estimate the target size for mutational disruption without the need for explicit models of free energy change. For example, mutations in the putative diguanylate cyclase PFLU0085 can lead to increased c-di-GMP production and a wrinkly spreader phenotype through mutations clustered in a particular region, including small deletions and amino acid substitutions, suggesting that any disrupting mutation in the region will do (Lind et al. 2015). Assuming that most adaptive mutations disrupt intermolecular interactions together with the observation that no mutations were found in a potential unknown interacting protein, suggests that the mutations disrupt interdomain interactions or oligomerization. In order to predict mutational targets in related species homology modelling of the structure of the monomer (Kelley et al. 2015) combined with mapping of mutations onto structure can then be used to select a likely dimeric model (Baek et al. 2017) where the interaction surface between the monomers can be estimated and likely mutations predicted (Fig. 4.6c).

Methods based on evolutionary conservation, including hybrid machine learning methods, implicitly incorporate interaction information and has been found to better predict fitness effects (Lind et al. 2017a) (Fig. 4.6d), but experimental fitness data for mutations in more complex proteins remain scarce. A cautionary example is that mutations in ribosomal proteins are predicted to be highly deleterious, but the experimental data shows them to be extraordinarily robust to mutations, presumable due to their large interaction surface with ribosomal RNA (Lind et al. 2010a, 2017a).

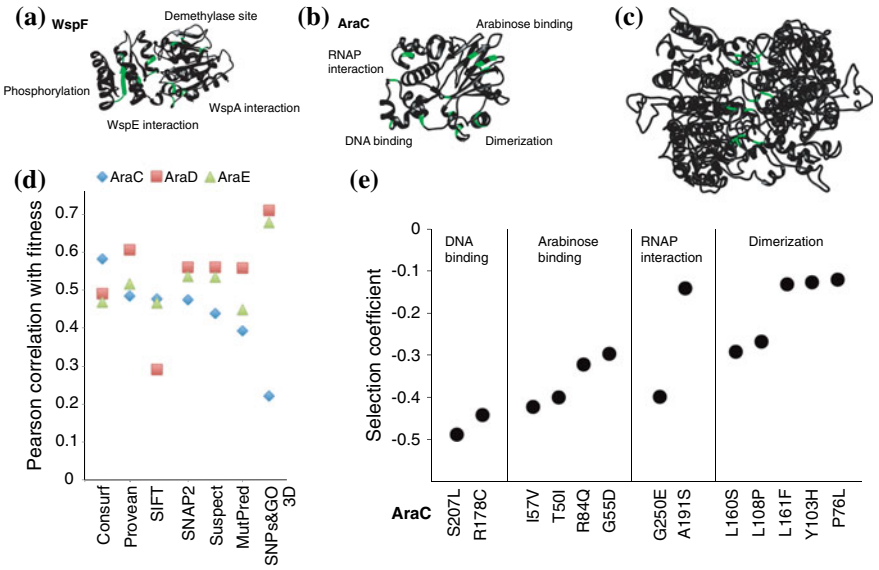


Fig. 4.6 Causes and prediction of mutational effects. **a** Loss-of-function mutations in WspF cause a wrinkly spreader phenotype. In addition to insertion and deletions, amino acid substitutions (in green) in four regions of the protein have been reported with each corresponding to disruption of a particular interaction. Adapted from Lind (2018). **b** Highly deleterious mutations the transcriptional activator AraC (in green) are clustered in four different regions each potentially disrupting a particular intermolecular interaction. **c** Disabling mutations in a small region of PFLU0085 are found in a surface exposed region away from diguanylate cyclase domain. By mapping mutations (green) onto a homology model of the same protein in *P. protegens* Pf-5 PFL_0087, a likely dimer structure can be produced, based on the assumption that mutations disrupt intermolecular interactions, and the size of the interaction surface and mutational target can be estimated. **d** Methods based on evolutionary conservation can produce significant predictions of mutational effects. Adapted from Lind et al. (2017a). **e** Fitness effects of disabling mutations in four distinct regions differ, but are often similar for mutations with similar function effect

When an adaptive phenotype can be realized by many different mutations in several different genes, it would be necessary to in some way predict the fitness of all these variants. Currently, there is no way to do this and experimental measurements of fitness for large sets of mutations are difficult even for microbes. A way forward could be to assume that mutations with similar molecular effects will have similar fitness effects. This in combination with the assumption that most large-effect mutations cause a complete loss of a functional interaction suggests that the distribution of fitness effects for each gene has a limited number of modes corresponding mainly to loss of each of its functions. This provides an explanation for the bimodal distribution of fitness effects of random mutations commonly observed with one mode close to neutrality and one close to deletion fitness (Eyre-Walker and Keightley 2007; Hietpas et al. 2011; Lind et al. 2017a; Lundin et al. 2017; Sanjuan 2010). For proteins with more complex interactions, it is conceivable that the fitness modes would correspond to loss of each functional interaction, so that for AraC with four distinct functional

regions, fitness effects of large-effect mutations are expected to be similar for mutations in the same region (Fig. 4.6e). If we can assume that experimental evolution will be dominated by large-effect mutations and each protein has only a few different functional phenotypes that can be reached by one or possibly a few mutations this could simplify forecasts. For a bacterium with a genome of 5 Mbp and 5000 genes, assuming only a few fitness modes per gene would dramatically reduce the dimensionality of the problem compared to considering all possible mutations.

Most likely, enabling mutations also play a major part in some contexts even though their rate is believed to be lower than disabling mutations. These could strengthen an interaction, for example with a new substrate or increase binding to a specific DNA motif. In many cases, however, an increase of one reaction rate can be the result of a decrease of another: disruption of one interaction will allow formation of another or relieve interdomain feedback control.

4.4.5 Conclusion—Desired Properties of Model Systems for Forecasting Experimental Evolution

A long-term goal for evolutionary forecasting is to be able to predict short-term evolutionary processes in natural environments based only on an organism's or cell's genome sequence combined with previous experimental data, bioinformatics and mathematical models. Machine learning approaches and fitness models are likely to dominate well-described scenarios of special interests such as cancer or infectious diseases, but there is also potential for mechanistic models to contribute, especially to forecasting of novel scenarios.

Experimental evolution could prove useful for development of new methods and to explicitly test forecasting ability under simplified conditions. A diverse set of model systems is required to reduce the impact of idiosyncrasies from specific strain backgrounds or specific selective conditions. There are some properties that should be taken into account when searching for suitable model systems for testing of evolutionary forecasting methods and no model is likely to satisfy them all. A key requirement is robustness, both in term of genetics (similar strains should evolve similar adaptive solutions) and environment (similar environments should give similar adaptive solutions because a dominant selective pressure can be established). This is necessary to be able to transfer forecasting models between related biological systems. The dominant selective pressure imposed should ideally be relevant to natural populations.

The most informative model systems are currently likely to be of intermediate complexity so that there are several phenotypic and genetic solutions, but these can be recognized beforehand and their cellular and molecular effects can be predicted to some degree. There should also be diversity between different strains or species, so that forecasts are not all identical, but a degree of conservation is also required to be able to use data from previous experiments to improve forecasts in an iterative

manner. Ideally, the adaptive solutions would be modular and non-essential so that a genetic of phenotypic solution can be genetically removed to allow finding rare or low fitness pathways. The desired properties of diversity, conservation and modularity might most readily be found for genes subject to high rates of horizontal gene transfer and gene loss.

The genotype-to-phenotype map should be understood well enough to allow parameterization and modelling to allow formal assessment of forecasting ability. Tools for genetic engineering must be available to reconstruct mutations and to create predicted mutants that are not observed after experimental evolution to elucidate the underlying causes. Fitness assays should be feasible under the experimental conditions and fitness of mutants should not be frequency dependent. Ideally, the mutation rate for each phenotype should be able to be measured without the influence of natural selection using fluctuation tests.

Clearly, it will be difficult to find one experimental system that ticks all the boxes. For the *P. fluorescens* wrinkly spreader system, more than one hundred mutations leading to distinct adaptive phenotypes have been described (Ferguson et al. 2013; Lind et al. 2017b, 2019; McDonald et al. 2009, 2011), including rare mutational pathways (Lind et al. 2015), and it is possible to obtain mutants from fluctuation assays that can be directly compared to predictions from models (Lind et al. 2019). However, disadvantages with the system include short timescale and the difficulty of frequency-dependent fitness effects (Lind et al. 2015, 2019). The wrinkly spreader system also has several properties that might allow it to be extended to other species of *Pseudomonas* (Lind 2018). A dominant selective pressure can be established by selection for colonization of the air–liquid interface that is independent of growth media, temperature, population size and growth vessel (Lind 2018). Phenotypic solutions are likely to be limited, but diverse (Lind 2018), and the conserved role of the c-di-GMP network (Romling et al. 2013) makes its activation a probable scenario across species. Both diguanylate cyclases and exopolysaccharides are commonly transferred by horizontal gene transfer and lost by deletions leading to different predictions even for closely related species, while similar molecular networks can be present in divergent species and models reused. This allows forecasts to be made based only on genome sequence and data from the original wrinkly spreader system on the levels of ecotype, phenotype, types of mutations, genetic networks, single genes and specific mutations (Lind 2018). No doubt many such predictions will fail, but hopefully they will fail in interesting and informative ways. When reasons for failure can be determined, this allows improvement of forecasting methods and the use of an iterative workflow of forecasting: experimental evolution—genome sequencing—characterization—model update. An initial experimental test with the closely related species *Pseudomonas protegens* Pf-5 suggests that forecasts can be extended to other species and that conservation of the genotype-to-phenotype map cause similar evolutionary outcomes (Lind 2018). The results also show that mutational biases were not conserved and it remains to be determined to what degree relative fitness effects of mutations are conserved between species (Lind 2018).

Understanding the fundamental limits of evolutionary forecasting and developing effective methodologies requires that not only positive data be published. Ideally,

forecasts should be published before conducting an experimental evolution study, which could also allow other researchers to develop alternative forecasts. Often these forecasts would be wrong, and nobody likes being wrong, but it may propel the evolutionary forecasting field forward faster than if only studies with impressive ad hoc explanations of repeatability are published.

References

Andersson DI, Hughes D (2009) Gene amplification and adaptive evolution in bacteria. Annu Rev Genet 43:167–195. https://doi.org/10.1146/annurev-genet-102108-134805

Baek M, Park T, Heo L, Park C, Seok C (2017) GalaxyHomomer: a web server for protein homo-oligomer structure prediction from a monomer sequence or structure. Nucleic Acids Res 45:W320–W324. https://doi.org/10.1093/nar/gkx246

Bailey SF, Bataillon T (2016) Can the experimental evolution programme help us elucidate the genetic basis of adaptation in nature? Mol Ecol 25:203–218. https://doi.org/10.1111/mec.13378

Bailey SF, Blanquart F, Bataillon T, Kassen R (2017) What drives parallel evolution?: how population size and mutational variation contribute to repeated evolution. BioEssays 39:1–9. https://doi.org/10.1002/bies.201600176

Barrick JE, Lenski RE (2013) Genome dynamics during experimental evolution. Nat Rev Genet 14:827–839. https://doi.org/10.1038/nrg3564

Bennett DE et al (2009) Drug resistance mutations for surveillance of transmitted HIV-1 drug-resistance: 2009 update. PLoS One 4:e4724. https://doi.org/10.1371/journal.pone.0004724

Blank D, Wolf L, Ackermann M, Silander OK (2014) The predictability of molecular evolution during functional innovation. Proc Natl Acad Sci USA 111:3044–3049. https://doi.org/10.1073/pnas.1318797111

Blattner FR et al (1997) The complete genome sequence of *Escherichia coli* K-12. Science 277:1453–1462

Bloom JD, Labthavikul ST, Otey CR, Arnold FH (2006) Protein stability promotes evolvability. Proc Natl Acad Sci USA 103:5869–5874. https://doi.org/10.1073/pnas.0510098103

Blount ZD, Barrick JE, Davidson CJ, Lenski RE (2012) Genomic analysis of a key innovation in an experimental *Escherichia coli* population. Nature 489:513–518. https://doi.org/10.1038/nature11514

Blount ZD, Lenski RE, Losos JB (2018) Contingency and determinism in evolution: replaying life's tape. Science 362. https://doi.org/10.1126/science.aam5979

Bromberg Y, Rost B (2007) SNAP: predict effect of non-synonymous polymorphisms on function. Nucleic Acids Res 35:3823–3835. https://doi.org/10.1093/nar/gkm238

Capriotti E, Calabrese R, Fariselli P, Martelli PL, Altman RB, Casadio R (2013) WS-SNPs&GO: a web server for predicting the deleterious effect of human protein variants using functional annotation. BMC Genomics 14(Suppl 3):S6. https://doi.org/10.1186/1471-2164-14-s3-s6

Celniker G et al (2013) ConSurf: using evolutionary data to raise testable hypotheses about protein function. Isr J Chem 53:199–206. https://doi.org/10.1002/ijch.201200096

Choi Y, Chan AP (2015) PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels. Bioinformatics 31:2745–2747. https://doi.org/10.1093/bioinformatics/btv195

Cooper VS, Schneider D, Blot M, Lenski RE (2001) Mechanisms causing rapid and parallel losses of ribose catabolism in evolving populations of *Escherichia coli* B. J Bacteriol 183:2834–2841. https://doi.org/10.1128/JB.183.9.2834-2841.2001

Daegelen P, Studier FW, Lenski RE, Cure S, Kim JF (2009) Tracing ancestors and relatives of *Escherichia coli* B, and the derivation of B strains REL606 and BL21(DE3). J Mol Biol 394:634–643. https://doi.org/10.1016/j.jmb.2009.09.022

de Visser JA, Krug J (2014) Empirical fitness landscapes and the predictability of evolution. Nat Rev Genet 15:480–490. https://doi.org/10.1038/nrg3744

de Visser J, Elena SF, Fragata I, Matuszewski S (2018) The utility of fitness landscapes and big data for predicting evolution. Heredity (Edinb) 121:401–405. https://doi.org/10.1038/s41437-018-0128-4

Deatherage DE, Kepner JL, Bennett AF, Lenski RE, Barrick JE (2017) Specificity of genome evolution in experimental populations of *Escherichia coli* evolved at different temperatures. Proc Natl Acad Sci USA 114:E1904–E1912. https://doi.org/10.1073/pnas.1616132114

Dehouck Y, Kwasigroch JM, Gilis D, Rooman M (2011) PoPMuSiC 2.1: a web server for the estimation of protein stability changes upon mutation and sequence optimality. BMC Bioinform 12:151. https://doi.org/10.1186/1471-2105-12-151

Eyre-Walker A, Keightley PD (2007) The distribution of fitness effects of new mutations. Nat Rev Genet 8:610–618. https://doi.org/10.1038/nrg2146

Ferguson GC, Bertels F, Rainey PB (2013) Adaptive divergence in experimental populations of *Pseudomonas fluorescens*. V. Insight into the Niche specialist "Fuzzy Spreader" compels revision of the model *Pseudomonas* radiation genetics. https://doi.org/10.1534/genetics.113.154948

Fischer A, Vazquez-Garcia I, Illingworth CJR, Mustonen V (2014) High-definition reconstruction of clonal composition in cancer. Cell Rep 7:1740–1752. https://doi.org/10.1016/j.celrep.2014.04.055

Flowers JM, Hanzawa Y, Hall MC, Moore RC, Purugganan MD (2009) Population genomics of the *Arabidopsis thaliana* flowering time gene network. Mol Biol Evol 26:2475–2486. https://doi.org/10.1093/molbev/msp161

Gerrish PJ, Lenski RE (1998) The fate of competing beneficial mutations in an asexual population. Genetica 102–103:127–144

Gerstein AC, Lo DS, Otto SP (2012) Parallel genetic changes and nonparallel gene-environment interactions characterize the evolution of drug resistance in yeast. Genetics 192:241–252. https://doi.org/10.1534/genetics.112.142620

Goldstein BP (2014) Resistance to rifampicin: a review. J Antibiot (Tokyo) 67:625–630. https://doi.org/10.1038/ja.2014.107

Good BH, Rouzine IM, Balick DJ, Hallatschek O, Desai MM (2012) Distribution of fixed beneficial mutations and the rate of adaptation in asexual populations. Proc Natl Acad Sci USA 109:4950–4955. https://doi.org/10.1073/pnas.1119910109

Good BH, McDonald MJ, Barrick JE, Lenski RE, Desai MM (2017) The dynamics of molecular evolution over 60,000 generations. Nature 551:45–50. https://doi.org/10.1038/nature24287

Gould SJ (1989) Wonderful life: the Burgess Shale and the nature of history, 1st edn. W.W. Norton, New York

Gullberg E, Cao S, Berg OG, Ilback C, Sandegren L, Hughes D, Andersson DI (2011) Selection of resistant bacteria at very low antibiotic concentrations. PLoS Pathog 7:e1002158. https://doi.org/10.1371/journal.ppat.1002158

Harpak A, Bhaskar A, Pritchard JK (2016) Mutation rate variation is a primary determinant of the distribution of allele frequencies in humans. PLoS Genet 12:e1006489. https://doi.org/10.1371/journal.pgen.1006489

Herron MD, Doebeli M (2013) Parallel evolutionary dynamics of adaptive diversification in *Escherichia coli*. PLoS Biol 11:e1001490. https://doi.org/10.1371/journal.pbio.1001490

Hietpas RT, Jensen JD, Bolon DN (2011) Experimental illumination of a fitness landscape. Proc Natl Acad Sci USA 108:7896–7901. https://doi.org/10.1073/pnas.1016024108

Hodgkinson A, Ladoukakis E, Eyre-Walker A (2009) Cryptic variation in the human mutation rate. PLoS Biol 7:e1000027. https://doi.org/10.1371/journal.pbio.1000027

Hudson RE, Bergthorsson U, Ochman H (2003) Transcription increases multiple spontaneous point mutations in *Salmonella enterica*. Nucleic Acids Res 31:4517–4522

Hughes D, Andersson DI (2017) Evolutionary trajectories to antibiotic resistance. Annu Rev Microbiol 71:579–596. https://doi.org/10.1146/annurev-micro-090816-093813

Johnson PL, Hellmann I (2011) Mutation rate distribution inferred from coincident SNPs and coincident substitutions. Genome Biol Evol 3:842–850. https://doi.org/10.1093/gbe/evr044
Kassen R, Bataillon T (2006) Distribution of fitness effects among beneficial mutations before selection in experimental populations of bacteria. Nat Genet 38:484–488. https://doi.org/10.1038/ng1751
Kawecki TJ, Lenski RE, Ebert D, Hollis B, Olivieri I, Whitlock MC (2012) Experimental evolution. Trends Ecol Evol 27:547–560. https://doi.org/10.1016/j.tree.2012.06.001
Kaznatcheev A (2019) Computational complexity as an ultimate constraint on evolution. Genetics. https://doi.org/10.1534/genetics.119.302000
Keightley PD, Eyre-Walker A (2010) What can we learn about the distribution of fitness effects of new mutations from DNA sequence data? Philos Trans R Soc Lond B Biol Sci 365:1187–1193. https://doi.org/10.1098/rstb.2009.0266
Kelley LA, Mezulis S, Yates CM, Wass MN, Sternberg MJ (2015) The Phyre2 web portal for protein modeling, prediction and analysis. Nat Protoc 10:845–858. https://doi.org/10.1038/nprot.2015.053
Knoppel A, Knopp M, Albrecht LM, Lundin E, Lustig U, Nasvall J, Andersson DI (2018) Genetic adaptation to growth under laboratory conditions in *Escherichia coli* and *Salmonella enterica*. Front Microbiol 9:756. https://doi.org/10.3389/fmicb.2018.00756
Koskiniemi S, Andersson DI (2009) Translesion DNA polymerases are required for spontaneous deletion formation in *Salmonella typhimurium*. Proc Natl Acad Sci USA 106:10248–10253. https://doi.org/10.1073/pnas.0904389106
Kovacs AT, Dragos A (2019) Evolved biofilm: review on the experimental evolution studies of *Bacillus subtilis* pellicles. J Mol Biol. https://doi.org/10.1016/j.jmb.2019.02.005
Krasovec R et al (2017) Spontaneous mutation rate is a plastic trait associated with population density across domains of life. PLoS Biol 15:e2002731. https://doi.org/10.1371/journal.pbio.2002731
Kumar MD, Bava KA, Gromiha MM, Prabakaran P, Kitajima K, Uedaira H, Sarai A (2006) ProTherm and ProNIT: thermodynamic databases for proteins and protein-nucleic acid interactions. Nucleic Acids Res 34:D204–D206. https://doi.org/10.1093/nar/gkj103
Lamrabet O, Plumbridge J, Martin M, Lenski RE, Schneider D, Hindre T (2019) Plasticity of promoter-core sequences allows bacteria to compensate for the loss of a key global regulatory gene. Mol Biol Evol. https://doi.org/10.1093/molbev/msz042
Lang GI, Desai MM (2014) The spectrum of adaptive mutations in experimental evolution. Genomics 104:412–416. https://doi.org/10.1016/j.ygeno.2014.09.011
Lässig M, Mustonen V, Walczak AM (2017) Predicting evolution. Nat Ecol Evol 1:0077. https://doi.org/10.1038/s41559-017-0077
Levinson G, Gutman GA (1987) Slipped-strand mispairing: a major mechanism for DNA sequence evolution. Mol Biol Evol 4:203–221. https://doi.org/10.1093/oxfordjournals.molbev.a040442
Liao HX et al (2013) Co-evolution of a broadly neutralizing HIV-1 antibody and founder virus. Nature 496:469–476. https://doi.org/10.1038/nature12053
Lind PA (2018) Evolutionary forecasting of phenotypic and genetic outcomes of experimental evolution in *Pseudomonas*. bioRxiv https://doi.org/10.1101/342261
Lind PA, Andersson DI (2008) Whole-genome mutational biases in bacteria. Proc Natl Acad Sci USA 105:17878–17883. https://doi.org/10.1073/pnas.0804445105
Lind PA, Berg OG, Andersson DI (2010a) Mutational robustness of ribosomal protein genes. Science 330:825–827. https://doi.org/10.1126/science.1194617
Lind PA, Tobin C, Berg OG, Kurland CG, Andersson DI (2010b) Compensatory gene amplification restores fitness after inter-species gene replacements. Mol Microbiol 75:1078–1089. https://doi.org/10.1111/j.1365-2958.2009.07030.x
Lind PA, Farr AD, Rainey PB (2015) Experimental evolution reveals hidden diversity in evolutionary pathways. Elife 4. https://doi.org/10.7554/elife.07074

Lind PA, Arvidsson L, Berg OG, Andersson DI (2017a) Variation in mutational robustness between different proteins and the predictability of fitness effects. Mol Biol Evol 34:408–418. https://doi.org/10.1093/molbev/msw239

Lind PA, Farr AD, Rainey PB (2017b) Evolutionary convergence in experimental *Pseudomonas* populations. ISME J 11:589–600. https://doi.org/10.1038/ismej.2016.157

Lind PA, Libby E, Herzog J, Rainey PB (2019) Predicting mutational routes to new adaptive phenotypes. Elife 8. https://doi.org/10.7554/elife.38822

Long A, Liti G, Luptak A, Tenaillon O (2015) Elucidating the molecular architecture of adaptation via evolve and resequence experiments. Nat Rev Genet 16:567–582. https://doi.org/10.1038/nrg3937

Lovett ST (2004) Encoded errors: mutations and rearrangements mediated by misalignment at repetitive DNA sequences. Mol Microbiol 52:1243–1253. https://doi.org/10.1111/j.1365-2958.2004.04076.x

Lovett ST, Hurley RL, Sutera VA Jr, Aubuchon RH, Lebedeva MA (2002) Crossing over between regions of limited homology in *Escherichia coli*. RecA-dependent and RecA-independent pathways. Genetics 160:851–859

Luksza M, Lassig M (2014) A predictive fitness model for influenza. Nature 507:57–61. https://doi.org/10.1038/nature13087

Lundin E, Tang PC, Guy L, Nasvall J, Andersson DI (2017) Experimental determination and prediction of the fitness effects of random point mutations in the biosynthetic enzyme HisA. Mol Biol Evol. https://doi.org/10.1093/molbev/msx325

Lynch M, Ackerman MS, Gout JF, Long H, Sung W, Thomas WK, Foster PL (2016) Genetic drift, selection and the evolution of the mutation rate. Nat Rev Genet 17:704–714. https://doi.org/10.1038/nrg.2016.104

MacLean RC, Buckling A (2009) The distribution of fitness effects of beneficial mutations in *Pseudomonas aeruginosa*. PLoS Genet 5:e1000406. https://doi.org/10.1371/journal.pgen.1000406

Maddamsetti R, Hatcher PJ, Green AG, Williams BL, Marks DS, Lenski RE (2017) Core genes evolve rapidly in the long-term evolution experiment with *Escherichia coli*. Genome Biol Evol. https://doi.org/10.1093/gbe/evx064

Maharjan RP, Ferenci T (2017) A shifting mutational landscape in 6 nutritional states: stress-induced mutagenesis as a series of distinct stress input-mutation output relationships. PLoS Biol 15:e2001477. https://doi.org/10.1371/journal.pbio.2001477

Malone JG (2015) Role of small colony variants in persistence of *Pseudomonas aeruginosa* infections in cystic fibrosis lungs. Infect Drug Resist 8:237–247. https://doi.org/10.2147/IDR.S68214

McCandlish DM, Stoltzfus A (2014) Modeling evolution using the probability of fixation: history and implications. Q Rev Biol 89:225–252

McDonald MJ, Gehrig SM, Meintjes PL, Zhang XX, Rainey PB (2009) Adaptive divergence in experimental populations of *Pseudomonas fluorescens*. IV. Genetic constraints guide evolutionary trajectories in a parallel adaptive radiation. Genetics 183:1041–1053. https://doi.org/10.1534/genetics.109.107110

McDonald MJ, Cooper TF, Beaumont HJE, Rainey PB (2011) The distribution of fitness effects of new beneficial mutations in *Pseudomonas fluorescens*. Biol Letters 7:98–100. https://doi.org/10.1098/Rsbl.2010.0547

Nasvall J, Sun L, Roth JR, Andersson DI (2012) Real-time evolution of new genes by innovation, amplification, and divergence. Science 338:384–387. https://doi.org/10.1126/science.1226521

Neher RA, Russell CA, Shraiman BI (2014) Predicting evolution from the shape of genealogical trees. Elife 3. https://doi.org/10.7554/elife.03568

Ng PC, Henikoff S (2003) SIFT: predicting amino acid changes that affect protein function. Nucleic Acids Res 31:3812–3814

O'Neill AJ, Huovinen T, Fishwick CW, Chopra I (2006) Molecular genetic and structural modeling studies of *Staphylococcus aureus* RNA polymerase and the fitness of rifampin resistance genotypes in relation to clinical prevalence. Antimicrob Agents Chemother 50:298–309. https://doi.org/10.1128/AAC.50.1.298-309.2006

Orgogozo V (2015) Replaying the tape of life in the twenty-first century. Interface Focus 5:20150057. https://doi.org/10.1098/rsfs.2015.0057
Orr HA (2003) The distribution of fitness effects among beneficial mutations. Genetics 163:1519–1526
Orr HA (2010) The population genetics of beneficial mutations. Philos Trans R Soc Lond B Biol Sci 365:1195–1201. https://doi.org/10.1098/rstb.2009.0282
Otwinowski J (2018) Biophysical inference of epistasis and the effects of mutations on protein stability and function. Mol Biol Evol 35:2345–2354. https://doi.org/10.1093/molbev/msy141
Perfeito L, Fernandes L, Mota C, Gordo I (2007) Adaptive mutations in bacteria: high rate and small effects. Science 317:813–815. https://doi.org/10.1126/science.1142284
Rainey PB, Travisano M (1998) Adaptive radiation in a heterogeneous environment. Nature 394:69–72. https://doi.org/10.1038/27900
Rainey PB, Remigi P, Farr AD, Lind PA (2017) Darwin was right: where now for experimental evolution? Curr Opin Genet Dev 47:102–109. https://doi.org/10.1016/j.gde.2017.09.003
Reams AB, Roth JR (2015) Mechanisms of gene duplication and amplification. Cold Spring Harb Perspect Biol 7:a016592. https://doi.org/10.1101/cshperspect.a016592
Reams AB, Kofoid E, Duleba N, Roth JR (2014) Recombination and annealing pathways compete for substrates in making *rrn* duplications in *Salmonella enterica*. Genetics 196:119–135. https://doi.org/10.1534/genetics.113.158519
Rokyta DR, Beisel CJ, Joyce P, Ferris MT, Burch CL, Wichman HA (2008) Beneficial fitness effects are not exponential for two viruses. J Mol Evol 67:368–376. https://doi.org/10.1007/s00239-008-9153-x
Romling U, Galperin MY, Gomelsky M (2013) Cyclic di-GMP: the first 25 years of a universal bacterial second messenger. Microbiol Mol Biol Rev 77:1–52. https://doi.org/10.1128/mmbr.00043-12
Sanjuan R (2010) Mutational fitness effects in RNA and single-stranded DNA viruses: common patterns revealed by site-directed mutagenesis studies. Philos Trans R Soc Lond B Biol Sci 365:1975–1982. https://doi.org/10.1098/rstb.2010.0063
Sankar TS, Wastuwidyaningtyas BD, Dong Y, Lewis SA, Wang JD (2016) The nature of mutations induced by replication-transcription collisions. Nature. https://doi.org/10.1038/nature18316
Savageau MA, Fasani RA (2009) Qualitatively distinct phenotypes in the design space of biochemical systems. FEBS Lett 583:3914–3922. https://doi.org/10.1016/j.febslet.2009.10.073
Shewaramani S, Finn TJ, Leahy SC, Kassen R, Rainey PB, Moon CD (2017) Anaerobically grown *Escherichia coli* has an enhanced mutation rate and distinct mutational spectra. PLoS Genet 13:e1006570. https://doi.org/10.1371/journal.pgen.1006570
Sommer MOA, Munck C, Toft-Kehler RV, Andersson DI (2017) Prediction of antibiotic resistance: time for a new preclinical paradigm? Nat Rev Microbiol 15:689–696. https://doi.org/10.1038/nrmicro.2017.75
Sousa A, Bourgard C, Wahl LM, Gordo I (2013) Rates of transposition in *Escherichia coli*. Biol Lett 9:20130838. https://doi.org/10.1098/rsbl.2013.0838
Spiers AJ, Kahn SG, Bohannon J, Travisano M, Rainey PB (2002) Adaptive divergence in experimental populations of *Pseudomonas fluorescens*. I. Genetic and phenotypic bases of wrinkly spreader fitness. Genetics 161:33–46
Spiers AJ, Bohannon J, Gehrig SM, Rainey PB (2003) Biofilm formation at the air-liquid interface by the *Pseudomonas fluorescens* SBW25 wrinkly spreader requires an acetylated form of cellulose. Mol Microbiol 50:15–27
Steenackers HP, Parijs I, Dubey A, Foster KR, Vanderleyden J (2016) Experimental evolution in biofilm populations. FEMS Microbiol Rev 40:373–397. https://doi.org/10.1093/femsre/fuw002
Stern DL (2013) The genetic causes of convergent evolution. Nat Rev Genet 14:751–764. https://doi.org/10.1038/nrg3483
Stoltzfus A, McCandlish DM (2017) Mutational biases influence parallel adaptation. Mol Biol Evol 34:2163–2172. https://doi.org/10.1093/molbev/msx180

Sun S, Berg OG, Roth JR, Andersson DI (2009) Contribution of gene amplification to evolution of increased antibiotic resistance in *Salmonella typhimurium*. Genetics 182:1183–1195. https://doi.org/10.1534/genetics.109.103028

Tenaillon O, Rodriguez-Verdugo A, Gaut RL, McDonald P, Bennett AF, Long AD, Gaut BS (2012) The molecular diversity of adaptive convergence. Science 335:457–461. https://doi.org/10.1126/science.1212986

Thulin E, Sundqvist M, Andersson DI (2015) Amdinocillin (Mecillinam) resistance mutations in clinical isolates and laboratory-selected mutants of *Escherichia coli*. Antimicrob Agents Chemother 59:1718–1727. https://doi.org/10.1128/aac.04819-14

Valderrama-Gomez MA, Parales RE, Savageau MA (2018) Phenotype-centric modeling for elucidation of biological design principles. J Theor Biol 455:281–292. https://doi.org/10.1016/j.jtbi.2018.07.009

Van den Bergh B, Swings T, Fauvart M, Michiels J (2018) Experimental design, population dynamics, and diversity in microbial experimental evolution. Microbiol Mol Biol Rev 82. https://doi.org/10.1128/mmbr.00008-18

van Ditmarsch D et al (2013) Convergent evolution of hyperswarming leads to impaired biofilm formation in pathogenic bacteria. Cell Rep 4:697–708. https://doi.org/10.1016/j.celrep.2013.07.026

Viswanathan M, Lacirignola JJ, Hurley RL, Lovett ST (2000) A novel mutational hotspot in a natural quasipalindrome in *Escherichia coli*. J Mol Biol 302:553–564. https://doi.org/10.1006/jmbi.2000.4088

Wang X, Zorraquino V, Kim M, Tsoukalas A, Tagkopoulos I (2018) Predicting the evolution of *Escherichia coli* by a data-driven approach. Nat Commun 9:3562. https://doi.org/10.1038/s41467-018-05807-z

Weinreich DM, Delaney NF, Depristo MA, Hartl DL (2006) Darwinian evolution can follow only very few mutational paths to fitter proteins. Science 312:111–114. https://doi.org/10.1126/science.1123539

Wiser MJ, Ribeck N, Lenski RE (2013) Long-term dynamics of adaptation in asexual populations. Science 342:1364–1367. https://doi.org/10.1126/science.1243357

Wong A, Rodrigue N, Kassen R (2012) Genomics of adaptation during experimental evolution of the opportunistic pathogen *Pseudomonas aeruginosa*. PLoS Genet 8:e1002928. https://doi.org/10.1371/journal.pgen.1002928

Yampolsky LY, Stoltzfus A (2001) Bias in the introduction of variation as an orienting factor in evolution. Evol Dev 3:73–83

Yona AH, Alm EJ, Gore J (2018) Random sequences rapidly evolve into de novo promoters. Nat Commun 9:1530. https://doi.org/10.1038/s41467-018-04026-w

Zhen Y, Aardema ML, Medina EM, Schumer M, Andolfatto P (2012) Parallel molecular evolution in an herbivore community. Science 337:1634–1637. https://doi.org/10.1126/science.1226630

Chapter 5
RetroSpect, a New Method of Measuring Gene Regulatory Evolution Rates Using Co-mapping of Genomic Functional Features with Transposable Elements

Daniil Nikitin, Maxim Sorokin, Victor Tkachev, Andrew Garazha, Alexander Markov and Anton Buzdin

Abstract Transposable elements (TEs) are selfish genetic sequences that proliferate in the host genomes by spreading their copies in new genomic locations. TEs reside in the genomes of all groups of living organisms. TE sequences may be recruited by the host cells to serve as regulatory sites for the neighboring genes. These regulatory sites can be transcription factor binding sites (TFBS), histone modification loci, DNase I hypersensitivity sites, etc. Insertion of a TE in a gene neighborhood changes an equilibrium of regulatory sequences controlling this gene functioning. The more regulatory sites can be identified within gene-proximate TEs, the faster should be the evolution of gene regulation. We proposed a method for measuring evolutionary rates of gene regulation based on relative quantitation of regulatory sites located within TEs next to gene transcriptional start sites. It allows interrogating regulatory evolution for organisms with TE-rich genomes. This method termed RetroSpect was applied first for studying human gene evolution using TFBS co-mapping with the human retroelements (REs). RE is a subgroup of TEs that was active in mammals before and after their radiation. We characterized human genes and molecular pathways either enriched or deficient in RE-linked TFBS regulation for 563 transcription factors in thirteen human cell lines. We found that major groups enriched by RE regulation deal with gene control by microRNAs, olfaction, color vision, fertilization, cellular immune response, amino acids and fatty acids metabolism and detoxication. The *deficient* groups were involved in protein translation, RNA transcription and processing, chromatin organization, and molecular signaling.

D. Nikitin · M. Sorokin · A. Buzdin (✉)
I.M. Sechenov First Moscow State Medical University, Moscow, Russia
e-mail: buzdin@oncobox.com

D. Nikitin · V. Tkachev · A. Garazha · A. Buzdin
Omicsway Corp., Walnut, CA, USA

M. Sorokin · A. Buzdin
Shemyakin-Ovchinnikov Institute of Bioorganic Chemistry, Moscow, Russia

D. Nikitin · A. Markov
Faculty of Biology, Moscow State University, Moscow, Russia

P. Pontarotti (ed.), *Evolution, Origin of Life, Concepts and Methods*,
https://doi.org/10.1007/978-3-030-30363-1_5

5.1 Background

Retroelements (REs) are the only class of transposable elements thought to be active following mammalian radiation. They occupy ~40% of human genome (Lander et al. 2001; Kazazian and Moran 2017) and can regulate human genes in many ways including providing functional transcription factor binding sites (TFBS) (Chuong et al. 2016; Feschotte 2008; Suntsova et al. 2015; ENCODE Project Consortium 2012). This makes REs an important force reshaping genome regulatory landscape (Chuong et al. 2016; Cordaux and Batzer 2009). RE sequences do not share sequence identity with the surrounding genomic regions and pre-integration genomic sites (Lander et al. 2001). Accumulation of RE-linked functional elements, therefore, can indicate rates of evolution of gene transcriptional regulatory modules (Cordaux and Batzer 2009, Thompson et al. 2015, Cheatle Jarvela and Hinman 2015).

Evolution of gene regulatory networks attracts a growing interest due to the development of advanced experimental techniques and bioinformatic methods of high-throughput genetic data analysis (Albert and Kruglyak 2015; Doucet-O'Hare et al. 2016). It can be investigated at different levels. For single genes, it can interrogate *structural* and *functional* features. The former comprise nucleotide substitutions or mutations in gene regulatory regions (Albert and Kruglyak 2015) including RE inserts (Albert and Kruglyak 2015; Badge et al. 2003). In turn, the *functional* features deal with DNA modifications like methylation (Varriale 2014; Zhong 2016) and hydroxymethylation (Kato and Iwamoto 2014), changes of chromatin structure (Meier and Brehm 2014; Turner 2014) marked by histone modifications (Maleszka et al. 2014), and with binding of transcriptional factors (Villar et al. 2014).

The currently available OMICS technologies allow to directly interrogate both types of these properties, e.g., massive profiling of changes in DNA structure by deep sequencing (Royer-Bertrand and Rivolta 2015), or directly finding TFBS and DNA/chromatin modifications using immunoprecipitation-linked isolation and sequencing of the respective DNA loci (Mundade et al. 2014). These methods generate big amounts of genetic data that can be analyzed using appropriate bioinformatic approaches (Kapitonov and Jurka 2008; Barrio et al. 2009; Garazha et al. 2015; Buzdin et al. 2017a, b).

The differential genes obtained using these methods can be classified by some specific features, e.g., involvement in molecular pathways (Borisov et al. 2014, 2017), co-expression patterns and complex gene signatures (Harris et al 2015; Yuryev 2015). Here, the approaches like calculation of *pathway activation level* (Buzdin et al. 2017a, b; Aliper et al. 2017) or quantization of *Gene Ontology* (GO) clustering scores (Ashburner et al. 2000; The Gene Ontology Consortium 2017) can be appropriate to characterize changes in functioning of gene networks (Artemov et al. 2015; Yin et al. 2016).

On the interface of both structural and functional aspects of gene evolution, we proposed a new method termed *RetroSpect* and applied it to analyze the impact of RE-linked regulation for all human genes at the level of TFBS (Nikitin et al. 2018). *RetroSpect* is based on a rationale of measuring proportion of gene-proximate TFBS

hosted by the REs. This proportion is thought to reflect rates of gene regulatory evolution, thus enabling to identify genes with quickly and slowly evolving regulatory modules (Nikitin et al. 2019). For every gene, we took a threshold of 10-kb neighborhood of its transcriptional start site (TSS) because TSS-proximal DNA is enriched in active regulatory modules, such as promoters and enhancers (Danino et al. 2015). The proportions of RE-linked *TFBS* were calculated using high-throughput experimental chromatin immunoprecipitation sequencing (ChIP-Seq) data published by the ENCODE project team for several human cell lines (https://www.encodeproject.org). ChIP-Seq enables direct mapping and quantitation of transcriptional factors binding with DNA, where bigger number of sequencing reads (=hits) means stronger binding for a particular site, and vice versa (Johnson et al. 2007). This approach was found efficient for both levels of data analysis, i.e., individual genes and molecular pathways (Nikitin et al. 2018).

We found that 56% of totally mapped TFBS overlapped with the RE sequences, thus confirming RE status as the major source of TFBS for human cells. All the cell lines investigated showed highly correlated patterns of NGRE distributions in pairwise comparisons. We did the analysis on two evolutionary scales, first roughly corresponding to mammalian radiation and second to divergence of human and New World monkey ancestries. For mammalian radiation, the main processes with enriched RE-linked regulation involved olfaction, color vision, spermatogenesis and fertilization, some kinds of immune and hormonal responses, intracellular molecular trafficking, amino acids, vitamins and fatty acids metabolism, metabolism and detoxication of xenobiotics. In turn, the processes with deficient RE-linked regulation included protein synthesis and ribosome biogenesis, RNA transcription and processing, nuclear chromatin organization, cell cycle, apoptosis, cell contacts, embryo development, basic signaling pathways, cellular stress response, oxidative phosphorylation in mitochondria, and some different aspects of immunity. Remarkably, the cohort of top genes enriched in RE regulation had ~threefold higher content of noncoding RNA genes than for the bottom cohort of the deficient genes. The results at the scale of human ancestry divergence with New World monkeys were highly congruent.

5.2 Methods

A metric (*1*) termed *Gene RE-linked TFBS Enrichment score* (*GRE* score*)* was introduced to quantitate the RE-associated regulatory impact of *individual genes. GRE* of a certain gene is the sum of RE-linked TFBS hits mapped in 10-kb neighborhood of its TSS that is normalized on the mean content of RE-linked TFBS hits for all genes. *GRE* enables to measure regulation of a gene by RE-linked TFBS. However, different genes may have very different numbers of TFBS hits (both RE-linked and not) in their TSS neighborhood. *GRE* score shows enrichment of a gene by RE-linked TFBS hits relatively to other genes, but it does not take into account enrichment of the same gene by the total hits, also including non-RE hits.

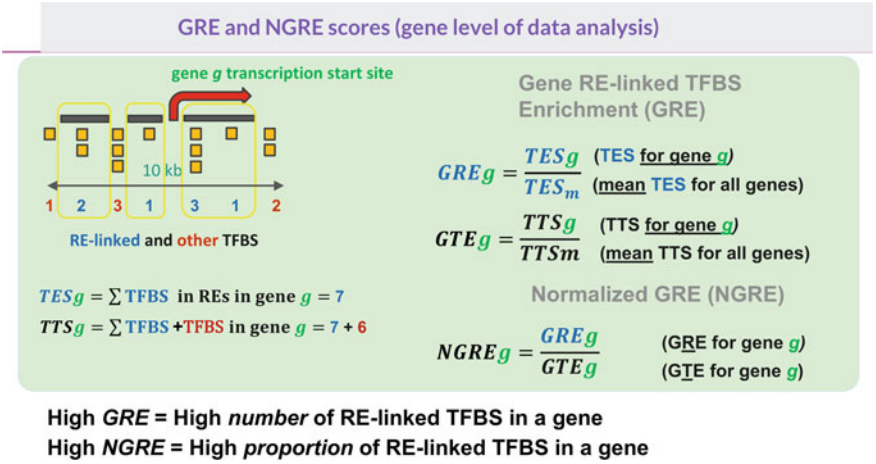


Fig. 5.1 RE insert in the proximity of TSS can create TFBSs and change regulation of gene expression. Cumulative (GRE) and normalized (NGRE) gene-level metrics for measuring RE-linked regulation are explained

Consequently, a double-normalized value was introduced that specifically measures if gene regulation is enriched by RE-linked hits relatively to total hits (*2*) termed *Normalized Gene RE-linked TFBS Enrichment score* (*NGRE*) (Fig. 5.1), equal to the sum of RE-linked hits for a gene under study, double-normalized on (i) average number of RE-linked hits for all genes and (ii) balanced total number of hits for this gene (Nikitin et al. 2018).

Conceptually, similar additive values were proposed at the level of *molecular pathways* (Nikitin et al. 2018) termed *Pathway Involvement Index* (*PII*) and *Normalized Pathway Involvement Index* (*NPII*). *PII* calculates total impact of RE-linked TFBS on regulation of a molecular pathway. The bigger is *PII*, the stronger is the impact of RE-linked hits on the overall regulation of a molecular pathway, and vice versa. However, *PII* is impossible for RE-linked regulation estimates in the context of total regulation of a pathway. We, therefore, introduced another metric termed *Normalized PII* (*NPII*). *NPII* estimates the *relative* RE-linked impact on the pathway regulation, where higher *NPII* means higher *relative* impact in the pathway regulation, and vice versa (Nikitin et al. 2018).

5.2.1 Quantitative Metrics of Genes and Pathways Regulatory Evolution

5.2.1.1 RE-Linked TFBS Impact at the Level of Individual Genes

For every individual gene, GRE score is calculated according to the formula (Eq. 5.1):

$$\mathrm{GRE_g} = \frac{\mathrm{TES_g}}{\frac{1}{n}\sum_{i=1}^{n}\mathrm{TES}_i} \tag{5.1}$$

where $\mathrm{GRE_g}$ is GRE score for a gene g; $\mathrm{TES_g}$ is number of RE-linked TFBS hits for a gene g; i is gene index and TES_i is number of RE-linked TFBS reads for a gene i; n is the total number of genes under study.

Normalized Gene RE-linked TFBS Enrichment score (NGRE score*).*

A double-normalized RE-specific enrichment measure for an individual gene termed NGRE was introduced for a gene g (Eq. 5.2):

$$\mathrm{NGRE_g} = \mathrm{GRE_g}/\mathrm{GTE_g} \tag{5.2}$$

Here, GRE value is calculated according to Formula (5.1) and GTE (*Gene TFBS Enrichment*) value characterizes gene-specific total TFBS content, expressed by the formula (Eq. 5.3):

$$\mathrm{GTE_g} = \frac{\mathrm{TTS_g}}{\mathrm{TTS_m}} \tag{5.3}$$

where $\mathrm{TTS_g}$ is the total number of TFBS hits mapped in the 10-kb neighborhood of gene g and $\mathrm{TTS_m}$ is the mean TTS for all genes under investigation.

5.2.1.2 RE-Linked TFBS Impact at the Level of Molecular Pathways

Pathway Involvement Index (PII) is expressed by the formula (Eq. 5.4):

$$\mathrm{PII_p} = \frac{\sum_{i=1}^{n}\mathrm{GRE}_i}{n} \tag{5.4}$$

where $\mathrm{PII_p}$ is PII score for a pathway p; GRE_i is GRE score for gene i; n is the total number of genes in pathway p.

Normalized Pathway Involvement Index (NPII). Normalized Pathway Involvement Index (NPII) is calculated as follows (Eq. 5.5):

$$\mathrm{NPII_p} = \mathrm{PII_p}/\mathrm{PGI_p} \tag{5.5}$$

where $\mathrm{PII_p}$ is PII for pathway p; $\mathrm{PGI_p}$ is *Pathway Gene-based TFBS Index* for pathway p introduced to assess the impact of total TFBS (not only RE-linked) on the regulation of molecular pathways. PGI for pathway p is expressed by the formula (Eq. 5.6):

$$\mathrm{PGI_p} = \frac{\sum_{i=1}^{n} \mathrm{GTE}_i}{n} \tag{5.6}$$

where GTE_i is GTE score for gene i; n is the number of genes in pathway p.

5.2.2 Extraction and Quantitation of RE-Specific TFBS

For the current application with counting human TFBS, genome-binding profiles of 563 transcription factor proteins were extracted from the ENCODE project repository (ENCODE database, transcription factors 2019a) for 13 human cell lines (K562, HepG2, HEK293, GM12878, MCF-7, A549, HeLa-s3, SK-N-SH, HCT116, Ishikawa, HEK293T, MCF-10A, GM12891) (ENCODE Project Consortium 2012). The reference human genome assembly 2009 (hg19) was indexed via Burrows-Wheeler algorithm using BWA software, version 0.7.10 (ENCODE database, BWA software 2019b). Concatenation of fastq files with single-end or pairwise reads, alignment to the reference genome and filtering were done (ENCODE database, BWA software 2019b). For every cell line, aligned TFBS reads were mapped on the RE sequences previously downloaded from the UCSC Browser (UCSC Browser, Human genome 2019b) and annotated by RepeatMasker software (RepeatMasker 2019). TFBS occurrence data were extracted (UCSC Browser, bedGraph files 2019a) containing conservative IDR-thresholded peaks according to the standard ENCODE ChIP-seq analysis pipeline (ENCODE ChIP-seq analysis pipeline 2019c). The fold change over control profiles for TFBS were built using Macs software (ENCODE ChIP-seq analysis pipeline 2019c). Quality control of TFBS peaks was performed using irreproducible discovery rate correction method (ENCODE ChIP-seq analysis pipeline 2019c).

5.2.3 Evolutionary Age of REs

In every RE family, average divergence from consensus sequence was taken as the measure of evolutionary age. In this study, REs with average divergence smaller than 8% formed *evolutionary younger fraction*, corresponding to evolution of human lineage after divergence with the lineage of New World monkeys (Giordano et al. 2007). An alternative group contained all REs and roughly reflected genome evolution since the origin of major eutherian clades (Giordano et al. 2007). Enrichment by RE-linked TFBS was calculated separately for *all* and *young* RE groups.

5.2.4 Gene Enrichment by RE–Linked TFBS

Coordinates of human genes were taken from the USCS Browser (UCSC Browser, Human genome 2019b). For every gene in each cell line, all REs overlapping with the 10-kb-long neighborhood of its reference TSS were extracted for further analysis. The 10-kb neighborhood covered an interval starting 5 kb upstream and ending 5 kb downstream the TSS. In every cell line, *GRE* and *NGRE* scores were calculated for 24,389 human genes.

5.2.5 Pathway Enrichment by RE–Linked TFBS

Gene compositions of the molecular pathways were extracted from the public databases NCI (2019), Reactome (2019), BioCarta (2019), KEGG (2019) and Pathway Central (2019) in .xml and .biopax formats and implemented in a pathway scoring computational algorithm (Buzdin et al. 2017a, b). Totally, *PII* and *NPII* scores were calculated for 3123 pathways.

5.2.6 Specific Cell Line Distribution Patterns of RE-Linked TFBS

To investigate correlations between RE-linked TFBS distribution patterns in different cell lines, we calculated pairwise Pearson correlation coefficients of *GRE* and *NGRE* values between all cell lines under investigation (Nikitin et al. 2018).

5.2.7 Enrichment Analysis for Groups of Differential Genes

We performed gene ontology (GO) analysis of the top gene sets that were enriched (RRE-enriched) or deficient (RRE-deficient) in RE-linked TFBS regulation using DAVID (DAVID analysis 2019) and Gorilla software (GOrilla 2019). The p values for significance of GO terms enrichment were calculated using a modified Fisher's exact test (Huang et al. 2009a, b), and the threshold for p values was set as 0.05.

5.2.8 Statistical Significance Tests

Significance of correlations was calculated as Pearson correlation coefficient with p value using the Seaborn package (Seaborn 2019). To investigate randomness of

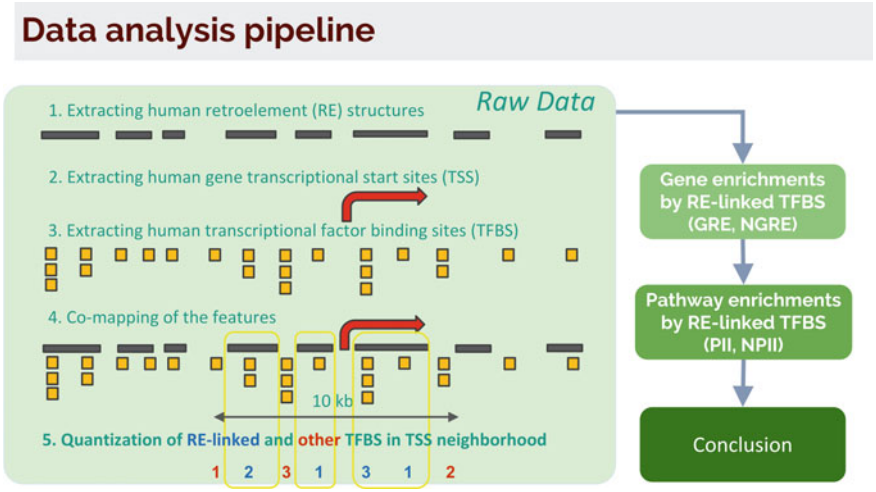


Fig. 5.2 RetroSpect analytic pipeline for TFBS data

observed patterns for RE-linked functional processes, we generated five hundred sets of randomly permuted GRE and NGRE scores across the cell lines tested by randomly rearranging gene names. For each permutation, we extracted a set of 1219 RRE-top and bottom genes (Nikitin et al. 2019). These gene sets were then analyzed using DAVID software and top-100 GO terms were selected for each set by the lowest *p* value for each random permutation. Finally, we compared distributions of *p* values for top-100 GO terms for the permuted and real gene sets. For each comparison, the lowest *p* values for the permuted and real GO terms sets were compared. This randomness test checks the degree of improbability of observed *p* values in RRE-enriched and RRE-deficient gene sets in comparison with random expectation. The data analysis pipeline is shown in Fig. 5.2.

5.3 RetroSpect Application to Detect Genes and Pathways with Quick and Slow Regulatory Evolution Using TFBS Data

5.3.1 Mapping of TFBS

We extracted TFBS data for 563 transcription factor proteins (Nikitin et al. 2019; ENCODE Project Consortium 2012; Sloan et al. 2016) in 13 human cell lines, K562, GM12891, GM12878, HeLa S3, Ishikawa, HEK293, HEK293T, MCF-7, MCF-10A, A549, HepG2, HCT116, and SK-N-SH, representing eight cancerous or transformed human tissues/organs: blood, cervix, kidney, adrenal gland, mammary gland, lung,

liver, and colon (Nikitin et al. 2019). Totally, we mapped ~277 million TFBS hits on human genome sequence. In agreement with previously published data (Lander et al. 2001), REs mapped by the same approach occupied ~45% of human DNA (Nikitin et al. 2019).

The RE class distribution of TSS-proximal TFBS hits was peculiar: ~30% were mapped on sequences SINE class of REs; ~17%—on LINEs and ~7%—on LTR retrotransposons and endogenous retroviruses. These proportions were different for the total fraction of RE-linked TFBS: 28, 26 and 12%, respectively (p values 0.074, 3.7×10^{-7} and 7.4×10^{-12}, respectively). The representation of TFBS associated with LINE and LTR retrotransposons/endogenous retroviruses was, therefore, ~1.5 times lower close to TSS, whereas it remained virtually the same for the SINE-associated TFBS (Nikitin et al. 2019).

For the different DNA-binding proteins, the distribution of TFBS hits also varied significantly. However, the thorough analysis here was impossible because the content of DNA binding proteins investigated was different for the different cell lines (Nikitin et al. 2019).

5.3.2 Gene- and Pathway-Specific Metrics of RE-Linked TFBS Content

We calculated the RE-linked TFBS absolute (*GRE*) and normalized (*NGRE*) enrichment scores for 25,075 human genes. Similarly, we calculated the absolute (*PII*) and normalized (*NPII*) enrichment scores for 3126 molecular pathways. All RE enrichment scores were calculated in two ways: for *all* REs and for *young* REs (those having mean divergence from their consensus sequence of 8% and lower). The scoring for the *young* REs was found less informative than for *all* REs. For example, in cell line K562 with the biggest available number of TFBS profiles, the number of non-informative genes having zero GRE scores was ~37 times bigger for *young* REs compared to *all* REs.

5.3.3 Uniform RE-Linked TFBS Distribution Patterns Among Different Cell Lines

We then compared *NGRE* scores the 13 cell lines investigated (Fig. 5.3). The *NGRE* scores for comparisons with the previously published profiles of leukemia cell line K562 were strongly correlated with Pearson correlation r varying 0.6–0.95 with median value of ~0.9, p values for all comparisons $<10^{-20}$ (Fig. 5.3b) (Nikitin et al. 2019). *NPII* scores were also highly correlated among the different cell lines tested (r 0.75–0.95 with median ~0.85, p values for all comparisons $<10^{-20}$), Fig. 5.3c.

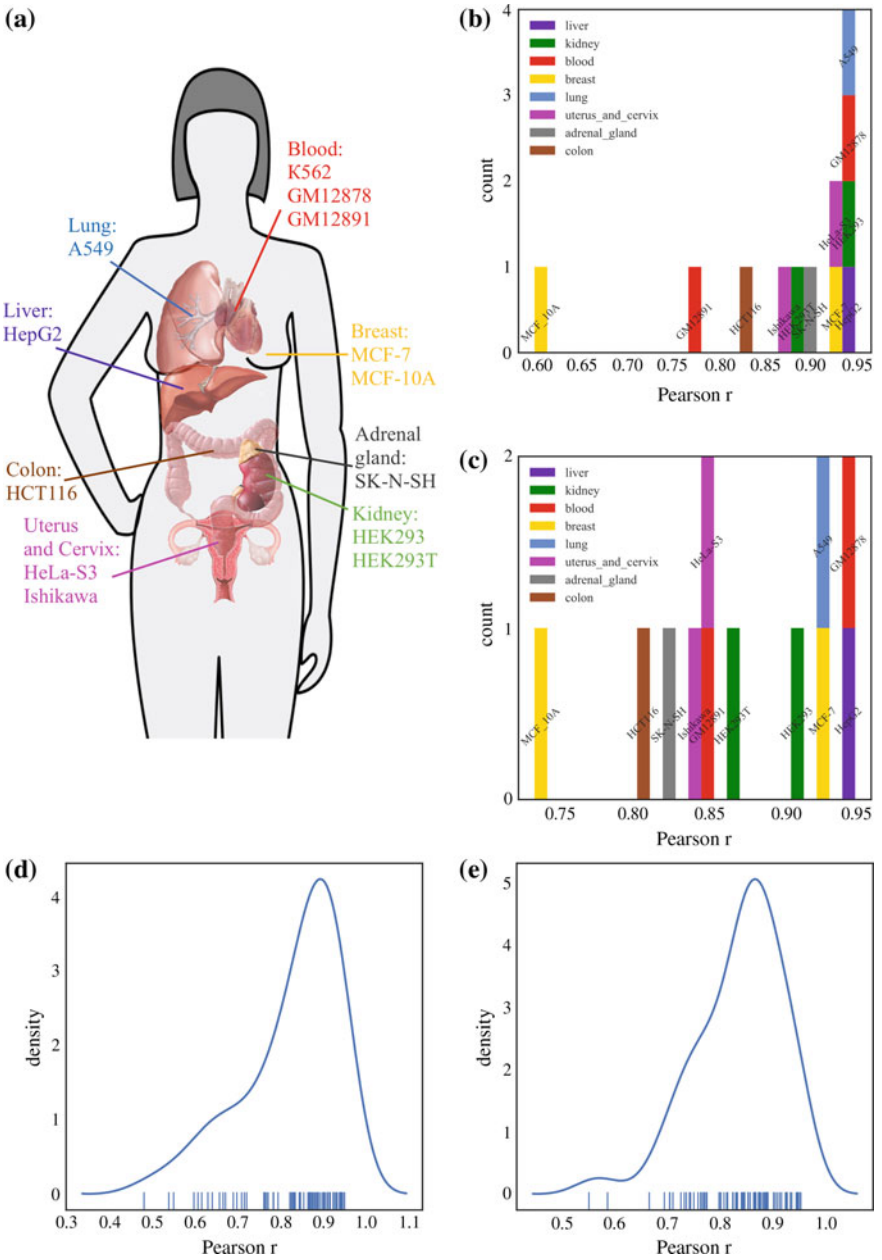


Fig. 5.3 Comparison of *NGRE* and *NPII* scores between 13 human cell lines. Colors denote organ origin of cell lines. **a** Anatomical map of cell line origins. **b** Distribution of Pearson correlation coefficients for *NGRE* score between K562 cell line and 12 other cell lines taken separately. **c** Distribution of Pearson correlation coefficients of *NPII* score between K562 cell line and 12 other cell lines taken separately. **d** Distribution of all possible pairwise Pearson correlation coefficients of *NGRE* scores for 13 cell lines tested. **e** Distribution of all possible pairwise Pearson correlation coefficients of *NPII* scores for 13 cell lines tested. Adapted after Nikitin et al. (2019)

We did not detect any tissue-specific patterns in correlations between K562 and other cell lines, at both gene and pathway levels (Fig. 5.3a–c). Then, we did pairwise comparisons between the 13 cell lines tested and observed a clear-cut unimodal distribution of Pearson correlations varying from ~0.5 to 0.95 with median of 0.88, both at the gene and the pathway levels (Fig. 5.3d and e, respectively, p values for all correlations $<10^{-20}$). This evidences that all the cell lines investigated had congruent RE-linked TFBS regulation patterns.

5.3.4 All RE-Linked TFBS Regulation

Using the specific metrics mentioned above, we aimed to identify top genes and molecular pathways enriched or deficient in RE-linked TFBS regulation (*RRE-enriched* or *RRE-deficient* genes and pathways). To this end, the genes and pathways were examined separately on scatter plots with *abscissa* axis showing *GRE* for genes and *PII* for pathways and *ordinate* axis showing *NGRE* for genes and *NPII* for pathways (Fig. 5.4). This projection makes it possible to segregate genes and pathways having different *RRE* impacts. Among the items with the same *GRE*/*PII* scores, those with higher *NGRE*/*NPII* value are considered *RRE*-enriched. In contrast, those having lower *NGRE*/*NPII* value are thought *RRE-deficient*.

Interestingly, we observed remarkably similar distribution trends among the different cell lines (Fig. 5.4). Moreover, in almost all cell lines investigated, we observed unusual yet very similarly inclined V-shaped distributions (Fig. 5.4). It represented two rays of higher and lower slope coming from a zero point. The possible explanation of this phenomenon is the presence of two biologically different groups of genes: genes with relatively fast RE-driven regulatory evolution and genes whose RE-driven regulation is developing relatively slowly. The upper and lower rays accumulated, therefore, relatively *RRE-enriched* and *RRE-deficient* genes, respectively.

Taking into account that different cell lines showed very similar tags of RE-linked regulation, we then aggregated scores equal to mean *GRE*, *NGRE, PII,* and *NPII* values for all cell lines investigated for further analyses (Fig. 5.4) (Nikitin et al. 2019). We then formalized identification of the top *RRE-enriched* and *deficient* genes and pathways. To this end, we randomly generated 1000 sets for 500 genes each and built a regression line for each of them (Numpy Least squares polynomial fit 2019). Then, two regression lines with the highest and the lowest slopes were identified, and (i) 5% (1219) genes lying above the highest slope regression line with the maximal Euclidean distance from it, or (ii) 1219 genes below the lowest slope regression line with maximal Euclidean distance were taken for further analysis. They were regarded, respectively, *RRE-enriched* and *RRE-deficient* (Fig. 5.4a).

For the 3123 molecular pathways, the same approach was used thus giving 5% (156 in each group) *RRE-enriched* and *RRE-deficient* pathways (Fig. 5.4c) (Nikitin et al. 2019).

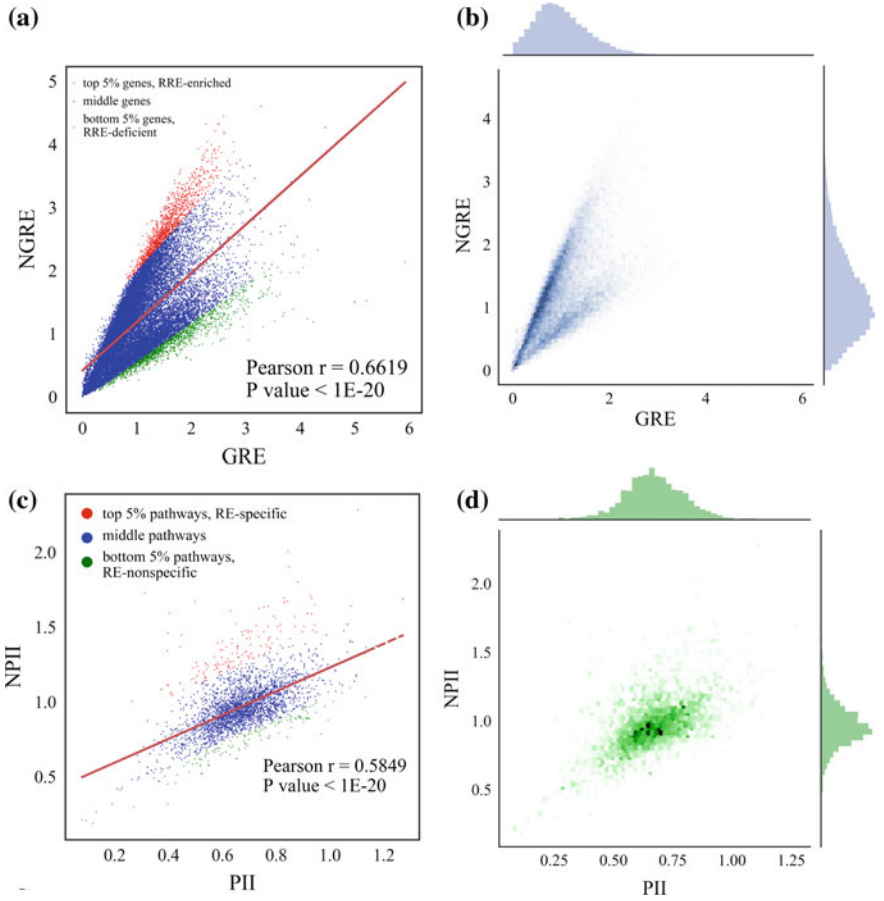


Fig. 5.4 Comparison of gene-level (*GRE, NGRE*) and pathway-level (*PII, NPII*) scores across cell lines. **a** Distribution of mean GRE and NGRE scores, each dot represents single gene, *GRE* and *NGRE* scores were averaged across all cell lines investigated. RRE-enriched genes are shown as red dots, RRE-deficient—as green dots. **b** Distribution of averaged *GRE* and *NGRE* scores. Color depth is congruent with local densities of single genes in each grain. Univariate distributions of *GRE* and *NGRE* are shown in plot margins. **c** Distribution of mean *PII* and mean *NPII* scores, each dot represents single pathway, *PII* and *NPII* scores were averaged across all cell lines investigated. RRE-enriched pathways are shown as red dots, RRE deficient—as green dots. **d** Distribution of averaged *PII* and *NPII* scores. Color depth is congruent with local densities of single pathways in each grain. Univariate distributions of *PII* and *NPII* are shown in plot margins. Adapted after Nikitin et al. (2019)

5.3.5 Top RRE-enriched and RRE-deficient genes (All REs)

To characterize if the selected top and bottom genes (identical numbers of genes were used for both groups) are enriched by the gene clusters involved in specific biological processes, we performed gene ontology (GO) analysis using DAVID software (Huang et al. 2009a, b; Eden et al. 2009). The results of this analysis (GO terms, p values and number of genes for each GO term) are available via the following link (https://is.gd/zuN3iF). Totally, for the *RRE-enriched* genes, we found 141 GO annotation terms ($p < 0.05$) compared to ~7 times bigger number (1022 GO terms) for the *RRE-deficient* genes. This finding may suggest that the number of evolutionary conserved processes prevails over the number of quickly evolving biological processes in human cells.

Manual inspection of these annotation terms enabled to classify them into 27 main functional groups (Table 5.1, Matthews correlation coefficient = 0.756). The *RRE-enriched* groups included the metabolism of amino acids, lipids and metals, detoxication and response to xenobiotics, sensory perception and neurotransmission, and the fertilization. More numerous *RRE-deficient* groups were for stress response and general virus response mechanisms, translation of protein, transcription of RNA, intracellular signaling, cell adhesion and intercellular interaction, progression of cell cycle, programmed cell death, carbohydrates and nucleic acids metabolism, phorbol acetate response, modifications of proteins, regulation of chromatin organization, electron transfer chain, and functioning of mitochondria.

Some groups showed contradictory trends. For example, there were 50 annotation terms linked with the immunity. They were distributed differently among the *RRE-enriched* and *deficient* genes in relation to their functional roles. The immune cells migration/activation and cellular immune response by T- and NK cells processes were *RRE-enriched*, whereas processes related to B cells functioning were *RRE-deficient* (Table 5.2).

An alternative way of GO analysis by Gorilla software with more stringent threshold (Eden et al. 2009) returned significantly smaller amount of hits. There was only one very strongly statistically significant *RRE-enriched* ($p < 10^{-9}$) and two *RRE-deficient* processes ($p < 10^{-3}$), respectively, for the "gene silencing by microRNAs" versus "Regulation of JNK cascade" and "Regulation of stress-activated MAPK cascade".

We then compared contents of noncoding RNA genes among the top *RRE-enriched* and *deficient* genes (Table 5.3). The *enriched* group had 177 microRNA (miR) genes versus 72 in the *deficient* group ($p < 10^{-17}$). Congruently, the number of long noncoding RNA (lncRNA) genes was 150 in the *enriched* versus only 18 in the *deficient* group ($p < 10^{-16}$; Table 5.3). In contrast, the *RRE-deficient* group was not enriched in both lncRNA and miR genes. These findings evidence that regulation of miR and lncRNA genes was quickly evolving in human molecular history.

Table 5.1 Groups of RRE-enriched and deficient processes by GO and molecular pathway analysis (*all* REs)

ID	Group	RRE enrichment, pathway analysis		RRE enrichment, GO analysis		Final status
		Enriched pws	Deficient pws	Enriched GO terms	Deficient GO terms	
1	Posttranscriptional silencing by small RNAs	1	0	1	0	Enriched
2	DNA repair	2	0	5	0	Enriched
3	Metabolism of aminoacids, peptides and polyamines	20	5	13	8	Enriched
4	Lipid metabolism	14	7	11	0	Enriched
5	Detoxication, metabolism of xenobiotics and rare molecules	13	0	4	0	Enriched
6	Sensory perception and neurotransmission	7	0	10	0	Enriched
7	Fertilization	1	0	9	0	Enriched
8	Cellular immunity (T cells and NK cells)	11	0	7	6	Enriched
9	Nucleic base, nucleosides and nucleotides metabolism	6	9	0	24	Deficient
10	DNA metabolism and chromatin structure	0	4	0	151	Deficient
11	Translation and protein quality control	0	12	8	130	Deficient
12	Intracellular signaling	22	94	5	48	Deficient
13	Response to viruses	0	3	0	17	Deficient
14	Vitamin metabolism	4	0	0	0	Enriched
15	Response to hormones	6	0	0	0	Enriched

(continued)

Table 5.1 (continued)

ID	Group	RRE enrichment, pathway analysis		RRE enrichment, GO analysis		Final status
		Enriched pws	Deficient pws	Enriched GO terms	Deficient GO terms	
16	Molecular transport	10	0	0	0	Enriched
17	Sulfur metabolism and linked redox reactions	5	0	0	0	Enriched
18	Metal metabolism	0	0	6	0	Enriched
19	Response to phorbol acetate	0	0	0	3	Deficient
20	Electron transfer reactions	0	0	5	17	Deficient
21	Mitochondria	0	0	5	17	Deficient
22	RNA synthesis and degradation	0	0	0	139	Deficient
23	Cell adhesion and interaction	0	0	0	15	Deficient
24	Cell cycle and mitosis	0	0	0	55	Deficient
25	Cell death	0	0	0	41	Deficient
26	Protein localisation and modification	0	0	0	19	Deficient
27	Response to physical and chemical stress	0	0	0	24	Deficient
28	Carbohydrates metabolism	5	3	0	9	Unclear
29	Immunity	36	16	23	45	Shown separately
30	Other/too general terms	0	0	13	17	N/A

5.3.6 Top RRE-Enriched and Deficient Pathways (All REs)

We then examined top *RRE-enriched* and *deficient* molecular pathways. The top five *RRE-enriched* and *deficient* pathways sorted by *NPII* are shown in Fig. 5.5. We analyzed sets of molecular pathways and sorted them according to the functional groups (Table 5.1).

The *RRE-enriched* pathways related to molecular transport, response to and production of hormones, detoxication and response on xenobiotics, metabolism of amino acids, vitamins, lipids, sulfur and carbohydrates, sensory perception and neurotransmission, and to fertilization. The *RRE-deficient* pathways were for the functional

Table 5.2 RRE-enrichment status of immunity processes by GO and molecular pathway analysis (*all* REs)

Group	RRE enrichment, pathway analysis		RRE enrichment, GO analysis		Final status
	Enriched pws	Deficient pws	Enriched GO terms	Deficient GO terms	
Autoimmunity	4	0	0	0	Enriched
Blood clotting	2	0	0	0	Enriched
Innate immunity	8	0	0	5	Ambiguous
Inflammation	3	5	0	0	Ambiguous
Cellular immune response (T cells and NK cells)	11	0	7	6	Enriched
Activation of antigen-presenting cells by T-helper cells	2	7	0	0	Deficient
Other/too general terms	6	1	8	11	Ambiguous
Immune cells migration and activation	0	0	7	0	Enriched
Activity and maturation of B cells	0	0	0	6	Deficient

Table 5.3 RRE-enriched and deficient microRNA and lncRNA genes (*all* REs)

Gene set	MiRs	Genes, totally	Hypergeometric *p* value	Hypothesis tested
RRE-enriched	177	1219	2.4160E−18	miRNA are enriched
RRE-deficient	72	1219	0.0138	miRNA are not enriched
Totally—1865 miRNA genes in 25,075 human genes				
Gene set	lncRNAs	Genes, totally	Hypergeometric *p* value	Hypothesis tested
RRE-enriched	150	1219	1.9500E−17	lncRNA are enriched
RRE-deficient	18	1219	2.42E−16	lncRNA are not enriched
Totally—1505 lncRNA genes in 25,075 human genes				

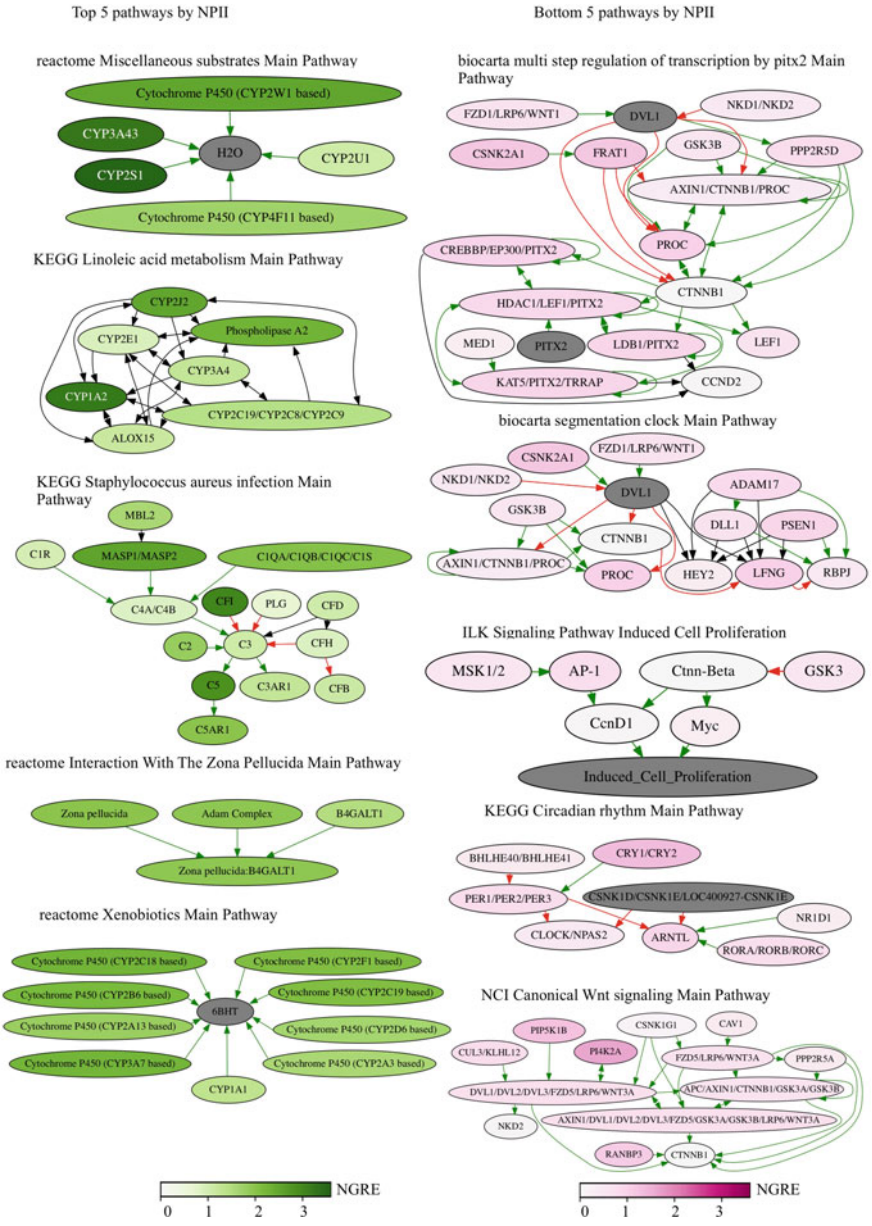


Fig. 5.5 Top ten *RRE-enriched* and *deficient* pathways for *all* REs. Adapted after Nikitin et al. (2019)

groups of protein translation, intracellular signaling including cell adhesion and intracellular interactions, cell cycle progression and programmed cell death, metabolism of nucleic acids, virus response mechanisms and control of chromatin organization. Remarkably, these functional groups showed ~74% overlap for the levels of genes and molecular pathways (Table 5.1, Matthews correlation coefficient = 0.756).

The pathways related to immunity showed differential trends depending on their functional roles (Nikitin et al. 2019). As for the *gene* level of data analysis, we found that the group of processes of cellular immune response by T- and NK cells was *RRE-enriched* (Table 5.2). Other *RRE-enriched* processes dealt with the blood clotting, autoimmunity, and innate immunity mechanisms. The *RRE-deficient* immunity pathways regulated the mechanisms of inflammation and the T-helper-dependent activation of antigen-presenting cells.

5.3.7 Intersection of Gene- and Pathway-Based Results (All REs)

The gene- and pathway-level results were compared to identify similarly regulated groups of processes (Table 5.1). The results were highly congruent and many processes showed consensus regulation patterns. The *RRE-enriched* groups of processes were: (1) silencing by small RNAs; (2) DNA repair; metabolism of (3) amino acids and (4) lipids; (5) detoxication and metabolism of xenobiotics; (6) sensory perception and neurotransmission; (7) fertilization; (8) T- and NK-cellular immune response.

Posttranscriptional silencing by small RNAs (1) was represented by a unique pathway and only one GO term for gene regulation by microRNAs, but according to GO analysis, it was the most statistically strongly enriched cluster among the *RRE-enriched* genes.

DNA repair (2) group of processes contained the pathway of Protein kinase dependent nonhomologous end joining and three GO terms of Mitotic recombination, Meiosis and Strand displacement and DNA synthesis involved in DNA repair.

Metabolism of amino acids (3) group had pathways of D-arginine and D-ornithine metabolism, of aspartate, asparagine, lysine, diphtamide, carnitine biosynthesis and of cysteine, proline, hydroxyproline, beta-alanine, tryptophan and L-kynurenine catabolism, of glutamate removal from folates, and of salicylate and benzoate conjugation with glycine. The GO terms identified dealt with the tryptophan catabolism and proline metabolism biochemical pathways.

Metabolism of lipids (4) group had many pathways regulating bile secretion and transport by SLC transporters, fatty acids cycling and acyl-CoA hydrolysis, alpha-linolenic acid metabolism, alpha-oxidation of phytanate and beta-oxidation of unsaturated fatty acids, lipoxin biosynthesis, ether lipid metabolism pathway, synthesis of (16–20)–hydroxyeicosatetraenoic, epoxyeicosatrienoic and dihydroxyeicosatrienoic acids, and phosphatidylinositol acyl chain remodeling. The GO terms identified

included lipid catabolic processes, arachidonic acid metabolism, epoxygenase and monooxygenase activity, carboxylic ester hydrolase activity and for lipase activities.

Detoxication and metabolism of xenobiotics (5) group had pathways for nicotine, heme, and bupropion degradation, for CYP2E1 reactions and cytochrome P450-dependent drugs metabolism, for caffeine metabolism and flavin-containing monooxygenase (FMO) oxidation of nucleophiles, for S-reticuline metabolism, formaldehyde oxidation, aflatoxin activation and detoxification. There were GO terms featuring epoxygenase P450 pathway, xenobiotic metabolic process and cellular response to xenobiotic stimuli.

Sensory perception and neurotransmission (6) group of processes included pathways for visual signal perception via cones, olfactory sensing and transduction, pathways for dopamine receptors functioning, GABA A (rho) receptor activation, and mechanisms of acetaminophen activity and toxicity. The GO terms included sensory perception of smell, odorant binding and olfactory receptor activity.

Fertilization (7) group had a unique pathway of interaction with the zona pellucida, and several GO terms: binding of sperm to zona pellucida, sperm-egg recognition, fertilization, single fertilization, sperm flagellum, and positive regulation of sperm motility.

T- and NK-cellular immune response (8) group had pathways regulating downstream signaling in naive CD8 T-cells (alpha, beta T-cell proliferation), phosphorylation of CD3 and T-cellular receptor zeta chains, and CD28 co-stimulation in T-cell homeostasis. The GO terms were identified for natural killer cell mediated immunity and its regulation, and for the regulation of leukocyte mediated cytotoxicity.

The *RRE-deficient* groups of processes were: (9) nucleotide and DNA metabolism; (10) maintaining and modulation of chromatin structure; (11) protein translation and ribosome biogenesis; (12) intracellular signaling pathways; (13) intracellular mechanisms of antiviral response.

Specifically, *Nucleotide and DNA metabolism* (9) group pathways regulated salvage of adenine and adenosine, UMP biosynthesis, biosynthesis of UDP-*N*-acetyl-D-galactosamine and of GDP-L-fucose, cleavage of the damaged purines and NADH repair. The GO terms identified dealt with metabolism of ATP, purine nucleosides and ribonucleotides, with adenine transport, GTP and pyrimidine binding.

Maintaining and modulation of chromatin structure (10) group included pathways of HDACs histone deacetylation and of HDAC proteasomal degradation, of PRC2-dependent methylation of histones and DNA, of histone arginine residues methylation by arginine methyltransferases (RMTs) and of G2/M DNA damage checkpoint. The GO terms identified were for the purine NTP-dependent helicase activity, histone exchange, histone lysine H3-K4 and H3-K9 methylation, acetylation and deacetylation, telomere formation and capping, for CENP-dependent centromere formation, for lamin–chromatin binding, and for signal transduction in response to DNA damage.

Protein translation and ribosome biogenesis (11) group pathways regulated transcription of ribosomal RNAs and transfer RNAs, processing of ribosomal and transfer

RNAs including chemical modifications such as biosynthesis of wybutosine and 7-3-amino-3-carboxypropyl-wyosine, tRNA aminoacetylation, ribosome assembly, ribosomal scanning, initiation, elongation and termination of translation, and nonsense mediated decay. The GO terms identified completely matched the above molecular pathways data.

Signaling pathways (12) group of molecular processes was the most numerous because it combined 94 different molecular pathways and 48 GO terms, which covered all major branches of intracellular molecular signaling (Table 5.1).

Cellular mechanisms of antiviral response (13) group covered pathways regulating interaction of host cells with retroviruses, including APOBEC3-mediated resistance to HIV-1 infection. The GO terms identified dealt with the cell-virus interplay in assembly of viral capsids, viral transcription, and viral translation, including IRES-dependent and cap-independent initiation of translation.

The above-mentioned functional groups of molecular processes had different numbers of pathways and GO terms (Table 5.1; scaled representation is shown in Fig. 5.6). There were also 14 other groups (50%) found to be either RRE-enriched or deficient using only one of the methods used (by *gene*- or *pathway*-RRE score). There was only one group showing ambiguous RRE trends according to *gene*- and *pathway* data that accounted for metabolism of carbohydrates (~4% of the number of groups identified; Table 5.1) (Nikitin et al. 2019).

These results are in a good agreement with many previous findings obtained using different methods by other investigators. Namely, the processes linked with protein translation, nucleotide metabolism and DNA synthesis were previously found to be especially highly conserved in all domains of life (Fox 2010; Caetano-Anollés et al. 2009). Here, we confirmed the slow regulatory evolution rates for these processes in human lineage. Intriguingly, the DNA repair regulatory network that is quickly developing according to our findings is known from the other side by its high redundancy and promiscuity of its enzymatic systems (O'Brien 2006). It has an explanation that it needs to be finely tuned to prevent proliferative disorders, especially in large long-living organisms (Hoeijmakers 2009). Similarly, quick changes in the regulation of lipids and amino acids metabolism are also in line with the high nutritional plasticity and the enclosing molecular mechanisms in the human ancestral lineage (Caetano-Anollés et al. 2009). In our study, the most fascinating evolutionary regulatory rates were observed for the mechanisms of gene silencing by noncoding RNAs. This is a major mechanism repressing newly inserted REs, viruses, and other intracellular pathogens (Hoeijmakers 2009). Its regulatory networks are permanently developing, being deeply implicated in the evolutionary arms race between the host genes and genomes, REs and internal pathogens. In contrast, the immunity-related processes showed contradictory trends: T- and NK-cellular immunities (primarily responsible for intercellular antiviral response) were found to be evolving more rapidly, whereas mechanisms of intracellular antiviral defense showed relatively slow rates of regulatory evolution. This was in line with the published reports uncovering that the cellular antiviral immune response, especially T-cell mediated adaptive response, is a more recent evolutionary innovation than the intracellular immunity mechanisms (Boehm and Swann 2014).

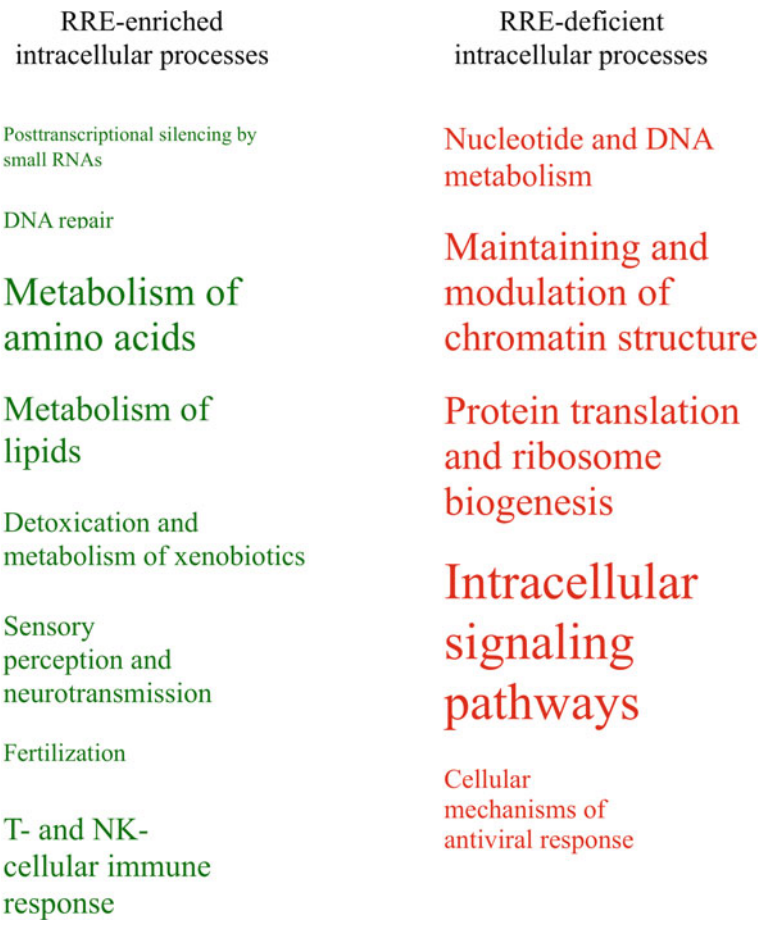


Fig. 5.6 *RRE-enriched* and *deficient* molecular processes for *all* REs. Adapted after Nikitin et al. (2019)

5.3.8 Randomness Test for RetroSpect Assay

The randomness test we performed utilized 500 random permutations generated for each of the GRE or NGRE value sets that have been averaged across cell lines by randomly rearranging gene names. We then analyzed the random gene sets using DAVID software and for each permutation selected top-100 GO terms by smallest p value. The distributions of p values for the *real* and *random* gene sets were then compared thus showing whether statistic characteristics observed for the *real* distribution outperform those for the *random* gene sets (Nikitin et al. 2019). This type of analysis evidenced the non-randomness of RRE-deficient molecular processes identified, because *random* and *real* distributions did not intersect (Nikitin et al. 2019). However, most of the *real* RRE-enriched GO terms showed an overlap with

the distribution for the *random* gene sets, although several *real* RRE-enriched hits had smaller p values than each of the *random* hits. Finally, none of the *random* permutations could generate GO terms with p values lower than those observed for the *real* enriched gene sets, thus returning overall confidence q values for both groups less than 0.002. This evidences high confidence level of the identified differential molecular processes.

5.3.9 RRE Analysis for the Young REs

In the same way, we analyzed top RRE-enriched and deficient molecular processes for the fraction of *young* human REs diverged less than 8% from their consensus sequence. The number of genes having non-zero GRE score was lower (only 44%) than for the case of *all* REs, and we identified top 673 RRE-enriched and 673 RRE-deficient genes. The subsequent GO analysis revealed 55 annotation terms for the *RRE-enriched* genes, and as much as 730 (~13 times more) terms for the *RRE-deficient* genes, thus evidencing higher number of evolutionary conserved than rapidly evolving molecular processes. As for the fraction of *all* REs, we manually curated the GO annotation terms identified and classified them into 24 major groups. Alternative analysis using Gorilla software showed statistically significant organization of GO terms only in *RRE-deficient* gene set where a network of processes linked with nervous system and reproductive organs development was identified, $p < 10^{-3}$ (Nikitin et al. 2019).

For the *young* REs, overlap of the processes identified at the gene and pathway levels was ~50% (Matthews correlation coefficient = 0), which was lower than in the case of *all* REs. At the pathway level, the processes identified for the *young* REs were highly congruent with the results for *all* REs. For example, the same groups of differential immunity processes were identified: innate immunity, T-cell mediated immunity and inflammation. In the latter two groups, more pathways were RRE-enriched than deficient (Nikitin et al. 2019).

We then compared RRE-enriched and deficient groups of processes identified at both gene and pathway levels. The consistent *RRE-enriched* or *deficient* processes identified for *young* REs were identical to those found for *all* REs, with only one exception of a new RRE-deficient group "cytoskeleton, cell adhesion and migration". It included pathways controlling actin cytoskeleton reorganization, cell adhesion and migration, intercellular junctions and fusion of myoblasts. This group represented molecular processes that, according RE-regulation landmarks, were developing especially slowly in the recent human lineage genome history compared to complete evolutional horizon.

5.4 Potentials and Limitations of RetroSpect Method

The RetroSpect method makes it possible to combine two levels of data analysis (genes and molecular pathways) to identify molecular processes impacted by the RE-linked functional features, on the different evolutionary time scales. For human TFBS data, the *gene* level analysis was performed by analyzing enrichment of gene ontology (GO) annotation terms. The analysis at the *pathway* level was performed by interrogating molecular pathway databases included in the Oncobox collection of molecular pathways. These approaches helped to identify different groups of molecular processes that are enriched or deficient in RE-linked gene regulation by TFBS. Based on RE divergence from the respective consensus sequences, two different evolutionary time scales were set for this type of analysis: at the level of mammalian radiation (*all* REs) and at the level of radiation of human ancestry with New World monkeys (*young* REs). Of them, for *all* REs, 13 groups of molecular processes (46%) showed identical trends for the both types of analysis, whereas 14 (50%) were identified using only one of the analytic levels used. Only one group of processes (4%) had contradictory trends at the different levels of data analysis.

For the fraction of *young* REs, eight processes (33%) showed similar trends for the two levels of data analysis used, whereas most of the processes ambiguous results were obtained. This may suggest that the RetroSpect method has different productivity for different evolutionary horizons. For the deeper time scale, such as mammalian radiation as investigated for *all* REs, the results were more robust and reproducible than for the relatively recent horizon of human ancestry—radiation with New World monkeys.

REs are one of the major sources of TFBS for the mammalian DNA (Lander et al. 2001; Suntsova et al. 2015; Cordaux and Batzer 2009; Nikitin et al. 2019), which makes them powerful pacemakers of human genome evolution (Cheatle Jarvela and Hinman 2015). Short polymorphisms and nucleotide substitutions account for permanent yet relatively slow alterations in the gene regulatory networks (Lynch et al. 2016). At the same time, every RE insert can result in a deep restructuring of genomic background by donating an up to 10-kb-long entirely new sequence (Burns and Boeke 2012) containing regulatory elements such as TFBS (Lavialle et al. 2013) influencing gene expression. Further mutations and epigenetic regulatory marks permanently transform these RE-linked regulatory features thus accelerating changes in gene regulation (Garazha et al. 2015; Nikitin et al. 2019). A high proportion of RE-linked regulatory features, therefore, relates to quickly evolving gene regulatory modules. Contrarily, low proportion of RRE features suggests slower evolution of gene regulation and, therefore, more conserved regulatory network. Using RetroSpect technique, therefore, enables finding genes and molecular processes that can be regarded as having outstandingly quick or slow rates of gene regulatory evolution, on the different evolutionary horizons.

In the previous applications, RetroSpect helped to explore evolution of gene regulatory modules at the level of TFBS. However, the same approach should be also worth applicable to high-throughput evolutionary profiling of other functional

genomic marks, e.g., DNA methylation, nucleosome positioning, histone modifications, and DNase hypersensitivity sites. Integration of these data can shed light on the real regulatory potential and evolutionary impact of TEs and measure it. RetroSpect can be applied to molecular data collected for various organisms, which can lead to comparison of gene regulation strategies in the different evolutionary clades. Feasibility of using RetroSpect in non-human organisms depends on (i) availability of high-throughput functional genomic data such as of TFBS profiles, (ii) knowledge and formalization of molecular pathways in the organisms under study, and (iii) abundance of REs or other transposable elements in the genome.

As shown for the human TFBS data, the RetroSpect approach had an important limitation that its resolution of different evolutionary horizons was relatively low, thus enabling investigation of only two key points, i.e., radiation of ancestries of humans and New World monkeys and radiation of major eutherian clades. Comparative analysis of the same evolutionary horizon between two or more different species can help understanding genomic determinants responsible for phenotypic macroevolution. We believe that future applications of RetroSpect for various organisms and genomic marks will help researchers accelerate studies of molecular evolution.

Acknowledgements We acknowledge Amazon and Microsoft Azure grants for cloud-based computations which helped us to complete this study. We thank Oncobox/OmicsWay research program in machine learning and digital oncology for providing access to software and pathway databases. The authors (A.B and M.S.) were supported by the Russian Science Foundation grant no. 18-15-00061.

Conflicts of Interests The authors declare that they have no competing interests.

References

Albert FW, Kruglyak L (2015) The role of regulatory variation in complex traits and disease. Nat Rev Genet 16(4):197–212. https://doi.org/10.1038/nrg3891

Aliper AM, Korzinkin MB, Kuzmina NB, Zenin AA, Venkova LS, Smirnov PY, Borisov NM (2017) Mathematical justification of expression-based pathway activation scoring (PAS). Methods Mol Biol 1613:31–51. https://doi.org/10.1007/978-1-4939-7027-8_3

Artemov A, Aliper A, Korzinkin M, Lezhnina K, Jellen L, Zhukov N, Buzdin A (2015) A method for predicting target drug efficiency in cancer based on the analysis of signaling pathway activation. Oncotarget 6(30):29347–29356. https://doi.org/10.18632/oncotarget.5119

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Sherlock G (2000) Gene ontology: tool for the unification of biology. Nat Genet 25(1):25–29. https://doi.org/10.1038/75556

Badge RM, Alisch RS, Moran JV (2003) ATLAS: a system to selectively identify human-specific L1 insertions. Am J Hum Genet 72(4):823–838. https://doi.org/10.1086/373939

Barrio AM, Lagercrantz E, Sperber GO, Blomberg J, Bongcam-Rudloff E (2009) Annotation and visualization of endogenous retroviral sequences using the distributed annotation system (DAS) and eBioX. BMC Bioinf 10(Suppl 6):S18. https://doi.org/10.1186/1471-2105-10-s6-s18

BioCarta (2019) Available online: https://cgap.nci.nih.gov/Pathways/BioCarta_Pathways. Cited 26 Mar 2019

Boehm T, Swann JB (2014) Origin and evolution of adaptive immunity. Annu Rev Anim Biosci 2(1):259–283. https://doi.org/10.1146/annurev-animal-022513-114201

Borisov N, Suntsova M, Sorokin M, Garazha A, Kovalchuk O, Aliper A, Buzdin A (2017) Data aggregation at the level of molecular pathways improves stability of experimental transcriptomic and proteomic data. Cell Cycle 16(19):1810–1823. https://doi.org/10.1080/15384101.2017.1361068

Borisov NM, Terekhanova NV, Aliper AM, Venkova LS, Smirnov PY, Roumiantsev S, Buzdin AA (2014) Signaling pathways activation profiles make better markers of cancer than expression of individual genes. Oncotarget 5(20):10198–10205. https://doi.org/10.18632/oncotarget.2548

Burns KH, Boeke JD (2012) Human transposon tectonics. Cell 149(4):740–752. https://doi.org/10.1016/j.cell.2012.04.019

Buzdin AA, Prassolov V, Garazha AV (2017a) Friends-enemies: endogenous retroviruses are major transcriptional regulators of human DNA. Front Chem 5. https://doi.org/10.3389/fchem.2017.00035

Buzdin AA, Prassolov V, Zhavoronkov AA, Borisov NM (2017b) Bioinformatics meets biomedicine: OncoFinder, a quantitative approach for interrogating molecular pathways using gene expression data. Methods Mol Biol 1613:53–83. https://doi.org/10.1007/978-1-4939-7027-8_4

Caetano-Anollés G, Yafremava LS, Gee H, Caetano-Anollés D, Kim HS, Mittenthal JE (2009) The origin and evolution of modern metabolism. Int J Biochem Cell Biol 41(2):285–297. https://doi.org/10.1016/j.biocel.2008.08.022

Cheatle Jarvela AM, Hinman VF (2015) Evolution of transcription factor function as a mechanism for changing metazoan developmental gene regulatory networks. Evodevo 6(1):3. https://doi.org/10.1186/2041-9139-6-3

Chuong EB, Elde NC, Feschotte C (2016) Regulatory evolution of innate immunity through co-option of endogenous retroviruses. Science 351(6277):1083–1087. https://doi.org/10.1126/science.aad5497

Cordaux R, Batzer MA (2009) The impact of retrotransposons on human genome evolution. Nat Rev Genet 10(10):691–703. https://doi.org/10.1038/nrg2640

Danino YM, Even D, Ideses D, Juven-Gershon T (2015) The core promoter: at the heart of gene expression. Biochim Biophys Acta Gene Regul Mech 1849(8):1116–1131. https://doi.org/10.1016/j.bbagrm.2015.04.003

DAVID (2019) DAVID functional annotation bioinformatics microarray analysis. Available online: https://david.ncifcrf.gov/. Cited 26 Mar 2019

Doucet-O'Hare TT, Sharma R, Rodić N, Anders RA, Burns KH, Kazazian HH (2016) Somatically acquired LINE-1 insertions in normal esophagus undergo clonal expansion in esophageal squamous cell carcinoma. Hum Mutat 37(9):942–954. https://doi.org/10.1002/humu.23027

Eden E, Navon R, Steinfeld I, Lipson D, Yakhini Z (2009) GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. BMC Bioinf 10(1):48. https://doi.org/10.1186/1471-2105-10-48

ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. Nature 489(7414):57–74. https://doi.org/10.1038/nature11247

ENCODE (2019a) ENCODE database, transcription factors. Available online: https://www.encodeproject.org/chip-seq/transcription_factor/ Cited 26 Mar 2019

ENCODE Database, BWA Software (2019b) Available online: https://www.encodeproject.org/pipelines/ENCPL220NBH/. Cited 26 Mar 2019

ENCODE ChIP-seq Analysis Pipeline (2019c) Available online: https://www.encodeproject.org/pipelines/ENCPL138KID/. Cited 26 Mar 2019

Feschotte C (2008) Transposable elements and the evolution of regulatory networks. Nat Rev Genet 9(5):397–405. https://doi.org/10.1038/nrg2337

Fox GE (2010) Origin and evolution of the ribosome. Cold Spring Harb Perspect Biol 2(9):a003483–a003483. https://doi.org/10.1101/cshperspect.a003483

Garazha A, Ivanova A, Suntsova M, Malakhova G, Roumiantsev S, Zhavoronkov A, Buzdin A (2015) New bioinformatic tool for quick identification of functionally relevant endogenous retroviral inserts in human genome. Cell Cycle 14(9):1476–1484. https://doi.org/10.1080/15384101.2015.1022696

Giordano J, Ge Y, Gelfand Y, Abrusán G, Benson G, Warburton PE (2007) Evolutionary history of mammalian transposons determined by genome-wide defragmentation. PLoS Comput Biol 3(7):e137. https://doi.org/10.1371/journal.pcbi.0030137

GOrilla (2019) GOrilla—a tool for identifying enriched GO terms. http://cbl-gorilla.cs.technion.ac.il. Cited 26 Mar 2019

Harris BHL, Barberis A, West CML, Buffa FM (2015) Gene expression signatures as biomarkers of tumour hypoxia. Clin Oncol 27(10):547–560. https://doi.org/10.1016/j.clon.2015.07.004

Hoeijmakers JHJ (2009) DNA damage, aging, and cancer. N Engl J Med 361(15):1475–1485. https://doi.org/10.1056/NEJMra0804615

Huang DW, Sherman BT, Lempicki RA (2009a) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res 37(1):1–13. https://doi.org/10.1093/nar/gkn923

Huang DW, Sherman BT, Lempicki RA (2009b) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc 4(1):44–57. https://doi.org/10.1038/nprot.2008.211

Johnson DS, Mortazavi A, Myers RM, Wold B (2007) Genome-wide mapping of in Vivo protein-DNA interactions. Science (80-)316(5830):1497–1502. https://doi.org/10.1126/science.1141319

Kapitonov VV, Jurka J (2008) A universal classification of eukaryotic transposable elements implemented in Repbase. Nat Rev Genet 9(5):411–412; author reply 414. https://doi.org/10.1038/nrg2165-c1

Kato T, Iwamoto K (2014) Comprehensive DNA methylation and hydroxymethylation analysis in the human brain and its implication in mental disorders. Neuropharmacology 80:133–139. https://doi.org/10.1016/j.neuropharm.2013.12.019

Kazazian HH Jr, Moran JV (2017) Mobile DNA in health and disease. N Engl J Med 377(4):361. https://doi.org/10.1056/NEJMRA1510092

KEGG (2019) Available online: http://www.genome.jp/kegg/. Cited 26 Mar 2019

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J (2001) International human genome sequencing consortium. Initial sequencing and analysis of the human genome. Nature 409(6822):860–921. https://doi.org/10.1038/35057062

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(1626):20120507. https://doi.org/10.1098/rstb.2012.0507

Lynch M, Ackerman MS, Gout JF, Long H, Sung W, Thomas WK, Foster PL (2016) Genetic drift, selection and the evolution of the mutation rate. Nat Rev Genet 17(11):704–714. https://doi.org/10.1038/nrg.2016.104

Maleszka R, Mason PH, Barron AB (2014) Epigenomics and the concept of degeneracy in biological systems. Brief Funct Genomics 13(3):191–202. https://doi.org/10.1093/bfgp/elt050

Meier K, Brehm A (2014) Chromatin regulation: how complex does it get? Epigenetics 9(11):1485–1495. https://doi.org/10.4161/15592294.2014.971580

Mundade R, Ozer HG, Wei H, Prabhu L, Lu T (2014) Role of ChIP-seq in the discovery of transcription factor binding sites, differential gene regulation mechanism, epigenetic marks and beyond. Cell Cycle 13(18):2847–2852. https://doi.org/10.4161/15384101.2014.949201

National Cancer Institute (2019) Available online: https://cactus.nci.nih.gov/ncicadd/about.htm. Cited 26 Mar 2019

Nikitin D, Garazha A, Sorokin M, Penzar D, Tkachev V, Markov A, Buzdin A (2019) Retroelement-linked transcription factor binding patterns point to quickly developing molecular pathways in human evolution. Cells 8(2):130. https://doi.org/10.3390/cells8020130

Nikitin D, Penzar D, Garazha A, Sorokin M, Tkachev V, Borisov N, Buzdin AA (2018) Profiling of human molecular pathways affected by retrotransposons at the level of regulation by transcription factor proteins. Front Immunol 9:30. https://doi.org/10.3389/fimmu.2018.00030

Numpy Least squares polynomial fit (2019) Available online: https://docs.scipy.org/doc/numpy-1.15.0/reference/generated/numpy.polyfit.html. Cited 26 Mar 2019

O'Brien PJ (2006) Catalytic promiscuity and the divergent evolution of DNA repair enzymes. Chem Rev 106(2):720–752. https://doi.org/10.1021/cr040481v

Pathway Central (2019) Available online: http://www.sabiosciences.com/pathwaycentral.php. Cited 26 Mar 2019

Reactome (2019) Available online: http://reactome.org. Cited 26 Mar 2019

RepeatMasker (2019) Available online: http://www.repeatmasker.org. Cited 26 Mar 2019

Royer-Bertrand B, Rivolta C (2015) Whole genome sequencing as a means to assess pathogenic mutations in medical genetics and cancer. Cell Mol Life Sci 72(8):1463–1471. https://doi.org/10.1007/s00018-014-1807-9

Seaborn (2019) Available online: http://seaborn.pydata.org/. Cited 26 Mar 2019

Sloan CA, Chan ET, Davidson JM, Malladi VS, Strattan JS, Hitz BC, Cherry JM (2016) ENCODE data at the ENCODE portal. Nucleic Acids Res 44(D1):D726–D732. https://doi.org/10.1093/nar/gkv1160

Suntsova M, Garazha A, Ivanova A, Kaminsky D, Zhavoronkov A, Buzdin A (2015) Molecular functions of human endogenous retroviruses in health and disease. Cell Mol Life Sci 72(19):3653–3675. https://doi.org/10.1007/s00018-015-1947-6

The Gene Ontology Consortium (2017) Expansion of the gene ontology knowledgebase and resources. Nucleic Acids Res 45(D1):D331–D338. https://doi.org/10.1093/nar/gkw1108

Thompson D, Regev A, Roy S (2015) Comparative analysis of gene regulatory networks: from network reconstruction to evolution. Annu Rev Cell Dev Biol 31(1):399–428. https://doi.org/10.1146/annurev-cellbio-100913-012908

Turner BM (2014) Nucleosome signalling; an evolving concept. Biochim Biophys Acta 1839(8):623–626. https://doi.org/10.1016/j.bbagrm.2014.01.001

UCSC Browser, bedGraph files (2019a) Available online: https://genome.ucsc.edu/goldenpath/help/bedgraph.html. Cited 26 Mar 2019

UCSC Browser, Human genome (2019b) Available online: https://genome.ucsc.edu/cgi-bin/hgs. Cited 26 Mar 2019

Varriale A (2014) DNA methylation, epigenetics, and evolution in vertebrates: facts and challenges. Int J Evol Biol 2014:475981. https://doi.org/10.1155/2014/475981

Villar D, Flicek P, Odom DT (2014) Evolution of transcription factor binding in metazoans—mechanisms and functional implications. Nat Rev Genet 15(4):221–233. https://doi.org/10.1038/nrg3481

Yin H, Wang S, Zhang Y-H, Cai Y-D, Liu H (2016) Analysis of important gene ontology terms and biological pathways related to pancreatic cancer. Biomed Res Int 2016:1–10. https://doi.org/10.1155/2016/7861274

Yuryev A (2015) Gene expression profiling for targeted cancer treatment. Expert Opin Drug Discov 10(1):91–99. https://doi.org/10.1517/17460441.2015.971007

Zhong X (2016) Comparative epigenomics: a powerful tool to understand the evolution of DNA methylation. New Phytol 210(1):76–80. https://doi.org/10.1111/nph.13540

Chapter 6
Methods to Detect and Associate Divergence in Cis-Regulatory Elements to Phenotypic Divergence

Juliana G. Roscito and Michael Hiller

Abstract Understanding which genomic changes are responsible for morphological differences between species is a long-standing question in biology. While evolutionary theory predicts that morphology largely evolves by changing expression of important developmental genes, finding the underlying regulatory mutations is inherently difficult. Here, we discuss how the integration of comparative and functional genomics has provided valuable insights into the regulatory changes involved in morphological changes. By comparing genomes of species exhibiting differences in a morphological trait, comparative genomic methods enable the systematic detection of regulatory elements with divergence in sequence or transcription factor binding sites. To narrow this set of diverged elements down to those that likely contribute to differences in the trait of interest, one can leverage knowledge about gene function to assess which elements are associated with genes known to control the development of this trait. In addition, functional genomics can further prioritize diverged genomic regions based on overlap with experimentally determined regulatory elements that are active in tissues relevant for the trait. Further experiments can then evaluate whether sequence or binding site divergence translates into regulatory differences and affects the development of the trait. Thus, combining comparative and functional genomic approaches provide a widely applicable strategy to reveal regulatory changes contributing to morphological differences, which will enhance our understanding of how nature's spectacular phenotypic diversity evolved.

J. G. Roscito · M. Hiller (✉)
Max Planck Institute of Molecular Cell Biology and Genetics, Dresden, Germany
e-mail: hiller@mpi-cbg.de

J. G. Roscito
e-mail: roscito@mpi-cbg.de

Max Planck Institute for the Physics of Complex Systems, Dresden, Germany

Center for Systems Biology, Dresden, Germany

P. Pontarotti (ed.), *Evolution, Origin of Life, Concepts and Methods*,
https://doi.org/10.1007/978-3-030-30363-1_6

6.1 Introduction

The diversity of life forms is one of the most fascinating aspects of biology. The different shapes of the vertebrate limb, the breathtaking diversity of marine invertebrates, and many modifications of body plans in fish are only a few examples of the extent to which the evolutionary process shaped morphology. The question of how this diversity arose is a long-standing question in biology. With recent technological advances in genetics and genomics, we are now better equipped to investigate which changes in the genomes are responsible for morphological differences between species.

Morphology is established during development and involves genes controlling the growth and patterning of different parts of the body. Most genes involved in development are highly pleiotropic, that is, they regulate multiple developmental processes. These genes often code for transcription factors or other signaling proteins that play central roles in various gene regulatory networks that govern the formation of different body structures. Because these high-level regulators are involved in making several structures, they have to be expressed at many tissues and timepoints in the developing embryo.

The specificity in expression is given by cis-regulatory elements (CREs), which directly control the transcriptional output of the gene in a spatial and temporal manner (Davidson et al. 2003; Howard and Davidson 2004). CREs can be promoters, enhancers, silencers, and insulators. While promoters are located in close proximity to the gene's transcription start site, enhancers, silencers, and insulators can be located far from the gene they regulate and can be located upstream or downstream of the transcription start site. CREs are characterized by the presence of shorter sequences to which transacting DNA binding proteins, or transcription factors (TFs), bind to. The binding of specific combinations of transcription factors results in the activation or repression of transcription. Distal CREs such as enhancers and silencers are often modular and independently regulate gene expression in specific tissues or developmental timepoints (Davidson et al. 2003; Howard and Davidson 2004).

Given that morphology is established during development, evolutionary changes in morphology typically require changes affecting developmental genes. Conceptually, such changes can be due to mutations in the protein-coding sequence, leading to changes in protein function, or mutations in the associated regulatory elements that alter the expression pattern of these genes. Coding mutations in pleiotropic developmental genes would likely have widespread effects on development, which is often deleterious for the organism. In contrast, regulatory mutations in modular CREs can have smaller-scale effects, i.e., expression of its target gene would only be altered in a specific tissue or timepoint. The relative contribution of mutations in coding regions and CREs to the evolution of morphology has been the focus of a long-standing debate (Carroll 2008; Wittkopp et al. 2004; Wray 2007), and there are clear examples showing that both coding and regulatory mutations contribute to morphological differences (Stern and Orgogozo 2008; Burga et al. 2017; Sharma et al. 2018). However, the general difference in the degree of pleiotropy of developmental

genes and the modular CREs that control their expression predicts that morphology largely evolves by changes in the CREs, affecting gene expression patterns.

While there is strong support that many morphological changes are associated with spatial or temporal shifts in gene expression patterns, finding the underlying genomic changes in CREs is challenging. Many examples where morphological differences have been mechanistically linked to changes in CREs come from comparative studies between *Drosophila* species. Among those are a variety of abdominal and wing pigmentation patterns that are due to mutations in CREs regulating the *yellow* gene (Gompel et al. 2005; Prud'homme et al. 2006; Wittkopp et al. 2002), the loss of dorsal cuticular hairs in larvae of *Drosophila sechellia* that is a result of changes in CREs of the *shavenbaby* gene (Frankel et al. 2011; McGregor et al. 2007), and the evolution of additional sex combs in the legs and loss of part of the sensory bristles in the genitalia of *Drosophila santomea* that is due to a single nucleotide substitution in an enhancer of the *scute* gene (Nagy et al. 2018). Changes in CREs are also associated with vertebrate morphological evolution. One example is the diversification of artiodactyl digit patterns, which arose by changes in a CRE regulating the *sonic hedgehog* receptor gene *Ptch1* (Lopez-Rios et al. 2014). A second example is the loss of limbs in snakes, which is associated with deletions in a crucial enhancer that regulates the *sonic hedgehog* morphogen in the developing limbs. Snake-specific deletions of TF binding sites (TFBS) in this enhancer result in reduced enhancer activity, which provides an explanation for the reduced expression of this key limb patterning factor in snakes (Kvon et al. 2016; Leal and Cohn 2016). Complete CRE deletions can also change morphology. Loss of pelvic spines in freshwater stickleback fish species is due to recurrent deletions of a pelvic enhancer regulating *Pitx1*, which abolishes expression of this developmental transcription factor in the pelvic region and leads to the loss of pelvic spines (Chan et al. 2010). Finally, as detailed below, the evolution of several human-specific traits is linked to the loss of enhancers (McLean et al. 2011).

While these examples highlight the importance of CRE changes to the evolution of morphology, there are many other morphological changes for which the underlying genetic cause has not been identified yet, despite observations that expression differences in developmental genes are involved. Prominent examples include shifts in the expression domains of *Hox* genes, which pattern the anterior-posterior (AP) axis. In crustaceans, AP shifts in the expression domain of *ultrabithorax* result in the modification of locomotor into feeding appendages (Averof and Patel 1997). In snakes, an anterior expansion of the expression domain of *HoxB* and *HoxC* genes is associated with the homeotic transformation of the cervical region toward a thoracic identity, resulting in the absence of forelimbs (Cohn and Tickle 1999). There are several other examples of divergent limb morphologies within vertebrates that are associated with expression changes of key developmental factors. For example, spatial and temporal changes in expression of many limb patterning genes in the developing hindlimb buds of cetaceans are associated with an arrested development and regression of hindlimbs (Thewissen et al. 2006), *Shh* expression changes are associated with digit reductions in lizards (Shapiro et al. 2003), and changes in *Fgf8*, *Shh,* and *Ptch1* expression in bat wings are associated with digit lengthening and

the extension of interdigital webbing (Hockman et al. 2008). Finally, the variety of beak sizes and shapes in Darwin finches is linked to different expression patterns of the *Bmp4* gene (Abzhanov et al. 2004). While these morphological differences are associated with spatial and/or temporal expression shifts of relevant developmental genes, the genomic changes that underlie the expression changes of these genes remain largely unknown.

We are now witnessing a genomics revolution that can help to identify the genomic changes responsible for morphological changes. Technological and methodological advances are facilitating the acquisition of high-throughput data for an increasing number of species, allowing for large-scale comparative genomic analysis. Furthermore, experimental approaches using high-throughput sequencing enable an in-depth profiling of cell/tissue-specific regulatory landscapes. In this chapter, we discuss how integrating comparative and functional genomics can be a powerful approach to identify the genomic basis of morphological evolution. In the following, we first present a brief overview of computational and functional approaches for genome-wide identification of CREs. Then, we describe how comparative methods can use this data to quantify CRE divergence and to associate divergence patterns to morphological differences.

6.2 Methods to Identify CREs

Approaches to identify CRE candidates genome-wide can be mainly divided into computational and experimental methods. It is important to note that each approach has their strengths and intrinsic biases; therefore, combining different methods provide a more complete picture of the regulatory landscape. In-depth information on each of the methods can be found in a number of reviews (Wasserman and Sandelin 2004; Hardison and Taylor 2012; Kleftogiannis et al. 2016; Bell et al. 2011; Pennacchio and Rubin 2001; Li et al. 2015; Noonan and McCallion 2010; Shlyueva et al. 2014; Yip et al. 2013; Aerts 2012; Schmitt et al. 2016).

6.2.1 Computational Identification of CREs

1. Evolutionary conservation

Predicting CRE candidates based on evolutionary conservation relies on the fact that functional genomic sequences often evolve under purifying selection, that is, deleterious mutations affecting the function of such sequences negatively affect fitness and are less likely to be fixed in the population. As a result, functional genomic regions accumulate fewer mutations compared to other non-functional regions of the genome. This principle is exploited in *phylogenetic footprinting* approaches (Tagle et al. 1988) that compare genomes of many species to identify functional genomic

regions (Lindblad-Toh et al. 2011). While conserved genomic regions overlapping coding exons are likely constrained to preserve the encoded protein sequence, conserved sequences in non-coding parts of the genome, referred here as conserved non-coding elements (CNEs), can reliably reveal CREs (Hardison 2000; Pennacchio and Rubin 2001; Woolfe et al. 2005), in particular, when non-coding regions are conserved over large evolutionary timescales (e.g., from human to fish). Indeed, many experimental studies showed that a substantial portion of CNEs have regulatory activity and direct tissue-specific expression patterns (Frazer et al. 2004; Grice et al. 2005; Pennacchio et al. 2006; Visel et al. 2008). While this approach will capture neither recently evolved nor species-specific CREs, sequence conservation is a powerful indicator of CREs.

2. In silico identification of transcription factor binding site clusters

CREs are activated by the binding of TFs to sequence motifs that are generally short (on average 8 bp long) and often degenerate. The observation that CREs are typically bound by multiple TFs made it possible to develop methods that scan the genome for clusters of putative TFBS (Hughes et al. 2000; Kim et al. 2010; Rajewsky et al. 2002; Sinha et al. 2003). Motif-based approaches for predicting CREs are influenced by the set of TFs and a sizeable portion of the TFBS-based predicted CREs are false positives.

6.2.2 Experimental Identification of CREs

Genomic regions comprising active CREs exhibit characteristic physical and chemical signatures that can be detected experimentally. Such functional genomic methods provide a reliable way to identify CREs that are active in the selected tissue or cell type.

1. Open chromatin

Activation of CREs depends on the binding of transcription factors to the CRE sequence. TF binding competes with histone binding, which results in an increased accessibility of the local DNA region. Thus, active CREs can be identified by nucleosome-depleted open chromatin regions. Open chromatin can be targeted with nucleases such as DNaseI (Cockerill 2011; Gross and Garrard 1988; Wu et al. 1979). Another widely used method to profile open chromatin is the ATAC method (assay for transposase-accessible chromatin) (Buenrostro et al. 2013) that uses the Tn5 transposase to digest DNA in areas of reduced nucleosome occupancy. Nucleosome-depleted regions can also be identified as by-product of the detection of nucleosome-bound regions using the micrococcal nuclease MNase (Yuan et al. 2005). Fragmentation-based detection of open chromatin areas is the principle underlying FAIRE (formaldehyde-assisted isolation of regulatory elements) (Simon et al. 2013), which involves the crosslinking and mechanical shearing of DNA and subsequent depletion of those fragments with bound nucleosomes.

2. Chromatin immunoprecipitation (ChIP)

ChIP is a technique that relies on crosslinking chromatin in its native state and the subsequent antibody-based selection of proteins (TFs) bound to DNA, or of specific histone modifications that are hallmarks of active CREs. Given an antibody for the protein of interest, this method enables the direct detection of binding sites of TFs, transcriptional co-activators such as MED or p300, and proteins associated with transcription initiation such as Pol II. It also allows the detection of active and repressed promoters and distal CREs based on the characteristic post-translational histone modifications such as acetylation and methylation of the N-terminal histone tail (Heintzman et al. 2007).

3. Chromosome conformation capture

Chromosome conformation capture assays target the physical interaction between CREs and gene promoters. The detection of interacting regions in a chromosome is a powerful method to map active distal CREs, which are thought to control transcription of their target gene by physically interacting with target promoter regions (de Laat and Duboule 2013; Montavon and Duboule 2012; Noordermeer and Duboule 2013). There are currently a wide variety of chromosome capture techniques (Schmitt et al. 2016), which differ mainly with respect to resolution and the scale of detected interactions. The most commonly used technique for genome-wide profiling of chromosomal interactions is Hi-C.

4. Enhancer RNA (eRNA)

Due to advancements in high-throughput sequencing, it became clear that the pool of RNA in a cell also comprises RNA molecules derived from enhancers (eRNAs) (Andersson et al. 2014; Azofeifa et al. 2018; Lam et al. 2014). While it is not clear whether these short and unstable eRNAs have a function themselves, they are a valuable marker of active CREs. Indeed, eRNA levels are a quantitative readout of CRE activity and correlate with the mRNA levels of the gene they regulate. Thus, selecting nascent RNAs followed by sequencing is another powerful method to determine active CREs in the tissue of interest.

5. Experimental profiling of regulatory activity

The ability to drive expression of a reporter gene is another way to experimentally test whether a CRE candidate has regulatory activity. While such reporter gene assays were previously limited to few elements, high-throughput screening techniques enable now testing the regulatory potential of a large set of DNA fragments (Inoue and Ahituv 2015). High-throughput analysis of regulatory activity involves massive parallel cloning of tagged DNA fragments into reporter vectors and quantifying enhancer activity by deep sequencing. In the Starr-seq method (Arnold et al. 2013), DNA fragments are inserted into a vector downstream of a minimal promoter and transfected into cells; if the DNA fragment acts as an enhancer it will transcribe itself. Sequencing the pool of cells is then used to quantify the transcriptional products, resulting in a parallel readout of enhancer activity for millions of fragments.

6.2.3 *Integrating Computational and Functional Genomics Data for Comprehensive CRE Annotation*

Despite the significant advancements of all above-mentioned methods, allowing for accurate identification of enhancers, each method produces false-positive candidates, and no single method is sufficient to detect all active CREs. By definition, conservation-based or TFBS-based methods will only identify deeply conserved CREs or CREs that exhibit clusters of binding sites for the specified TFs. Likewise, functional genomics methods will only identify CREs active in the tissue or cell type and timepoint where the assay was performed. For these reasons, integrating evidences from multiple approaches can significantly reduce biases and false-positives coming from individual methods alone. A straightforward approach is to combine both computational and experimental predictions by selecting computationally predicted CREs that overlap one or more sets of experimentally discovered CREs. In case many experimental datasets are available, machine learning approaches can be used to model, annotate, and classify CREs genome-wide, which has been shown to increase sensitivity and accuracy in annotating CREs (Erwin et al. 2014; Kleftogiannis et al. 2015; Li et al. 2015; Monti et al. 2017; Narlikar et al. 2010; van Duijvenboden et al. 2016).

6.3 Detecting and Quantifying CRE Divergence

Identifying CREs is the first step in identifying the regulatory changes that are associated with morphological changes. However, narrowing the list down to those CRE candidates actually involved in phenotypic changes is challenging, since each of the above-discussed approaches typically yield hundreds, or even thousands, of CRE candidates. The search for the relevant CREs, i.e., those associated with the morphological change, requires a comparative approach that involves either finding newly evolved CREs specific to species that show the trait of interest, or finding ancestral CREs that changed or lost function in these species, which can be detected as sequence or TF binding site divergence.

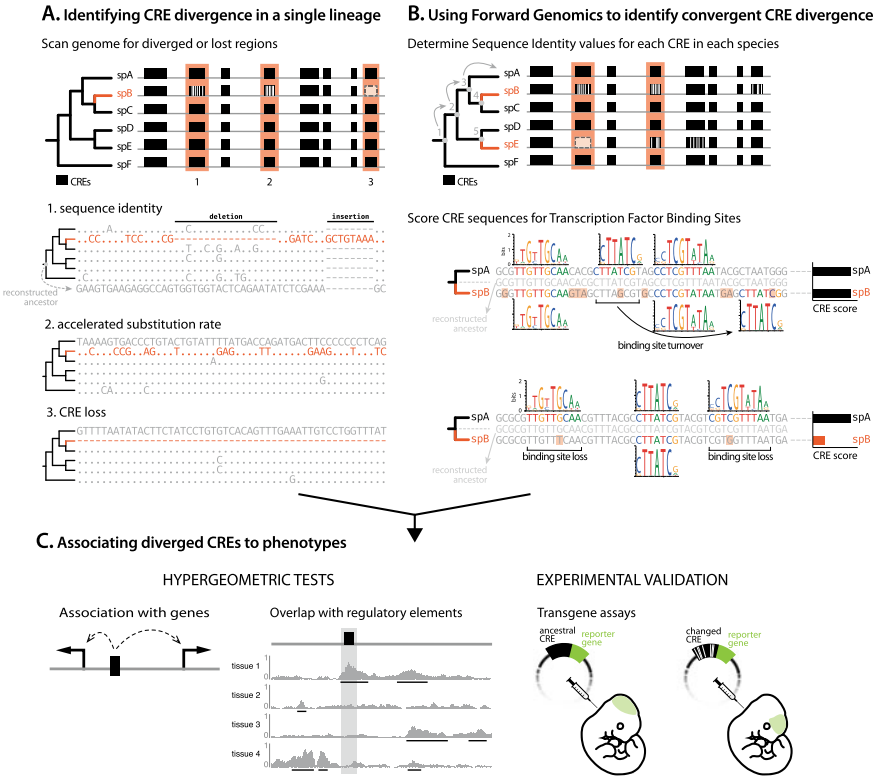
Changes in gene regulation can be due to the evolution of novel CREs in a particular species. Detection of such lineage-specific CREs with computational methods is challenging. Thus, obtaining experimental data for tissues relevant for the trait of interest is often necessary. Newly evolved CREs that may contribute to a newly evolved or changed trait can then be identified by comparing the regulatory landscape of species exhibiting the ancestral phenotype with that of species exhibiting the derived phenotype. However, evolution often also tinkers with existing functional elements. Thus, in addition to newly evolved CREs, changes or losses in conserved CREs can contribute to changes in gene expression and consequently changes in morphological structures. Hence, detecting and quantifying changes in ancestral CREs can highlight regulatory changes that are associated with morphological changes.

In the following, we will describe comparative genomic approaches to detect and quantify sequence and TFBS changes in ancestral CRE candidates whose sequences align across species, and methods to associate divergence patterns to differences in morphological traits. Figure 6.1 illustrates different types of sequence divergence described below, statistical tests to find characteristic enrichments of the identified CREs, and experimental tests to assess whether sequence divergence translates into regulatory changes.

6.3.1 Loss of CREs

The most radical change affecting an ancestral CRE is its complete loss, which can be computationally detected by lineage-specific absences of CNEs (Fig. 6.1a). Following the principle that evolutionary conservation should reflect function, deletion or absence of any aligning sequence in a particular lineage for an otherwise highly conserved CNE can occur when there is a relaxation of the selective pressures to maintain its function. This can happen as a result of different scenarios. First, a CRE can be lost when a new CRE arises at a different locus and functionally replaces the ancestral one, releasing the latter from selective constraint. Second, CREs can be lost following the loss of their target gene. Third, CRE loss can be associated with loss of a particular body structure, reflecting the loss of expression of the CRE's target gene that was once needed for the development of that structure. It should be noted that the absence of conserved sequence in a multiple genome alignment is not always due to CNE loss, as assembly incompleteness of alignment issues can mimic CNE loss. CNE losses should therefore be carefully validated (Hiller et al. 2012a).

CNE loss is not a rare event. A genome-wide screen showed that about 5% of mammalian CNEs are lost in at least one mammal species (Hiller et al. 2012a). More than 600 of those CNE losses are shared by more than one lineage and experiments showed that one of these independently lost CNEs functions as a spinal cord enhancer of the developmental *Gdf11* gene (Hiller et al. 2012a). CNE losses have been linked to lineage-specific morphological changes. For example, a genomic analysis showed that the seahorse has lost a substantially higher number of CNEs compared to other percomorph fish, and several of these CNE losses likely played a role in the evolution of the spectacular seahorse morphology (Lin et al. 2016). McLean et al. performed a screen to detect CNEs that are completely deleted in humans but present in chimpanzees, other primates and other mammals (McLean et al. 2011). Experimental tests of the selected CNEs with transgenic reporter assays in mouse (Fig. 6.1c) allowed them to associate the loss of an enhancer of the androgen receptor gene with the loss of sensory vibrissae and penile spines in humans. Furthermore, they found that humans have lost an enhancer regulating the tumor suppressor gene *Gadd45g*, which could be associated with the expansion of the neocortex in humans. Thus, complete loss of CNEs may affect gene expression and, consequently, morphological traits; however, experimental characterization of the consequences of the loss of a particular CNE are fundamental to reveal whether and which morphological traits are affected.


A. Identifying CRE divergence in a single lineage
Scan genome for diverged or lost regions
spA
spB
spC
spD
spE
spF
CREs
1
2
3
1. sequence identity
deletion
insertion
reconstructed ancestor
2. accelerated substitution rate
3. CRE loss
B. Using Forward Genomics to identify convergent CRE divergence
Determine Sequence Identity values for each CRE in each species
Score CRE sequences for Transcription Factor Binding Sites
reconstructed ancestor
binding site turnover
binding site loss
binding site loss
CRE score
CRE score
C. Associating diverged CREs to phenotypes
HYPERGEOMETRIC TESTS
Association with genes
Overlap with regulatory elements
tissue 1
tissue 2
tissue 3
tissue 4
EXPERIMENTAL VALIDATION
Transgene assays
ancestral CRE
reporter gene
changed CRE
reporter gene

◀**Fig. 6.1** Comparative genomic approaches to identify CRE divergence and methods to associate diverged CREs to functions and phenotypes. **a** To identify lineage-specific CRE divergence, illustrated here for species B (spB, orange), one can screen for CREs that are significantly diverged in this lineage but well conserved in the other species. The most general approach to identify sequence divergence takes the different types of mutations (substitutions, insertions, and deletions) into account and measures sequence divergence between extant species and their reconstructed ancestor (1). Alternatively, one can restrict the screen to identify CREs with accelerated substitution rates (2) or CREs that are completely lost or deleted (3). DNA bases that are identical to the reference are shown as dots in the alignments. **b** The Forward Genomics framework can identify associations between a convergently evolved phenotype and CRE divergence in the respective independent lineages, illustrated here for species B and E (orange). The two orange bars highlight CREs for which a genome-wide screen found a match between the CRE divergence pattern and the phenotypic pattern. These CREs exhibit a significantly higher divergence in species with the derived or lost phenotype (B, E) compared to the other species with the ancestral phenotype (A, C, D, F). The general Forward Genomics framework can consider either sequence or TFBS divergence. Both divergence measurements rely on reconstructing ancestral sequence states for each node in the phylogeny and using both extant and ancestral sequences to compute a branch-specific divergence measure by comparing the start and end node of each branch (illustrated by the gray arrows on the phylogeny). While sequence divergence reflects the amount of nucleotide changes (similar to A1), TFBS divergence reflects changes in binding sites independently of the number of mutations. The top alignment example illustrates similar TFBS scores of species A and B despite a high number of substitutions in species B that still retains all three binding sites because of motif degeneracy (substitutions that do not affect TF binding) and binding site turnover. In contrast, the bottom alignment illustrates that very few mutations can destroy binding sites, leading to a decreased TFBS score in species B. **c** Enrichment tests can be used to test whether diverged CREs are statistically associated with specific functions. Using non-diverged CREs as a control, hypergeometric tests can assess whether diverged CREs are significantly associated with genes belonging to certain functional groups (based on gene ontology terms, pathways, or knockout phenotypes) or whether diverged CRE candidates significantly overlap regulatory elements active in specific tissues. Finally, transgene assays can test whether CRE candidates have regulatory activity and whether CRE divergence translates into expression differences

6.3.2 CREs with Accelerated Substitution Rates

While complete loss of a CRE is the most radical type of change, changes in CRE function can also be detected by divergence in the CRE sequence. For example, a screen for CNEs that exhibit significantly elevated substitution rates in a particular lineage (Fig. 6.1a) can reveal elements evolving under positive selection that are associated with traits specific to this lineage.

This approach has been widely used (Cotney et al. 2013; Lindblad-Toh et al. 2011) and, together with follow-up experimental characterizations (Fig. 6.1c), revealed insights into regulatory changes involved in many traits. For example, a screen for CNEs with accelerated substitution rates in the placental mammal ancestor identified almost 5000 loci (Holloway et al. 2016). Subsequent transgenic reporter assays revealed that some of these accelerated CREs are associated with general mammalian morphological and physiological characteristics, such as hormonal regulation of milk production and uterine contractions, and development of the central nervous system

(in particular the visual system) (Holloway et al. 2016). Screens for CNEs with accelerated substitution rates in the human lineage revealed CREs regulating expression of genes involved in brain and neuronal development, including the RNA gene *Har1f* (Capra et al. 2013; Pollard et al. 2006a, b; Prabhakar et al. 2006). Experiments also revealed that a human-specific accelerated CRE drives stronger expression in the developing limbs than the chimpanzee or rhesus macaque sequence, which likely contributes to human-specific aspects of limb and digit morphology such as shortening of the toes and increase in thumb length (Prabhakar et al. 2008). Screens for accelerated CNEs also contributed to understanding traits in birds and bats. For example, vocal learning in independent bird lineages is linked to convergent acceleration of CNEs that likely regulate genes expressed in the song nuclei of the brain, such as the speech gene *FoxP1* (Zhang et al. 2014). The evolution of the bat wing is associated with bat-specific accelerated CREs regulating *HoxD* and other genes (Booker et al. 2016). Reporter assays showed that the sequence of these CREs often drives a novel expression pattern in the developing limbs compared to the orthologous mouse sequence (Booker et al. 2016).

In addition to detecting a gain or change of CRE function by identifying those CREs evolving under positive selection in specific lineages, accelerated substitution rates can also be used to detect otherwise conserved CREs that evolve under relaxed selection or neutrally in particular lineages. For example, several CNEs near *Pax6* and other transcription factors important for eye development exhibited accelerated substitution rates in subterranean mammals that have highly degenerated eyes (Partha et al. 2017). This acceleration was largely the result of neutral evolution, consistent with an association between the reduced visual system and relaxation of constraint on eye-regulatory elements (Partha et al. 2017).

6.3.3 Sequence Identity

By definition, accelerated substitution rates consider only substitutions and ignore insertions and deletions that can also affect CRE function. An exclusive focus on substitutions is a severe limitation when screening for neutrally evolving CREs as the clearest indication of loss of function is the complete deletion of a CRE. To take substitutions, insertions, and deletions into account when measuring sequence divergence, we previously developed an approach that is based on reconstructing the sequence of common ancestor of the species of interest (Hiller et al. 2012b; Prudent et al. 2016). This ancestral sequence is then aligned to sequences of extant species to determine the percent of identical bases (Fig. 6.1a). If the ancestral sequence is unchanged in an extant species, percent identify will be 100%. Complete divergence or a deletion of the ancestral sequence will result in percent identity of 0%. A framework termed "Forward Genomics" uses this sequence identity measure to perform a genome-wide screen to identify those conserved regions that exhibit elevated sequence divergence in those lineages where a particular trait was independently changed or lost (Fig. 6.1b) (Hiller et al. 2012b). For example, screening for genomic

regions that are preferentially diverged in mammals that are unable to synthesize vitamin C successfully identified genomic loci overlapping the vitamin C synthesizing *Gulo* gene (Hiller et al. 2012b).

To increase the power of the Forward Genomics framework, we developed the "branch method" to control for phylogenetic relatedness between species and differences in their evolutionary rates (Prudent et al. 2016). This method reconstructs the sequence of all ancestral states, represented as internal nodes in the phylogenetic tree (Fig. 6.1b), and computes branch-specific sequence divergence values by comparing the sequence at the start and end of each branch. These values are normalized by the length of the branch to control for differences in evolutionary rates. Since each branch represents independent evolution, branch-specific sequence divergence values circumvent the problem that the sequences of extant species are phylogenetically related (Prudent et al. 2016). As a consequence, this measure can accurately highlight specific branches of the tree along which the given sequence significantly accumulated mutations. Finally, to associate sequence divergence with a given trait, the branch method tests, separately for each CRE, whether the sequence identity values of branches along which the trait was changed or lost are lower than the values of branches along which the ancestral trait was preserved.

We applied this approach to perform the first genome-wide screen for CREs associated with the convergent eye degeneration in subterranean mammals (Roscito et al. 2018). This screen revealed more than 9000 CNEs with significantly increased sequence divergence on the branches leading to subterranean mammals. Using computational enrichment tests (Fig. 6.1c), we found that these CNEs are preferentially located near genes involved in eye development and eye function. To test whether these diverged CNEs correspond to eye-regulatory elements, we performed ATAC-seq in the developing mouse eye tissues and found that diverged CNEs significantly overlap eye CREs (Roscito et al. 2018) (Fig. 6.1c). Corroborating these results, diverged CNEs also significantly overlap functional genomics datasets obtained for adult mouse eye tissues. Together, our genome-wide analysis integrating functional and comparative genomics demonstrated that eye degeneration in subterranean mammals is associated with a widespread sequence divergence in the eye-regulatory landscape.

We used a similar ancestral reconstruction-based approach to investigate regulatory changes associated with the loss of limbs in snakes. Our genome-wide screen detected more than 5000 CNEs that exhibit significantly higher sequence divergence in snakes in comparison to other limbed species. Among those snake-diverged CNEs is the *Shh* limb enhancer that was experimentally shown to have significantly decreased its limb regulatory activity in snakes (Kvon et al. 2016; Leal and Cohn 2016). By combining ATAC-seq profiling and computational enrichment tests, we could show that snake-diverged CNEs are preferentially located near genes involved in limb development and significantly overlap CREs active in the developing limbs. This provides evidence that loss of limbs in the snake lineage is associated with genome-wide divergence of the limb regulatory landscape. Together with our results

on eye degeneration, these genome-wide screens suggest that the widespread divergence of the trait-specific regulatory landscape is a general evolutionary principle following the loss of complex morphological traits.

Another study devised a computational "Reverse Genomics" approach to associate sequence divergence to changes in morphological traits (Marcovitz et al. 2016). Instead of using ancestral sequence reconstruction, sequence divergence was identified based on a pairwise comparison between each mammalian species and human, which served as the reference. The study identified CNEs that were independently lost during placental mammalian evolution and used a large phenotypic character matrix to associate CRE loss patterns to trait change patterns. This screen uncovered numerous cases where independently lost CNEs are located near genes that are known to affect the development of the changed morphological structure. Striking examples include diverged CNEs associated with differences in pelvic or forelimb skeleton or differences in cochlea and brain morphology (Marcovitz et al. 2016).

Taken together, associating CNE sequence divergence in independent lineages with convergently evolved traits represents a powerful approach to reveal regulatory changes involved in morphological changes.

6.3.4 Scoring Gains and Losses of Transcription Factor Binding Sites

CREs are comprised of short-sequence motifs bound by TFs that determine regulatory activity. Gains, losses, and changes in TFBS, therefore, have strong potential to modify the regulatory activity of a CRE. Hence, investigating sequence instead of TFBS changes in a CRE may not always be informative about predicting functional CRE changes. For example, very few nucleotide changes (irrespective of whether they are substitutions or insertions/deletions) may affect key TFBS and thus change function, whereas numerous nucleotide changes may leave TFBS unaffected or result in TFBS turnover (creation of new equivalent TFBS and destruction of the original site, Fig. 6.1b) (Dermitzakis and Clark 2002; Huang et al. 2007; Otto et al. 2009; Villar et al. 2014). An analogy to that are nucleotide changes in protein-coding regions where few non-synonymous (amino-acid changing) differences can have a large impact on protein function, whereas numerous synonymous differences result in the same protein. Consequently, methods that take TFBS divergence into account can provide a more reliable way of identifying CRE changes that are more likely to affect regulatory activity and result in morphological changes across species.

A key requirement for identifying TFBS changes across species is predicting TFBS in a given CRE sequence. The binding motif of a TF can often be described by a position weight matrix (PWM), which is a representation of the per-base affinity of a TF to the respective binding site. Given that TFs of related species often have highly similar binding motifs (Boyle et al. 2014; Nitta et al. 2015), the same PWM can be used to predict TFBS in orthologous CRE sequences of multiple species. Applied

genome-wide, this information can highlight CREs with altered TFBS and can reveal statistical associations between TFBS differences and trait changes (Fig. 6.1c). For example, significant TFBS differences were observed in the promoter of genes associated with social behavior in bees and these differences correlated with differential expression of these genes in different bee castes (Kapheim et al. 2015; Sinha et al. 2006). Widespread gains and losses of limb-related TFBS in bat-accelerated CREs corroborate an association between changes in CRE function and the evolution of the bat wing (Booker et al. 2016). Another study demonstrated that genome-wide detection of TFBS losses combined with enrichment tests can reveal associations between TFBS losses and morphological differences (Berger et al. 2018). For example, TFBS preferentially lost in cetaceans and sirenians that lack hindlimbs are significantly associated with genes involved in hindlimb development, and TFBS preferentially lost in subterranean mammals are significantly associated with genes involved in eye development (Berger et al. 2018).

To pinpoint individual CREs where changes in clusters of TFBS are associated with convergent trait changes in independent lineages, we developed the REforge (regulatory element forward genomics) method (Langer et al. 2018). REforge is built on the Forward Genomics framework explained above, but measures TFBS divergence instead of sequence divergence. Given a set of TFs that may be relevant for the trait of interest (e.g., eye-related TFs could be relevant for eye degeneration), REforge performs a genome-wide screen to find CREs for which divergence of binding sites of these TFs is statistically associated with a given convergent trait change. In detail, REforge first reconstructs all ancestral sequence states of each CRE and uses Stubb (Sinha et al. 2003, 2006) to compute per-sequence TF binding scores. For a single (extant or ancestral) sequence, this score reflects the collective binding affinity of the input TF set. Then, for each branch in the phylogeny, REforge computes the difference between the scores for the sequences at the start and end of each branch, which reflects evolutionary gains and losses of TFBS along this branch (Fig. 6.1b). While the TFBS scores of sequences of extant species are not independent as species share evolutionary histories, the branch-specific scores are independent, enabling a direct comparison between branches. REforge then uses the Forward Genomics principle to test whether the given CRE exhibits a TFBS divergence pattern that is associated with the convergent change of the trait of interest. If this CRE is involved in the trait change, one would expect that the branches leading to species with the derived trait exhibit a stronger divergence of the TFBS ensemble than branches leading to lineages in which the ancestral trait is present. This can be directly tested by comparing the branch-specific scores of the two groups of branches.

We applied REforge to identify CREs that preferentially lost TFBS in subterranean mammals exhibiting convergent eye degeneration. Given a set of eye-related TFs, REforge identified thousands of CREs with a significant TFBS loss in subterranean mammals (Langer et al. 2018). These CREs are preferentially located near genes known to be involved in eye development and significantly overlap eye-specific regulatory elements identified in mouse by several experimental methods. A direct comparison to the sequence identity-based Forward Genomics approach showed that REforge has a substantially improved ability to detect functionally relevant CRE

divergence. Overall, this shows that convergent degeneration of the visual system in these species involved the loss of binding sites for many eye-relevant TFs in gene regulatory regions (Langer et al. 2018).

The input set of TFs influences the results of REforge. The choice of TFs relevant for the phenotype of interest can come from the literature, expression data of TFs (reasoning that relevant TFs should be expressed in the respective tissue), or TF knockout studies in model organisms, as we have done for analyzing eye degeneration. For cases where prior knowledge of relevant TFs is limited, we developed TFforge (transcription factor forward genomics) (Langer and Hiller 2019) that allows for the discovery of trait-involved TFs by large-scale divergences of their binding sites in species in which the trait has changed. As for REforge, TFforge relies on ancestral sequence reconstruction and scoring TFBS changes along individual branches. In contrast to REforge, TFforge jointly considers a large set of CNEs and screens all given TF motifs for those that exhibit a preferential and widespread binding site divergence in the CREs of species with the modified phenotype. Applying TFforge to genomic regions bound by the eye TFs CRX and NRL revealed that subterranean mammals have not only lost a significant number of binding sites for these two TFs but also other eye-related TFs that interact or co-bind with CRX and NRL (Langer et al. 2018). The joint application of TFforge and REforge has strong potential to reveal TFBS changes in CREs that could contribute to convergent changes in morphology.

6.4 Concluding Remarks

The identification of genomic changes associated with changes in morphological structures provides a foundation to understand how nature's phenotypic diversity evolved. While there is strong support that mutations in CRE sequences are an important source for morphological evolution, finding the regulatory mutations that contribute to morphological differences is difficult. Indeed, identifying CREs, detecting differences that likely affect CRE function and associating such differences to morphological change is not straightforward. Nevertheless, facilitated by technological and methodological advances, the combination of comparative and functional genomics approaches makes it possible to reveal regulatory differences associated with morphological traits. Specifically, a variety of comparative genomic approaches has been developed to detect differences in the sequence of CRE candidates such as accelerated substitution rates, increased sequence divergence or complete loss. Furthermore, recent methods are able to detect TFBS differences in CREs, which is a more powerful way of identifying changes likely affecting regulatory activity. Results from such genome-wide screens can be intersected with functional genomics data to reveal which of the diverged regulatory element candidates are active in a relevant tissue or timepoint, and thus may contribute to the trait of interest. To explore whether CRE divergence truly translates into regulatory differences, reporter assays

can subsequently test whether the CRE sequence of different species drives different expression patterns. Finally, genome engineering can be used to assess whether introducing the modified CRE in a model organism recapitulates trait changes seen in natural species, and thus establish a causal connection between differences in CREs and the trait. In summary, combining comparative and functional genomics provides a general and widely applicable strategy to reveal insights into the genomic basis of morphological differences.

References

Abzhanov A, Protas M, Grant BR, Grant PR, Tabin CJ (2004) Bmp4 and morphological variation of beaks in Darwin's finches. Science 305(5689):1462–1465. https://doi.org/10.1126/science.1098095

Aerts S (2012) Computational strategies for the genome-wide identification of cis-regulatory elements and transcriptional targets. Curr Top Dev Biol 98:121–145. https://doi.org/10.1016/B978-0-12-386499-4.00005-7

Andersson R, Gebhard C, Miguel-Escalada I, Hoof I, Bornholdt J, Boyd M, Chen Y, Zhao X, Schmidl C, Suzuki T, Ntini E, Arner E, Valen E, Li K, Schwarzfischer L, Glatz D, Raithel J, Lilje B, Rapin N, Bagger FO, Jorgensen M, Andersen PR, Bertin N, Rackham O, Burroughs AM, Baillie JK, Ishizu Y, Shimizu Y, Furuhata E, Maeda S, Negishi Y, Mungall CJ, Meehan TF, Lassmann T, Itoh M, Kawaji H, Kondo N, Kawai J, Lennartsson A, Daub CO, Heutink P, Hume DA, Jensen TH, Suzuki H, Hayashizaki Y, Muller F, Forrest ARR, Carninci P, Rehli M, Sandelin A (2014) An atlas of active enhancers across human cell types and tissues. Nature 507(7493):455–461. https://doi.org/10.1038/nature12787

Arnold CD, Gerlach D, Stelzer C, Boryn LM, Rath M, Stark A (2013) Genome-wide quantitative enhancer activity maps identified by STARR-seq. Science 339(6123):1074–1077. https://doi.org/10.1126/science.1232542

Averof M, Patel NH (1997) Crustacean appendage evolution associated with changes in Hox gene expression. Nature 388(6643):682–686. https://doi.org/10.1038/41786

Azofeifa JG, Allen MA, Hendrix JR, Read T, Rubin JD, Dowell RD (2018) Enhancer RNA profiling predicts transcription factor activity. Genome Res. https://doi.org/10.1101/gr.225755.117

Bell O, Tiwari VK, Thoma NH, Schubeler D (2011) Determinants and dynamics of genome accessibility. Nat Rev Genet 12(8):554–564. https://doi.org/10.1038/nrg3017

Berger MJ, Wenger AM, Guturu H, Bejerano G (2018) Independent erosion of conserved transcription factor binding sites points to shared hindlimb, vision and external testes loss in different mammals. Nucleic Acids Res 46(18):9299–9308. https://doi.org/10.1093/nar/gky741

Booker BM, Friedrich T, Mason MK, VanderMeer JE, Zhao J, Eckalbar WL, Logan M, Illing N, Pollard KS, Ahituv N (2016) Bat accelerated regions identify a bat forelimb specific enhancer in the HoxD locus. PLoS Genet 12(3):e1005738. https://doi.org/10.1371/journal.pgen.1005738

Boyle AP, Araya CL, Brdlik C, Cayting P, Cheng C, Cheng Y, Gardner K, Hillier LW, Janette J, Jiang L, Kasper D, Kawli T, Kheradpour P, Kundaje A, Li JJ, Ma L, Niu W, Rehm EJ, Rozowsky J, Slattery M, Spokony R, Terrell R, Vafeados D, Wang D, Weisdepp P, Wu YC, Xie D, Yan KK, Feingold EA, Good PJ, Pazin MJ, Huang H, Bickel PJ, Brenner SE, Reinke V, Waterston RH, Gerstein M, White KP, Kellis M, Snyder M (2014) Comparative analysis of regulatory information and circuits across distant species. Nature 512(7515):453–456. https://doi.org/10.1038/nature13668

Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ (2013) Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat Methods 10(12):1213–1218. https://doi.org/10.1038/nmeth.2688

Burga A, Wang W, Ben-David E, Wolf PC, Ramey AM, Verdugo C, Lyons K, Parker PG, Kruglyak L (2017) A genetic signature of the evolution of loss of flight in the Galapagos cormorant. Science 356(6341). https://doi.org/10.1126/science.aal3345

Capra JA, Erwin GD, McKinsey G, Rubenstein JL, Pollard KS (2013) Many human accelerated regions are developmental enhancers. Philos Trans R Soc Lond B Biol Sci 368(1632):20130025. https://doi.org/10.1098/rstb.2013.0025

Carroll SB (2008) Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution. Cell 134(1):25–36. https://doi.org/10.1016/j.cell.2008.06.030

Chan YF, Marks ME, Jones FC, Villarreal G Jr, Shapiro MD, Brady SD, Southwick AM, Absher DM, Grimwood J, Schmutz J, Myers RM, Petrov D, Jonsson B, Schluter D, Bell MA, Kingsley DM (2010) Adaptive evolution of pelvic reduction in sticklebacks by recurrent deletion of a Pitx1 enhancer. Science 327(5963):302–305. https://doi.org/10.1126/science.1182213

Cockerill PN (2011) Structure and function of active chromatin and DNase I hypersensitive sites. FEBS J 278(13):2182–2210. https://doi.org/10.1111/j.1742-4658.2011.08128.x

Cohn MJ, Tickle C (1999) Developmental basis of limblessness and axial patterning in snakes. Nature 399(6735):474–479. https://doi.org/10.1038/20944

Cotney J, Leng J, Yin J, Reilly SK, DeMare LE, Emera D, Ayoub AE, Rakic P, Noonan JP (2013) The evolution of lineage-specific regulatory activities in the human embryonic limb. Cell 154(1):185–196. https://doi.org/10.1016/j.cell.2013.05.056

Davidson EH, McClay DR, Hood L (2003) Regulatory gene networks and the properties of the developmental process. Proc Natl Acad Sci USA 100(4):1475–1480. https://doi.org/10.1073/pnas.0437746100

de Laat W, Duboule D (2013) Topology of mammalian developmental enhancers and their regulatory landscapes. Nature 502(7472):499–506. https://doi.org/10.1038/nature12753

Dermitzakis ET, Clark AG (2002) Evolution of transcription factor binding sites in Mammalian gene regulatory regions: conservation and turnover. Mol Biol Evol 19(7):1114–1121. https://doi.org/10.1093/oxfordjournals.molbev.a004169

Erwin GD, Oksenberg N, Truty RM, Kostka D, Murphy KK, Ahituv N, Pollard KS, Capra JA (2014) Integrating diverse datasets improves developmental enhancer prediction. PLoS Comput Biol 10(6):e1003677. https://doi.org/10.1371/journal.pcbi.1003677

Frankel N, Erezyilmaz DF, McGregor AP, Wang S, Payre F, Stern DL (2011) Morphological evolution caused by many subtle-effect substitutions in regulatory DNA. Nature 474(7353):598–603. https://doi.org/10.1038/nature10200

Frazer KA, Tao H, Osoegawa K, de Jong PJ, Chen X, Doherty MF, Cox DR (2004) Noncoding sequences conserved in a limited number of mammals in the SIM2 interval are frequently functional. Genome Res 14(3):367–372. https://doi.org/10.1101/gr.1961204

Gompel N, Prud'homme B, Wittkopp PJ, Kassner VA, Carroll SB (2005) Chance caught on the wing: cis-regulatory evolution and the origin of pigment patterns in Drosophila. Nature 433(7025):481–487. https://doi.org/10.1038/nature03235

Grice EA, Rochelle ES, Green ED, Chakravarti A, McCallion AS (2005) Evaluation of the RET regulatory landscape reveals the biological relevance of a HSCR-implicated enhancer. Hum Mol Genet 14(24):3837–3845. https://doi.org/10.1093/hmg/ddi408

Gross DS, Garrard WT (1988) Nuclease hypersensitive sites in chromatin. Annu Rev Biochem 57:159–197. https://doi.org/10.1146/annurev.bi.57.070188.001111

Hardison RC (2000) Conserved noncoding sequences are reliable guides to regulatory elements. Trends Genet 16(9):369–372

Hardison RC, Taylor J (2012) Genomic approaches towards finding cis-regulatory modules in animals. Nat Rev Genet 13(7):469–483. https://doi.org/10.1038/nrg3242

Heintzman ND, Stuart RK, Hon G, Fu Y, Ching CW, Hawkins RD, Barrera LO, Van Calcar S, Qu C, Ching KA, Wang W, Weng Z, Green RD, Crawford GE, Ren B (2007) Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. Nat Genet 39(3):311–318. https://doi.org/10.1038/ng1966

Hiller M, Schaar BT, Bejerano G (2012a) Hundreds of conserved non-coding genomic regions are independently lost in mammals. Nucleic Acids Res 40(22):11463–11476. https://doi.org/10.1093/nar/gks905

Hiller M, Schaar BT, Indjeian VB, Kingsley DM, Hagey LR, Bejerano G (2012b) A "forward genomics" approach links genotype to phenotype using independent phenotypic losses among related species. Cell Rep 2(4):817–823. https://doi.org/10.1016/j.celrep.2012.08.032

Hockman D, Cretekos CJ, Mason MK, Behringer RR, Jacobs DS, Illing N (2008) A second wave of Sonic hedgehog expression during the development of the bat limb. Proc Natl Acad Sci USA 105(44):16982–16987. https://doi.org/10.1073/pnas.0805308105

Holloway AK, Bruneau BG, Sukonnik T, Rubenstein JL, Pollard KS (2016) Accelerated evolution of enhancer hotspots in the mammal ancestor. Mol Biol Evol 33(4):1008–1018. https://doi.org/10.1093/molbev/msv344

Howard ML, Davidson EH (2004) Cis-Regulatory control circuits in development. Dev Biol 271(1):109–118. https://doi.org/10.1016/j.ydbio.2004.03.031

Huang W, Nevins JR, Ohler U (2007) Phylogenetic simulation of promoter evolution: estimation and modeling of binding site turnover events and assessment of their impact on alignment tools. Genome Biol 8(10):R225. https://doi.org/10.1186/gb-2007-8-10-r225

Hughes JD, Estep PW, Tavazoie S, Church GM (2000) Computational identification of cis-regulatory elements associated with groups of functionally related genes in *Saccharomyces cerevisiae*. J Mol Biol 296(5):1205–1214. https://doi.org/10.1006/jmbi.2000.3519

Inoue F, Ahituv N (2015) Decoding enhancers using massively parallel reporter assays. Genomics 106(3):159–164. https://doi.org/10.1016/j.ygeno.2015.06.005

Kapheim KM, Pan H, Li C, Salzberg SL, Puiu D, Magoc T, Robertson HM, Hudson ME, Venkat A, Fischman BJ, Hernandez A, Yandell M, Ence D, Holt C, Yocum GD, Kemp WP, Bosch J, Waterhouse RM, Zdobnov EM, Stolle E, Kraus FB, Helbing S, Moritz RF, Glastad KM, Hunt BG, Goodisman MA, Hauser F, Grimmelikhuijzen CJ, Pinheiro DG, Nunes FM, Soares MP, Tanaka ED, Simoes ZL, Hartfelder K, Evans JD, Barribeau SM, Johnson RM, Massey JH, Southey BR, Hasselmann M, Hamacher D, Biewer M, Kent CF, Zayed A, Blatti C, 3rd, Sinha S, Johnston JS, Hanrahan SJ, Kocher SD, Wang J, Robinson GE, Zhang G (2015) Social evolution. Genomic signatures of evolutionary transitions from solitary to group living. Science 348(6239):1139–1143. https://doi.org/10.1126/science.aaa4788

Kim J, Cunningham R, James B, Wyder S, Gibson JD, Niehuis O, Zdobnov EM, Robertson HM, Robinson GE, Werren JH, Sinha S (2010) Functional characterization of transcription factor motifs using cross-species comparison across large evolutionary distances. PLoS Comput Biol 6(1):e1000652. https://doi.org/10.1371/journal.pcbi.1000652

Kleftogiannis D, Kalnis P, Bajic VB (2015) DEEP: a general computational framework for predicting enhancers. Nucleic Acids Res 43(1):e6. https://doi.org/10.1093/nar/gku1058

Kleftogiannis D, Kalnis P, Bajic VB (2016) Progress and challenges in bioinformatics approaches for enhancer identification. Brief Bioinform 17(6):967–979. https://doi.org/10.1093/bib/bbv101

Kvon EZ, Kamneva OK, Melo US, Barozzi I, Osterwalder M, Mannion BJ, Tissieres V, Pickle CS, Plajzer-Frick I, Lee EA, Kato M, Garvin TH, Akiyama JA, Afzal V, Lopez-Rios J, Rubin EM, Dickel DE, Pennacchio LA, Visel A (2016) Progressive loss of function in a limb enhancer during snake evolution. Cell 167(3):633–642, e611. https://doi.org/10.1016/j.cell.2016.09.028

Lam MT, Li W, Rosenfeld MG, Glass CK (2014) Enhancer RNAs and regulated transcriptional programs. Trends Biochem Sci 39(4):170–182. https://doi.org/10.1016/j.tibs.2014.02.007

Langer BE, Hiller M (2019) TFforge utilizes large-scale binding site divergence to identify transcriptional regulators involved in phenotypic differences. Nucleic Acids Res 47(4):e19. https://doi.org/10.1093/nar/gky1200

Langer BE, Roscito JG, Hiller M (2018) REforge associates transcription factor binding site divergence in regulatory elements with phenotypic differences between species. Mol Biol Evol 35(12):3027–3040. https://doi.org/10.1093/molbev/msy187

Leal F, Cohn MJ (2016) Loss and re-emergence of legs in snakes by modular evolution of sonic hedgehog and HOXD enhancers. Curr Biol 26(21):2966–2973. https://doi.org/10.1016/j.cub.2016.09.020

Li Y, Chen CY, Kaye AM, Wasserman WW (2015) The identification of cis-regulatory elements: a review from a machine learning perspective. Biosystems 138:6–17. https://doi.org/10.1016/j.biosystems.2015.10.002

Lin Q, Fan S, Zhang Y, Xu M, Zhang H, Yang Y, Lee AP, Woltering JM, Ravi V, Gunter HM, Luo W, Gao Z, Lim ZW, Qin G, Schneider RF, Wang X, Xiong P, Li G, Wang K, Min J, Zhang C, Qiu Y, Bai J, He W, Bian C, Zhang X, Shan D, Qu H, Sun Y, Gao Q, Huang L, Shi Q, Meyer A, Venkatesh B (2016) The seahorse genome and the evolution of its specialized morphology. Nature 540(7633):395–399. https://doi.org/10.1038/nature20595

Lindblad-Toh K, Garber M, Zuk O, Lin MF, Parker BJ, Washietl S, Kheradpour P, Ernst J, Jordan G, Mauceli E, Ward LD, Lowe CB, Holloway AK, Clamp M, Gnerre S, Alfoldi J, Beal K, Chang J, Clawson H, Cuff J, Di Palma F, Fitzgerald S, Flicek P, Guttman M, Hubisz MJ, Jaffe DB, Jungreis I, Kent WJ, Kostka D, Lara M, Martins AL, Massingham T, Moltke I, Raney BJ, Rasmussen MD, Robinson J, Stark A, Vilella AJ, Wen J, Xie X, Zody MC, Broad Institute Sequencing P, Whole Genome Assembly T, Baldwin J, Bloom T, Chin CW, Heiman D, Nicol R, Nusbaum C, Young S, Wilkinson J, Worley KC, Kovar CL, Muzny DM, Gibbs RA, Baylor College of Medicine Human Genome Sequencing Center Sequencing T, Cree A, Dihn HH, Fowler G, Jhangiani S, Joshi V, Lee S, Lewis LR, Nazareth LV, Okwuonu G, Santibanez J, Warren WC, Mardis ER, Weinstock GM, Wilson RK, Genome Institute at Washington U, Delehaunty K, Dooling D, Fronik C, Fulton L, Fulton B, Graves T, Minx P, Sodergren E, Birney E, Margulies EH, Herrero J, Green ED, Haussler D, Siepel A, Goldman N, Pollard KS, Pedersen JS, Lander ES, Kellis M (2011) A high-resolution map of human evolutionary constraint using 29 mammals. Nature 478(7370):476–482. https://doi.org/10.1038/nature10530

Lopez-Rios J, Duchesne A, Speziale D, Andrey G, Peterson KA, Germann P, Unal E, Liu J, Floriot S, Barbey S, Gallard Y, Muller-Gerbl M, Courtney AD, Klopp C, Rodriguez S, Ivanek R, Beisel C, Wicking C, Iber D, Robert B, McMahon AP, Duboule D, Zeller R (2014) Attenuated sensing of SHH by Ptch1 underlies evolution of bovine limbs. Nature 511(7507):46–51. https://doi.org/10.1038/nature13289

Marcovitz A, Jia R, Bejerano G (2016) "Reverse genomics" predicts function of human conserved noncoding elements. Mol Biol Evol 33(5):1358–1369. https://doi.org/10.1093/molbev/msw001

McGregor AP, Orgogozo V, Delon I, Zanet J, Srinivasan DG, Payre F, Stern DL (2007) Morphological evolution through multiple cis-regulatory mutations at a single gene. Nature 448(7153):587–590. https://doi.org/10.1038/nature05988

McLean CY, Reno PL, Pollen AA, Bassan AI, Capellini TD, Guenther C, Indjeian VB, Lim X, Menke DB, Schaar BT, Wenger AM, Bejerano G, Kingsley DM (2011) Human-specific loss of regulatory DNA and the evolution of human-specific traits. Nature 471(7337):216–219. https://doi.org/10.1038/nature09774

Montavon T, Duboule D (2012) Landscapes and archipelagos: spatial organization of gene regulation in vertebrates. Trends Cell Biol 22(7):347–354. https://doi.org/10.1016/j.tcb.2012.04.003

Monti R, Barozzi I, Osterwalder M, Lee E, Kato M, Garvin TH, Plajzer-Frick I, Pickle CS, Akiyama JA, Afzal V, Beerenwinkel N, Dickel DE, Visel A, Pennacchio LA (2017) Limb-enhancer genie: an accessible resource of accurate enhancer predictions in the developing limb. PLoS Comput Biol 13(8):e1005720. https://doi.org/10.1371/journal.pcbi.1005720

Nagy O, Nuez I, Savisaar R, Peluffo AE, Yassin A, Lang M, Stern DL, Matute DR, David JR, Courtier-Orgogozo V (2018) Correlated evolution of two copulatory organs via a single cis-regulatory nucleotide change. Curr Biol 28(21):3450–3457, e3413. https://doi.org/10.1016/j.cub.2018.08.047

Narlikar L, Sakabe NJ, Blanski AA, Arimura FE, Westlund JM, Nobrega MA, Ovcharenko I (2010) Genome-wide discovery of human heart enhancers. Genome Res 20(3):381–392. https://doi.org/10.1101/gr.098657.109

Nitta KR, Jolma A, Yin Y, Morgunova E, Kivioja T, Akhtar J, Hens K, Toivonen J, Deplancke B, Furlong EE, Taipale J (2015) Conservation of transcription factor binding specificities across 600 million years of bilateria evolution. Elife 4. https://doi.org/10.7554/elife.04837

Noonan JP, McCallion AS (2010) Genomics of long-range regulatory elements. Annu Rev Genomics Hum Genet 11:1–23. https://doi.org/10.1146/annurev-genom-082509-141651

Noordermeer D, Duboule D (2013) Chromatin looping and organization at developmentally regulated gene loci. Wiley Interdiscip Rev Dev Biol 2(5):615–630. https://doi.org/10.1002/wdev.103

Otto W, Stadler PF, Lopez-Giraldez F, Townsend JP, Lynch VJ, Wagner GP (2009) Measuring transcription factor-binding site turnover: a maximum likelihood approach using phylogenies. Genome Biol Evol 1:85–98. https://doi.org/10.1093/gbe/evp010

Partha R, Chauhan BK, Ferreira Z, Robinson JD, Lathrop K, Nischal KK, Chikina M, Clark NL (2017) Subterranean mammals show convergent regression in ocular genes and enhancers, along with adaptation to tunneling. Elife 6. https://doi.org/10.7554/elife.25884

Pennacchio LA, Ahituv N, Moses AM, Prabhakar S, Nobrega MA, Shoukry M, Minovitsky S, Dubchak I, Holt A, Lewis KD, Plajzer-Frick I, Akiyama J, De Val S, Afzal V, Black BL, Couronne O, Eisen MB, Visel A, Rubin EM (2006) In vivo enhancer analysis of human conserved non-coding sequences. Nature 444(7118):499–502. https://doi.org/10.1038/nature05295

Pennacchio LA, Rubin EM (2001) Genomic strategies to identify mammalian regulatory sequences. Nat Rev Genet 2(2):100–109. https://doi.org/10.1038/35052548

Pollard KS, Salama SR, King B, Kern AD, Dreszer T, Katzman S, Siepel A, Pedersen JS, Bejerano G, Baertsch R, Rosenbloom KR, Kent J, Haussler D (2006a) Forces shaping the fastest evolving regions in the human genome. PLoS Genet 2(10):e168. https://doi.org/10.1371/journal.pgen.0020168

Pollard KS, Salama SR, Lambert N, Lambot MA, Coppens S, Pedersen JS, Katzman S, King B, Onodera C, Siepel A, Kern AD, Dehay C, Igel H, Ares M Jr, Vanderhaeghen P, Haussler D (2006b) An RNA gene expressed during cortical development evolved rapidly in humans. Nature 443(7108):167–172. https://doi.org/10.1038/nature05113

Prabhakar S, Noonan JP, Paabo S, Rubin EM (2006) Accelerated evolution of conserved noncoding sequences in humans. Science 314(5800):786. https://doi.org/10.1126/science.1130738

Prabhakar S, Visel A, Akiyama JA, Shoukry M, Lewis KD, Holt A, Plajzer-Frick I, Morrison H, Fitzpatrick DR, Afzal V, Pennacchio LA, Rubin EM, Noonan JP (2008) Human-specific gain of function in a developmental enhancer. Science 321(5894):1346–1350. https://doi.org/10.1126/science.1159974

Prud'homme B, Gompel N, Rokas A, Kassner VA, Williams TM, Yeh SD, True JR, Carroll SB (2006) Repeated morphological evolution through cis-regulatory changes in a pleiotropic gene. Nature 440(7087):1050–1053. https://doi.org/10.1038/nature04597

Prudent X, Parra G, Schwede P, Roscito JG, Hiller M (2016) Controlling for phylogenetic relatedness and evolutionary rates improves the discovery of associations between species' phenotypic and genomic differences. Mol Biol Evol 33(8):2135–2150. https://doi.org/10.1093/molbev/msw098

Rajewsky N, Vergassola M, Gaul U, Siggia ED (2002) Computational detection of genomic cis-regulatory modules applied to body patterning in the early Drosophila embryo. BMC Bioinf. 3:30

Roscito JG, Sameith K, Parra G, Langer BE, Petzold A, Moebius C, Bickle M, Rodrigues MT, Hiller M (2018) Phenotype loss is associated with widespread divergence of the gene regulatory landscape in evolution. Nat Commun 9(1):4737. https://doi.org/10.1038/s41467-018-07122-z

Schmitt AD, Hu M, Ren B (2016) Genome-wide mapping and analysis of chromosome architecture. Nat Rev Mol Cell Biol 17(12):743–755. https://doi.org/10.1038/nrm.2016.104

Shapiro MD, Hanken J, Rosenthal N (2003) Developmental basis of evolutionary digit loss in the Australian lizard Hemiergis. J Exp Zool B Mol Dev Evol 297(1):48–56

Sharma V, Lehmann T, Stuckas H, Funke L, Hiller M (2018) Loss of RXFP2 and INSL3 genes in Afrotheria shows that testicular descent is the ancestral condition in placental mammals. PLoS Biol 16(6):e2005293. https://doi.org/10.1371/journal.pbio.2005293

Shlyueva D, Stampfel G, Stark A (2014) Transcriptional enhancers: from properties to genome-wide predictions. Nat Rev Genet 15(4):272–286. https://doi.org/10.1038/nrg3682

Simon JM, Giresi PG, Davis IJ, Lieb JD (2013) A detailed protocol for formaldehyde-assisted isolation of regulatory elements (FAIRE). Curr Protoc Mol Biol Chap 21:Unit 21 26. https://doi.org/10.1002/0471142727.mb2126s102

Sinha S, Ling X, Whitfield CW, Zhai C, Robinson GE (2006) Genome scan for cis-regulatory DNA motifs associated with social behavior in honey bees. Proc Natl Acad Sci USA 103(44):16352–16357. https://doi.org/10.1073/pnas.0607448103

Sinha S, van Nimwegen E, Siggia ED (2003) A probabilistic method to detect regulatory modules. Bioinformatics 19(Suppl 1):i292–i301

Stern DL, Orgogozo V (2008) The loci of evolution: how predictable is genetic evolution? Evolution 62(9):2155–2177. https://doi.org/10.1111/j.1558-5646.2008.00450.x

Tagle DA, Koop BF, Goodman M, Slightom JL, Hess DL, Jones RT (1988) Embryonic epsilon and gamma globin genes of a prosimian primate (*Galago crassicaudatus*). Nucleotide and amino acid sequences, developmental regulation and phylogenetic footprints. J Mol Biol 203 (2):439–455

Thewissen JG, Cohn MJ, Stevens LS, Bajpai S, Heyning J, Horton WE Jr (2006) Developmental basis for hind-limb loss in dolphins and origin of the cetacean bodyplan. Proc Natl Acad Sci USA 103(22):8414–8418. https://doi.org/10.1073/pnas.0602920103

van Duijvenboden K, de Boer BA, Capon N, Ruijter JM, Christoffels VM (2016) EMERGE: a flexible modelling framework to predict genomic regulatory elements from genomic signatures. Nucleic Acids Res 44(5):e42. https://doi.org/10.1093/nar/gkv1144

Villar D, Flicek P, Odom DT (2014) Evolution of transcription factor binding in metazoans—mechanisms and functional implications. Nat Rev Genet 15(4):221–233. https://doi.org/10.1038/nrg3481

Visel A, Prabhakar S, Akiyama JA, Shoukry M, Lewis KD, Holt A, Plajzer-Frick I, Afzal V, Rubin EM, Pennacchio LA (2008) Ultraconservation identifies a small subset of extremely constrained developmental enhancers. Nat Genet 40(2):158–160. https://doi.org/10.1038/ng.2007.55

Wasserman WW, Sandelin A (2004) Applied bioinformatics for the identification of regulatory elements. Nat Rev Genet 5(4):276–287. https://doi.org/10.1038/nrg1315

Wittkopp PJ, Haerum BK, Clark AG (2004) Evolutionary changes in cis and trans gene regulation. Nature 430(6995):85–88. https://doi.org/10.1038/nature02698

Wittkopp PJ, Vaccaro K, Carroll SB (2002) Evolution of yellow gene regulation and pigmentation in Drosophila. Curr Biol 12(18):1547–1556

Woolfe A, Goodson M, Goode DK, Snell P, McEwen GK, Vavouri T, Smith SF, North P, Callaway H, Kelly K, Walter K, Abnizova I, Gilks W, Edwards YJ, Cooke JE, Elgar G (2005) Highly conserved non-coding sequences are associated with vertebrate development. PLoS Biol 3(1):e7. https://doi.org/10.1371/journal.pbio.0030007

Wray GA (2007) The evolutionary significance of cis-regulatory mutations. Nat Rev Genet 8(3):206–216. https://doi.org/10.1038/nrg2063

Wu C, Bingham PM, Livak KJ, Holmgren R, Elgin SC (1979) The chromatin structure of specific genes: I. Evidence for higher order domains of defined DNA sequence. Cell 16(4):797–806

Yip KY, Cheng C, Gerstein M (2013) Machine learning and genome annotation: a match meant to be? Genome Biol 14(5):205. https://doi.org/10.1186/gb-2013-14-5-205

Yuan GC, Liu YJ, Dion MF, Slack MD, Wu LF, Altschuler SJ, Rando OJ (2005) Genome-scale identification of nucleosome positions in *S. cerevisiae*. Science 309(5734):626–630. https://doi.org/10.1126/science.1112178

Zhang G, Li C, Li Q, Li B, Larkin DM, Lee C, Storz JF, Antunes A, Greenwold MJ, Meredith RW, Odeen A, Cui J, Zhou Q, Xu L, Pan H, Wang Z, Jin L, Zhang P, Hu H, Yang W, Hu J, Xiao J, Yang Z, Liu Y, Xie Q, Yu H, Lian J, Wen P, Zhang F, Li H, Zeng Y, Xiong Z, Liu S, Zhou L, Huang Z, An N, Wang J, Zheng Q, Xiong Y, Wang G, Wang B, Wang J, Fan Y, da Fonseca RR, Alfaro-Nunez A, Schubert M, Orlando L, Mourier T, Howard JT, Ganapathy G, Pfenning A, Whitney O, Rivas MV, Hara E, Smith J, Farre M, Narayan J, Slavov G, Romanov MN, Borges R, Machado JP, Khan I, Springer MS, Gatesy J, Hoffmann FG, Opazo JC, Hastad O, Sawyer RH, Kim H, Kim KW, Kim HJ, Cho S, Li N, Huang Y, Bruford MW, Zhan X, Dixon A, Bertelsen MF, Derryberry E, Warren W, Wilson RK, Li S, Ray DA, Green RE, O'Brien SJ, Griffin D, Johnson WE, Haussler D, Ryder OA, Willerslev E, Graves GR, Alstrom P, Fjeldsa J, Mindell DP, Edwards SV, Braun EL, Rahbek C, Burt DW, Houde P, Zhang Y, Yang H, Wang J, Avian Genome C, Jarvis ED, Gilbert MT, Wang J (2014) Comparative genomics reveals insights into avian genome evolution and adaptation. Science 346(6215):1311–1320. https://doi.org/10.1126/science.1251385

Chapter 7
A Look Back Over 20 Years of Evo-Devo Studies on Sponges: A Challenged View of Urmetazoa

Quentin Schenkelaars, Amélie Vernale, Laura Fierro-Constaín, Carole Borchiellini and Emmanuelle Renard

Abstract The majority of extant metazoans are animals with bilateral symmetry, a gut, a central nervous system and a head; they are grouped in a taxon named Bilateria. As a consequence, most of the zoological definitions are based on observations in bilaterian animals. To understand how these anatomical features emerged, it is now necessary for the Evo-Devo field to undertake studies on the four extant lineages that emerged earlier, namely Cnidaria, Placozoa, Ctenophora and Porifera. Being the sister group of bilaterians, cnidarians have received much attention compared to the other three phyla but, given the disparity of body plans between these four non-bilaterian phyla, comparative analyses are needed not only to trace back the origin and evolution of genetic mechanisms involved in animal development but also to understand the evolutionary processes that gave rise to such a huge animal body plan diversity. This chapter focuses on the evo-devo approaches applied to Porifera and how the data have changed our view of these uncommon animals and how it challenges previous concepts of the last common metazoan ancestor. We discuss here the current and future steps that need to be undertaken to ensure that sponges join the laboratory "model organisms club."

7.1 Introduction

The majority of extant metazoans (i.e. animals) belong to bilaterians (Bilateria). Bilateria is a taxon of animals showing a left-right symmetry and often possessing a gut, a central nervous system and head (Gilbert 2010). While bilaterians have massively

Q. Schenkelaars · A. Vernale · L. Fierro-Constaín · C. Borchiellini · E. Renard (✉)
Aix Marseille Univ, Avignon Université, CNRS, IRD, IMBE, Marseille, France, Station Marine d'Endoume, Rue de la Batterie des Lions, 13007 Marseille, France
e-mail: Emmanuelle.renard@univ-amu.fr

Q. Schenkelaars
Department of Genetics and Evolution, Faculty of Sciences, Institute of Genetics and Genomics in Geneva (IGe3), University of Geneva, Geneva, Switzerland

A. Vernale · E. Renard
Aix Marseille Univ, CNRS, IBDM, UMR 7288, Marseille, France

P. Pontarotti (ed.), *Evolution, Origin of Life, Concepts and Methods*,
https://doi.org/10.1007/978-3-030-30363-1_7

conquered all marine, freshwater and terrestrial ecosystems, non-bilaterians namely cnidarians (Phylum Cnidaria), placozoans (Phylum Placozoa), ctenophores (Phylum Ctenophora) and poriferans (Phylum Porifera) are mainly restricted to seawater. Therefore, the latter phyla often remain less widely known. As a direct consequence, most of the concepts that have been developed in zoology and developmental biology are based on the study of a few bilaterian models. However, to understand how the morpho-anatomical features of animals emerged, comparative studies embracing a larger panel of animal phyla are needed. The "Evo-Devo" field (a field of biology aiming to cross evolution and development) aims to address these issues is already in progress (Jenner and Wills 2007; Lanna 2015). While the Gastrea theory (Haeckel 1874) proposed that all animals originated from a single hypothetical ancestor resembling the gastrula stage (found in the development of all metazoans including sponges) (Eerkes-Medrano and Leys 2006; Reynolds 2019), the idea of a single origin of animal multicellularity only became recognized at the end of the 1990s [e.g., (Whittaker 1959; Brusca and Brusca 2003; Kotpal 2012)] (Fig. 7.1).

Accordingly, trying to imagine the features of the last common ancestor of extant animals, often named the "Urmetazoa" or the "Archimetazoa," is a long-standing

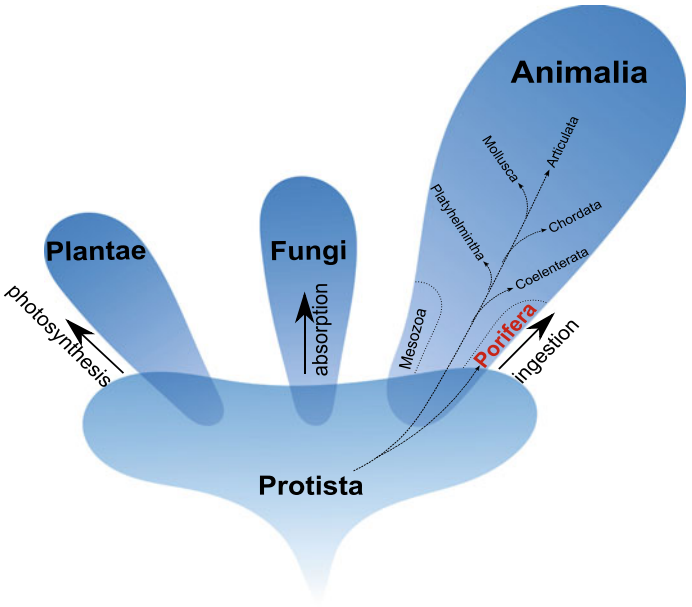


Fig. 7.1 Simplified schema of Whittaker's evolutionary scenario. Due to the huge diversity of animal phyla and their respective body plans, the idea of a single origin of animal multicellularity was far from being obvious until the end of the 1990's. In this famous representation of eukaryote relationships (Whittaker 1959), Mesozoa (e.g., Dyciemida and Orthonectida) and Porifera would have emerged from different unicellular ancestors compared to other animals (e.g., mollusks, flatworms, annelids, arthropods or chordates)

quest for zoologists. The present chapter traces back and discusses the main breakthroughs achieved by the spongiologist evo-devo community during the twenty-first century in order to understand one of the major evolutionary transitions, namely the origin of animal multicellularity (Sebé-Pedrós et al. 2017).

7.2 Sponges Appear to Be "Simple" Animals

Sponge diversity largely surpasses the famous bathroom sponge since more than 9000 species have been described to date (see World Porifera Database (http://www.marinespecies.org/porifera/ for an updated source of sponge diversity). Recent analyses have challenged the previous view of poriferan relationships, and it is now widely accepted that poriferans represent a monophyletic group that is split into four different lineages (i.e., Calcarea, Demospongiae, Hexactinellida and Homoscleromorpha), each with different characteristics (summarized in Fig. 7.2) (Ereskovsky 2010; Gazave et al. 2013; Leys and Hill 2012; Philippe et al. 2009; Renard et al. 2013, 2018).

Nevertheless, some common features have been established. Sponges are mainly marine animals. They do not possess any proper digestive tract but mainly behave as sessile filter feeders thanks to a net of canals, chambers and openings (i.e., ostia and osculum), which together form the aquiferous system (Fig. 7.3). Interestingly, sponge symmetry is often not as clear as in bilaterians or cnidarians. Instead, a sponge body plan relies on the organization of the aquiferous system and often on a skeleton consisting of an assembly of tiny mineral pieces named spicules. This leads either to radial symmetry and a clear apico-basal orientation as seen in the *Sycon* genus (class Calcarea) or to a more anarchic shape (asymmetry) in which the apico-basal polarity is less obvious as observed in the *Oscarella* genus, class Homoscleromorpha.

Irrespective of their body plans, sponges have two cellular layers (except for the so-called glass sponges corresponding to the class Hexactinellida that present an uncommon syncytial organization). On the one hand, there is a simple squamous epithelium named the pinacoderm which covers both the surface of the animal (the exopinacoderm composed of exopinacocytes) and the canals of the aquiferous system (the endopinacoderm composed of endopinacocytes). On the other hand, the choanocyte chambers (i.e., the filtering units composed of choanocytes) are defined as the choanoderm. Between these two tissues, there is a mesenchymal-like connective tissue named mesohyle that is composed of extracellular matrix and different cell types (about 2–12 cell types depending on the species considered). Among these mesohylar cells, some pluripotent stem cells (the archeocytes) and secretory cells (for example, vacuolar cells) have been described, but there is no evidence of nerve cells.

Although sponges are devoid of muscles and nerve cells, they are able to react to various stimuli and to contract thanks to the presence of sensory cilia, neurotransmitters or action potentials (Dunn et al. 2015; Elliott and Leys 2010; Ellwanger et al. 2007; Johnson et al. 2017; Leys 2015; Ludeman et al. 2014; Mah and Leys 2017;

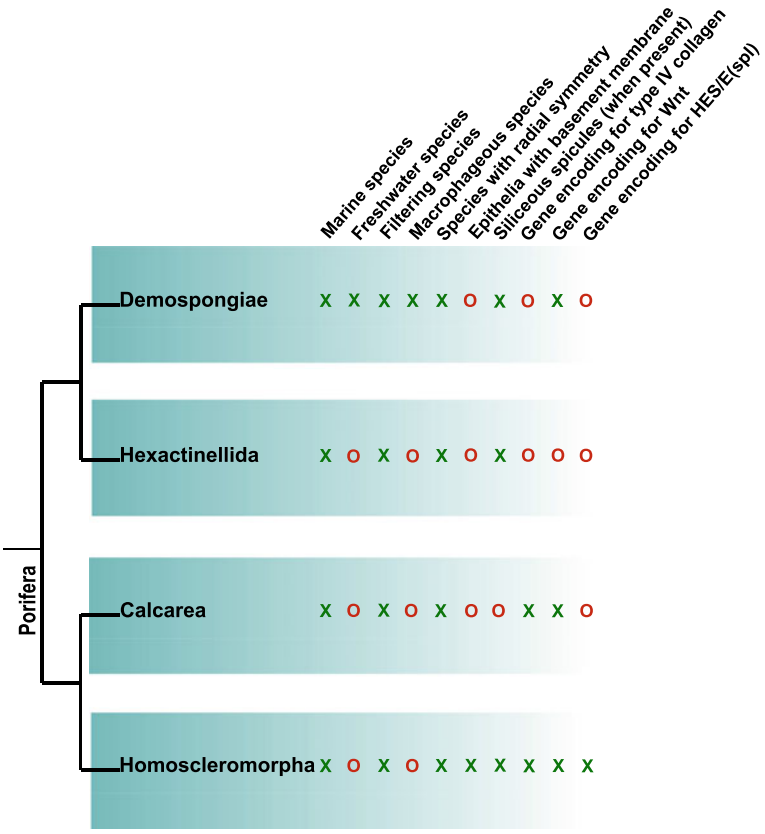


Fig. 7.2 According to the most recent molecular phylogenetic and phylogenomic studies, sponges form a monophyletic group (phylum Porifera) that is split into four lineages or classes (Gazave et al. 2013), each of which harbors distinct ecological, morphological, developmental and molecular features. This schema provides the phylogenetic relationships among the four sponge classes together with a few distinct features [for more details see (Ereskovsky 2010; Renard et al. 2018)]

Nakanishi et al. 2015; Nickel 2010; Renard et al. 2009; Tompkins-Macdonald et al. 2009; Wells et al. 2012).

Three of the four poriferan classes lack a basement membrane composed of type IV collagen, an observation that was thought to indicate that sponges were devoid of epithelia (Boute et al. 1996). However, there is growing evidence that sponge cell layers have similar physiological properties (e.g., a mechanical barrier against environment and ion composition regulation) to cnidarian and bilaterian epithelia (Adams et al. 2010; Leys and Hill 2012; Leys and Riesgo 2012; Leys et al. 2009; Renard et al. in press.). Because the zoologist community mainly focused on the missing elements in sponges when compared to other animals, they were considered as "lower" and "primitive" animals and thus set aside from "true" metazoans for too

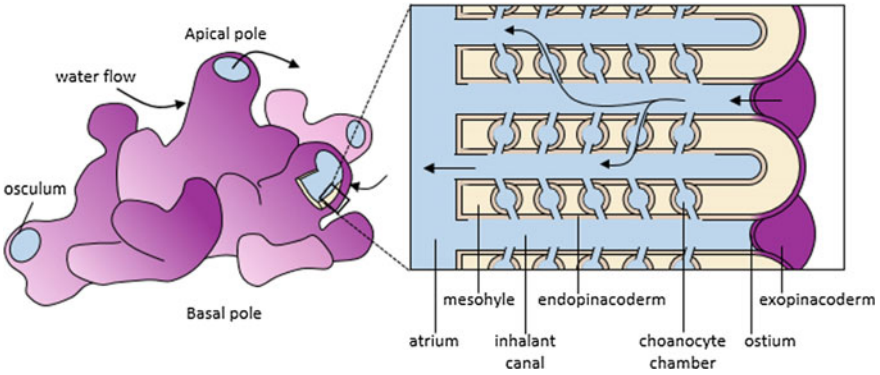


Fig. 7.3 At the adult stage, all sponges are sessile with a basal-apical pole (more or less obvious depending on the considered species). Most sponge species are filter feeders at the adult stage (except the anecdotal carnivorous Cladorhizidae family, Hestetun et al. 2017). The filtering activity relies on the presence of a network of canals and chambers named the aquiferous system. The body surface is usually covered by an epithelium named exopinacoderm, the inhalant and exhalant canals (including the large atrial cavity or atrium) are lined by the endopinacoderm. The pumping and filtering units of the aquiferous system are named choanocyte chambers and form the choanoderm. Between these epithelial layers, the mesohyle is composed of variable cell types, organic components (i.e., the extracellular matrix and collagen fibers) and the mineral elements of the skeleton (i.e., the spicules)

long (see next section). In other words, the apparent simplicity of sponge morpho-anatomy and the difficulty in establishing clear anatomical homologies between sponges and other animals, partially explain the gap in biological knowledge regarding these intriguing animals (Dunn et al. 2015).

7.3 Twenty Years of Uncertainty at the Base of the Animal Tree

Since the end of the 1990s, molecular phylogenetics and, more recently, phylogenomics, have considerably challenged our view of the tree of life, including the diversity of species and their relationships. All zoologists, for example, certainly remember the debates challenging the relationships among bilaterians that led to the replacement of Acoelomata, Pseudocoelomata, Coelomata and Articulata taxa by Ecdysozoa and Lophotrochozoa. Subsequently, the evolution of the bilaterian body plan was reinterpreted in light of the newly inferred relationships (Adoutte et al. 2000; Aguinaldo et al. 1997; Glenner et al. 2004; Halanych et al. 1995). Then, in the same period, molecular phylogenies demonstrated the monophyly of Metazoa including sponges (Borchiellini et al. 1998, 2000; Hanelt et al. 1996; Kim et al. 1999; Kobayashi et al. 1996; Kruse et al. 1998; Morris 1993; Müller 1995, 1998; Pawlowski et al. 1996; Zrzavý et al. 1998). This important result definitively invalidated the status of sponges as "parazoa" (i.e., "beside animal") and ended a debate

that had been present since the middle of the twentieth century (Brien 1967; Hadzi 1953; Hyman 1940, 1951).

This renewed vision of the animal tree fueled discussions in the evo-devo field and raised questions regarding the origins of bilaterians and the characteristics of the last common ancestor of Bilateria. This promoted comparative developmental studies between bilaterian models and other bilaterians but, also, promoted a comparison with non-bilaterian animals (Jenner and Wills 2007; Lanna 2015; Russell et al. 2017; Tickle and Urrutia 2017). In addition, increased efforts were made to establish a reliable phylogenetic context to interpret the acquired data. Therefore, while the ancestral features of bilaterians were more clearly defined (Deutsch and Le Guyader 1998; Slack et al. 1993), phylogenetic studies have shifted the debate to the nature of the last common ancestor of all extant animals.

From 1996 to 2009, the central aim was to decipher whether Porifera was monophyletic or not. Several molecular studies supported sponge paraphyly (Borchiellini et al. 2001; Cavalier-Smith et al. 1996; Collins 1998; Kruse et al. 1998; Medina et al. 2001; Peterson and Butterfield 2005; Peterson and Eernisse 2001; Sperling et al. 2007, 2009) leading to the conclusion that the last common ancestor of metazoans would have had sponge-like morpho-anatomical features and that the characters that defined sponges would have been plesiomorphic (Borchiellini et al. 2001; Cavalier-Smith 2017; Philippe et al. 2009; Sperling et al. 2009). Consequently, sponge characters would have been lost in the ancestor of eumetazoans, while the digestive tract and the nervous system would have been secondarily acquired during animal evolution. While this view was fully consistent with the traditional scenario of body plan complexification (Sperling et al. 2009), the hypothesis of sponge paraphyly no longer receives support since the rise of phylogenomic approaches (Philippe et al. 2009; Pick et al. 2010). Thus, the debate has shifted again and is now focused on the most basal positions in the animal tree. Indeed, in 2008, Casey Dunn and colleagues proposed that instead of sponges, comb jellies (Ctenophora) should be considered as the sister group of all other animals (Dunn et al. 2008) (Fig. 7.4).

Depending on the hypothesis that will be retained (i.e., Porifera-first or Ctenophora-first), the scenario of early animal evolution is totally different (King and Rokas 2017) (Fig. 7.4). If sponges are the sister group of all other animals, we can equally postulate a simple (no nerves, no muscles, no digestive tract) or a complex animal ancestor. In contrast, if comb jellies emerged first, then the last animal ancestor must have been complex and the last common ancestor of extant animals would have had a digestive tract, muscles and maybe a nervous system [some authors have proposed the convergent emergence of the nervous system (Jékely et al. 2015; Marlow and Arendt 2014; Moroz 2015; Moroz and Kohn 2015, 2016; Moroz et al. 2014; Ryan 2014; Ryan and Chiodin 2015)]. In this second case, sponges underwent a profound body simplification. Such a simplification is fully reliable, according to other well-known examples of sessile or parasitic animals: such as cirripeds, urochordates and myxozoans for example (Atkinson et al. 2018; Fu et al. 2008; Holland 2016; Mouchel-Vielh et al. 1998).

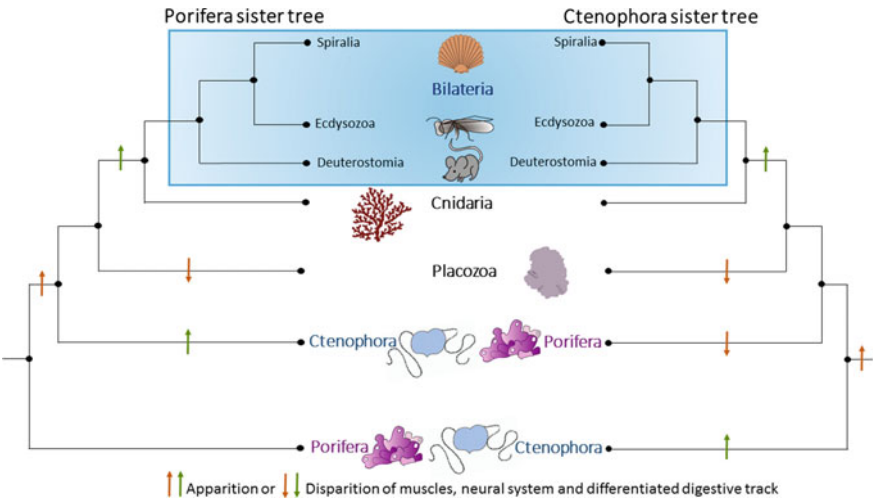


Fig. 7.4 For ten years, the debate concerning the animal tree has been focused on the relationships between the four non-bilaterian lineages (including Porifera whose basal-most position has been challenged). Depending on the relative positioning of ctenophores and sponges, the scenario of the evolution of some key animal features varies (Dunn et al. 2015; Halanych et al. 2016; King and Rokas 2017; Pisani et al. 2015; Simion et al. 2017; Whelan et al. 2017). Orange and green arrows represent alternative scenarios for neurons, muscles and a digestive tract according to the Porifera-first hypothesis (left) or according to the Ctenophora-first hypothesis (right)

Since 2008, a scientific debate opposes teams having strong backgrounds and experience, and having used the most complete datasets and the best methods available so far (Feuda et al. 2017; Simion et al. 2017; Whelan et al. 2015, 2017). It is, therefore, extremely difficult for most zoologists to dispassionately agree with one hypothesis or the other.

According to the quite strong support obtained for much more ancient nodes within the eukaryote tree (Adl et al. 2018), the irresolution of deep branches of the metazoan tree cannot only be attributed to limitations in the current methodology when trying to resolve deep nodes, as suggested by some authors (Halanych 2015; Nosenko et al. 2013). Indeed, because the taxa of interest diverged approximately 700 million years ago (Dohrmann and Wörheide 2017), multiple substitution events have likely occured in each lineage, hence making the phylogenetic signal partly masked by saturation and noise. To tackle this issue, phylogeneticists commonly vary the choice of the genetic markers included, the number of species compared and the choice of the outgroup. However, these classical remedies have failed to solve the current problem efficiently. We comment on the main unresolved biases that have been identified to date (clearly discussed in (King and Rokas 2017; Nosenko et al. 2013; Pett et al. 2019; Philippe et al. 2011))

– *Taxa sampling and the lack of data:* The number of transcriptomes, or even genomes, of non-bilaterians remains too limited to fully reflect their real diversity (Eitel et al. 2018; Renard et al. 2018). In addition, increasing the number of available databases for each taxon should allow us to find slow-evolving representatives that may help to retrace early events in animal evolution. This approach, for example, has been used to define the Ecdysozoa clade thanks to the choice of slow-evolving nematode sequences (Aguinaldo et al. 1997). Finding a glass sponge and a ctenophore that evolve more slowly, therefore, may be important since these two lineages often show long branches. Hence, sequencing even more non-bilaterian species would help to limit both saturation and long-branch attraction biases (Felsenstein 1978).
– *Choice of the model*: Choosing the best model of substitution to describe the evolution of sequences is another key step of phylogenetic analyses. It is well known that there is a variation of substitution rates between the sites of a sequence and between branches (Uzzell and Corbin 1971). For this reason, site-heterogeneous models usually provide a better description of the biological reality (e.g., codon redundancy and different constraints of functional domains). However, the more complex a model, the more parameters have to be evaluated and the longer the calculation time. This can be a major difficulty regarding the constant increase in the size of datasets. In this context, it is worth noting that using site-homogeneous models tend to favor the ctenophore-first hypothesis, while choosing site-heterogeneous models tend to favor the sponge-first hypothesis (King and Rokas 2017). Moreover, since the best fitting models differ from one gene family to another, not only a model taking into account site-heterogeneity is needed, but, also, data partitioning is needed to analyze them appropriately, meaning as biological entities with their own histories and functional constraints.
– *Orthology assessment*: Most animal genes are members of multigenic families. This means that, in a given genome, several homologous genes (i.e., that have descended from an ancestral copy) cohabit. Genes that have emerged via duplication events are named paralogs, while similar copies inherited by different species (i.e., speciation events) are named orthologs. A basic rule in phylogeny is to ensure the comparison of homologous characters. In the case of sequences, the characters to compare are the sequence sites (i.e., the amino acids), and the only way to ensure the comparison of homologous sites is to properly align orthologous sequences. Orthology assignment is thus essential to avoid phylogenic bias (Pett et al. 2019; Simion et al. 2017). However, identifying proper orthology groups in non-bilaterians is often difficult to do with confidence because of high sequence divergence and in-paralogs that are restricted to some taxa (i.e., often less copies in non-bilaterian animals). To tackle this issue, only single-copied genes or confidently assigned genes should be integrated into the analyses. To do so, however, time-consuming analyses (because not fully automatable) would be a prerequisite to infer the evolution of each gene (e.g., Fierro-Constaín et al. 2017; Schenkelaars et al. 2015).

– *Horizontal or Lateral Gene Transfer (HGT or LGT) events*: HGTs are events during which a sequence is transferred from one species to another. It was first described in the 1980s (Syvanen 1984) and leads to "disturbance" in gene phylogenies (Philippe et al. 2011). There is growing evidence that HGTs are more frequent in eukaryotes, including metazoans, than previously thought (Dunning et al. 2019; Dunning Hotopp 2018; Leger et al. 2018; Martin 2017; Nedelcu 2019; Ramulu et al. 2012). But, to our knowledge, no study has examined the impact of HGTs in the sponge-first versus ctenophore-first debate.

We should be confident that not only additional genomes (Renard et al. 2018) but, also, new methods, tools and innovative procedures (Kalyaanamoorthy et al. 2017; Oxusoff et al. 2018) may provide new opportunities to tackle this complex and exciting issue. Meanwhile, the scientific community should keep in mind this uncertainty and discuss their results in both contexts (i.e., Porifera- and Ctenophora-first).

In conclusion, after an exciting ten years of debate, the controversy as to whether Porifera or Ctenophora comes first in the animal tree is ongoing. The major positive outcome of this debate is that it has shed light on two neglected animal phyla and has renewed interest in studying them (Belahbib et al. 2018; Dunn et al. 2015; Fidler et al. 2017; Fierro-Constaín et al. 2017; Miller et al. 2018; Schenkelaars et al. 2015, 2016a, b, 2017; Schippers et al. 2018). Such comparative studies are undoubtedly a prerequisite to understanding early animal evolution (see Sect. 7.3).

7.4 An Ancestral Molecular Toolkit More Complex Than Imagined Twenty Years Ago

Until the 1970s, it was commonly thought that the genome size directly reflected body size and complexity. The *c*-value paradox concept, however, postulated that cellular DNA content did not correlate with body complexity (Eddy 2012). Furthermore, the number of genes is actually insufficient to explain the differences in genome size in eukaryotes (*g*-value paradox). These differences are mainly due to variations in non-coding DNA (pseudogenes, satellites, regulatory sequences, etc.) (Hahn and Wray 2002). Sponges illustrate well these two key concepts of genome evolution since they possess a simpler body organization than most other invertebrates, while their genome size and gene number seem similar to most non-chordate invertebrates. Nevertheless, the size and the gene content of sponge genomes appear to be much higher than their unicellular relatives (Fernandez-Valverde and Degnan 2016; Renard et al. 2018; Sebé-Pedrós et al. 2018).

Early work on the gene content of sponges (by PCR with degenerated primers or small EST datasets) showed that Porifera (at least Demospongiae, Homoscleromorpha and Calcarea) have in common with all other animals, genes involved in key developmental processes such as transcription factors (Degnan et al. 1993, 2009; Gazave et al. 2008; Larroux et al. 2007, 2008; Manuel and Le Parco 2000) and

members of signaling pathways (Adell et al. 2007; Gazave et al. 2009; Lapébie et al. 2009; Richards et al. 2008; Suga et al. 1999). This important finding was then largely confirmed and completed by the acquisition of larger transcriptomic and genomic datasets (Fortunato et al. 2015; Francis et al. 2017; Riesgo et al. 2014; Srivastava et al. 2010). Today, we can assert that the last common ancestor of sponges and the last common ancestor of all metazoans already possessed:

– *The genes to sustain at least six signaling pathways:* The core components of the nuclear receptor (NR), Wnt, Notch, TGF-β, Hippo and nitric oxide pathways were retrieved (Adamska et al. 2007a, b, 2010; Borisenko et al. 2016; Gazave et al. 2009; Hall et al. 2019; Lapébie et al. 2009; Nichols et al. 2006; Richards and Degnan 2012; Schenkelaars et al. 2015, 2016a, b, 2017; Ueda et al. 2016; Windsor and Leys 2010; Windsor Reid et al. 2018) [for review see: (Adamska 2016; Adamska et al. 2011; Babonis and Martindale 2017; Renard et al. 2018; Richards and Degnan 2009)], but, according to present available data, Hedgehog, Fibroblast Growth Factor (FGF) or HIF core components are missing. (Mills et al. 2018; Riesgo et al. 2014). Nevertheless, when the minimal set of members enabling to predict the establishment of a functional pathway is retrieved, sponges often lack some of the agonists and antagonists described in bilaterian. This suggests a more "basic" or "linear" signaling pathway (with less tuning possibilities) instead of complex interaction networks as described in bilaterians (Babonis and Martindale 2017; Gazave et al. 2009; Hinck et al. 2016; Schenkelaars et al. 2017) (Fig. 7.5).

It is cautionary to note that, despite the presence of required genes, some authors have recently questioned the role of a canonical Wnt signaling in the establishment of body polarity in sponges (Schenkelaars et al. 2017; Windsor Reid et al. 2018). In addition, the ancestral presence of functional canonical notch signaling (targeting the TF *Hes*) is unsure because of the functional results today available in Cnidarians (Babonis and Martindale 2017).

– *A great variety of transcription factors (TFs):* A recent study (Sebé-Pedrós et al. 2018) has yielded to the annotation of 232 TFs (pertaining to various families bHLH, bZIP, homeobox, Fox, T-box, etc.) in the sponge *Amphimedon queenslandica*, of which 168 showed cell-type-specific expression patterns. This exhaustive and unbiased analysis performed by the single-cell transcriptomic approach massively extends our previous knowledge of TF diversity in sponges (Fortunato et al. 2012, 2014, 2015, 2016; Mah and Leys 2017; Pastrana et al. 2019; Sebé-Pedrós et al. 2013, 2017). According to these new data obtained on the demosponge *A. queenslandica,* sponges would have a similar diversity of TFs as ctenophores and placozoans. In contrast, TFs are less diverse in Eumycota and holozoan unicellular relatives and more diverse in cnidarians and bilaterians. Nevertheless considering the differences in TF content reported previously between sponge classes (Fortunato et al. 2015, 2016; Gazave et al. 2014; Riesgo et al. 2014), it would be worth acquiring similar data from other sponge lineages.

Despite its present incompleteness, the information on the developmental toolkit of sponges that have accumulated during the last two decades enables us to partly

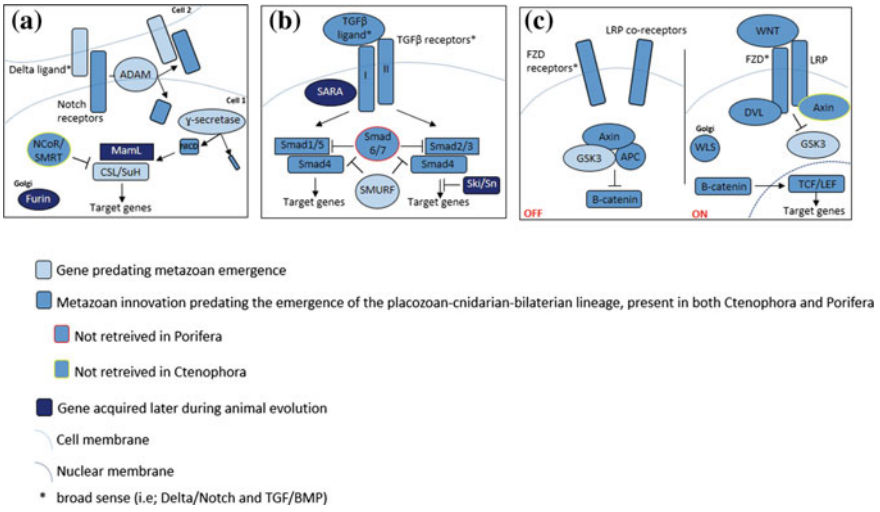


Fig. 7.5 Scenarios for the origin and evolution of three main signaling pathways involved in metazoan development (canonical Notch, TGF beta and canonical Wnt pathways) according to Babonis and Martindale (2017). These schemas show that (i) few members were already present before the emergence of the metazoan lineage, (ii) a minimal set of genes was already present in the last common ancestor of metazoans and enables us to envisage functional pathways (even though experimental evidences are still missing), and (iii) these minimal pathways were enriched by additional components after the divergence of Porifera and Ctenophora

answer to L. Caroll's questioning (Carroll 2005) concerning the developmental genetic toolkit of non-bilaterians. Sponges do lack some major components of the bilaterian developmental toolkit (e.g., *Hox* genes, *FGF* and *Hh* signaling), but, despite their morpho-anatomical simplicity, they do possess a large variety of molecular tools. Unfortunately, the most exhaustive studies on signaling pathways and TFs were mainly performed on transcriptomes and on a few sponge species (Borisenko et al. 2016; Fortunato et al. 2016; Gazave et al. 2009; Riesgo et al. 2014; Schenkelaars et al. 2016a; Sebé-Pedrós and Ruiz-Trillo 2017; Sebé-Pedrós et al. 2013). Therefore, we currently have a limited view of the gene content in sponges. This is even more plausible given that several gene losses have already been reported accordingly to the monophyly of sponges. For instance, the *Hes* and *Vang* genes were retrieved only Homoscleromorpha and would have been lost in other sponge classes (Gazave et al. 2014; Schenkelaars et al. 2016a), the *Wnt* and *Dsh* genes were secondarily lost in the glass sponge lineage (Riesgo et al. 2014; Schenkelaars et al. 2017; Windsor Reid et al. 2018), and the losses of the gene encoding for type IV collagen in silicious sponges (Fidler et al. 2017) and the Hox/ParaHox in some or all sponges have been documented (Ferrier 2016; Fortunato et al. 2014; Pastrana et al. 2019). Non-comprehensive analyses on the four sponge classes, therefore, may lead to an underestimation of the TF diversity and of the number of functional pathways that were present in the last common ancestor of all extant sponges.

7.5 The Need for Comparisons Among and Outside Animals

The importance of performing comparative studies between sponge classes is highlighted by differences in developmental, ecological, histological and molecular characteristics (for review see Renard et al. 2018). This bias should be considered not only for sponges but also for other pre-bilaterian lineages. Indeed, there is a similar lack of completeness and representativity in available data for ctenophores and placozoans regarding the diversity and heterogeneity of these taxa (Eitel et al. 2018; Fidler et al. 2017). Consequently, we cannot fully exclude the possibility that the molecular complexity of the last common ancestor of metazoans may still be underestimated. These remains much work to do, therefore, concerning non-bilaterian comparative biology at different levels of organization (Dunn et al. 2015; Lanna 2015). Encouragingly, the number of comparative studies that include several sponge classes and other non-bilaterian phyla has increased during the last two years (Belahbib et al. 2018; Fidler et al. 2017; Fierro-Constaín et al. 2017; Schenkelaars et al. 2016a, 2017; Sebé-Pedrós et al. 2018), and recent work has answered a number of fundamental questions related to the origin of multicellularity, the evolution of key body plan features and the origin of gastrulation (King and Rokas 2017; Lanna 2015).

Thanks to these extensive efforts, the portrait of the last common ancestor of all metazoans is becoming clearer. Current data suggest that the ancestor of animals was probably marine, acquiring multicellularity and epithelialization via segmentation and reaching a larval stage via gastrulation. Furthermore, the adult stage probably possessed a polarized body and was able to reproduce both sexually and asexually (Fig. 7.6).

If we assume that the above features are ancestral in metazoans, it remains to discover how and when the transition from unicellularity with a transient colonial association to permanent pluricellularity occurred. To answer this question, we need to explore the genomes and biology of metazoan unicellular relatives. The studies of two non-metazoan holozoan lineages, Filasterea *(Capsaspora owczarzaki*) and Choanoflagellata (*Monosiga brevicollis, Salpingoeca rosetta*) were started in 2001 (King and Carroll 2001). These organisms possess, for example, some genes encoding for so-called epithelial proteins such as cadherins and collagens (Brunet and King 2017; Grau-Bové et al. 2017; Renard et al. in press; Sebé-Pedrós et al. 2017) together with certain components of signaling pathways. It appears that metazoan multicellularity was partly built by reusing or rearranging ancient molecular tools. Unfortunately, to date the unicellular–multicellular transition remains a mystery because the function of these proteins is poorly investigated in these unicellular lineages. Much is therefore expected from the newly developed transfection methods raised in these organisms (Booth et al. 2018; Parra-Acero et al. 2018; Sebé-Pedrós et al. 2017).

In a similar way to the non-bilaterian lineages, it is interesting to note that comparative data have shown that filasterean or choanoflagellate species do not possess the same gene content. Focusing on only one species, therefore, would have resulted

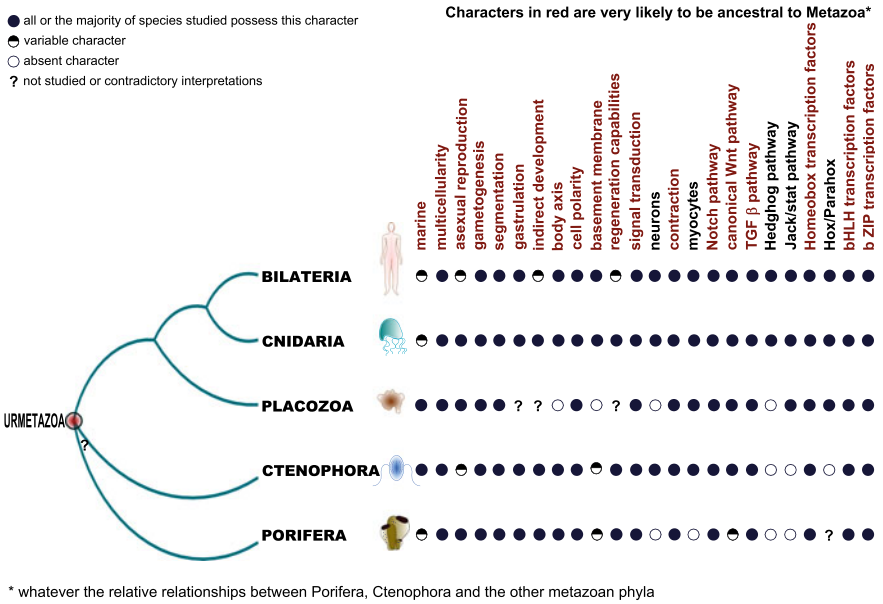


Fig. 7.6 Recent biological observations and genome surveys in various sponge species and other non-bilaterian taxa offer a new interpretation of the complexity of the last common ancestor of all extant metazoans

in erroneous evolutionary interpretations (Fairclough et al. 2013; Grau-Bové et al. 2017; Hoffmeyer and Burkhardt 2016; Sebé-Pedrós et al. 2017). However, these (still incomplete) genomic analyses have already enabled the determination of molecular synapomorphies that support the node of Metazoa (about 200 according to Suga et al. 2013). For example, genes involved in the establishment of epithelia, cell-cell communication and cell differentiation are metazoan innovations: classical E-cadherin, beta- and delta catenins, Wnt, Notch, TGF beta signaling pathways, Nanos, vasa, Par 3, Par 6, aPKC, scribble, LGL and the laminins (see previous references for signaling pathways+ Belahbib et al. 2018; Fahey and Degnan 2010, 2012; Fierro-Constaín et al. 2017; Ganot et al. 2015; Le Bivic 2013; Miller et al. 2013, 2018; Murray and Zaidel-Bar 2014; Renard et al. 2018; Richter et al. 2018; Schippers et al. 2018). In addition to gene inventories, these data provided information about the evolution of genome organization between non-metazoan and metazoan lineages (the length and abundance of introns, the number and proximity of coding sequences, the abundance of repetitive elements, etc.). It is now clear that massive events of expansion of gene families and of domain reorganization occurred during the key unicellular–multicellular transition that gave rise to metazoans, even though sponges seem to have retained a relatively compact genome (Babonis and Martindale 2017; Fernandez-Valverde and Degnan 2016; Grau-Bové et al. 2017).

7.6 The Present Limits to Be Overtaken

The main limit of these more and more exhaustive gene surveys is that they fail to explain histological or morpho-anatomical differences between taxa. For example: (i) sponges possess pre- and post-synaptic genes, whereas proper synapses were never observed in these animals, (ii) calcarean sponges possess a gene encoding for type IV collagen, whereas they do not have any basement membrane, (iii) all sponges possess genes needed to establish a cadherin–catenin complex (CCC), whereas adherens junctions were described only in calcareous and homoscleromorpha sponges (Adams et al. 2010; Belahbib et al. 2018; Leys 2015; Leys and Hill 2012; Leys and Riesgo 2012; Leys et al. 2009; Mah and Leys 2017; Renard et al. 2009).

This apparent inconsistency probably results from the gaps in our knowledge of sponge biology at different levels:

– *Cell-type homology and cell functions:*

As previously noted, the number of cell types varies between sponge species, and, until recently the nomenclature of cell types was mainly based on cytological features (De Vos et al. 1991; Ereskovsky 2010). However, it has become clear from different molecular and experimental approaches, together with advances in cell staining, that (i) some cells with similar functions but slightly different cytological differences received different names (e.g., archeocytes/archeocyte-like/type 2 vacuolar cells (Fierro-Constaín et al. 2017); "Globular" or "flask" cells (Renard et al. 2009), and that (ii) similar cells of the same sponge species can be divided into different subpopulations, each of which express different genes and, therefore, probably play different roles (Sebé-Pedrós et al. 2018). While certain cell types have obvious functions (e.g., an epithelial role for pinacocyte; a skeleton synthesis role for sclerocytes), little is known about the exact functions fulfilled by each cell type. The recent single-cell (SC) transcriptomic data on *A. queenslandica* (Sebé-Pedrós et al. 2018) confirm the multitask character of sponge cells, suspected many years ago by spongiologists (Simpson 1984). In addition, recent experimental data also seem to confirm the lability of sponge cell types because they are able to transdifferentiate into other cell types (Ereskovsky et al. 2015; Funayama 2018; Funayama et al. 2010; Nakanishi et al. 2014; Sogabe et al. 2016). To fully evaluate the homology, functions and lability of sponge cell types, both single-cell transcriptomic approaches in different sponge species (currently in progress), and the development of new techniques to stain and track live cells during morphogenetic processes are essential (Borchiellini et al. in press). To date only choanocytes can be specifically stained and tracked in different sponge species (Alié et al. 2015; Nakanishi et al. 2014; Sogabe et al. 2016; Rocher et al. in prep.; Borchiellini et al. in press.)

– *Gene functions*:

Co-option or neofunctionalization are now recognized as powerful and relatively common evolutionary events, with important consequences for organism evolution (Brosius 2018; Oakley 2017). For this reason, retrieving a gene in a sponge genome does not mean that this gene plays the same role in sponges as in bilaterians. The

previously cited examples (neuronal genes, AJs and basement membrane) suggest that some proteins (E-cadherin, type IV collagen for example) might play different roles in sponges (at least in some classes). The new era of sponge evo-devo now consists in the development of experimental tools to explore gene functions (Lanna 2015; Sebé-Pedrós et al. 2017). Localization of structural proteins may be a first approach, but to decipher finely the function of most proteins we now need efficient methods for disturbing gene transcription or translation. At present, these whole set of experimental tools is mastered only in the demosponge *Ephydatia muelleri* (Hall et al. 2019; Rivera et al. 2013, 2011; Windsor Reid et al. 2018) (Table 7.1). Unfortunately, these tools are still missing in other sponge classes.

Because *E. muelleri* is a freshwater sponge (a rare ecological characteristic compared to the diversity of marine sponge species), at the present time, it cannot be excluded that this lineage has developed particular adaptations and functions, not present in all sponges, to deal with the unusual osmotic conditions. To date, successful gene transfection has only been reported in two marine species *Suberites domuncula* and *Tethya wilhelma* (Revilla-I-Domingo et al. 2018; Rivera et al. 2011), but the transfection efficiency reported in these species is too low to perform efficient knockdown or ectopic expression studies. Further, most marine sponges tolerate only low salinity or osmotic changes, and, therefore, altering the salinity of the media to ensure RNA stability or to enable electroporation [as seen in other marine animals such *Ciona* (Stolfi and Christiaen 2012)] seems difficult. Similar difficulties in raising functional tools in marine species were previously faced for other animal phyla, for instance in Cnidaria for which gene silencing was developed earlier in the freshwater genus *Hydra* than in its marine relatives (Galliot 2012; Rentzsch et al. 2017). So to summarize, at present, as soon as its genome is available (currently in progress as mentioned in Renard et al. 2018) only the freshwater sponge species *E. muelleri* seems nearly ready to enter the post-genomic era.

7.7 Main Conclusions

This chapter offers a summary of the main advances made during the two last decades (with more emphasis on the last ten years), concerning sponge studies from an evolutionary perspective. The data that have been accumulated concerning sponge phylogeny, genomics, cell and developmental biology, in addition to the experimental tools mastered for several sponge species, indicate that only few species (Table 7.1) are ready to be considered as "non-model organisms", the study of which is desperately necessary to answer key biological evolutionary questions (Russell et al. 2017). Nevertheless, what we have learned about sponges during recent years has profoundly changed our view on these uncommon animals and has challenged the way in which early animal evolution is viewed. Indeed:

– Molecular data have deeply challenged our view on Porifera, long considered as "not really" animals.

Table 7.1 List of sponge species for which comparative cell and developmental studies have been performed

Class	Genus	Species	Country	Genome (draft)	Transcriptome/EST	In situ hybridization	Immunolocalization	Pharmacological assays	Transfection	RNAi silencing
Calcarea	*Leucosolenia*	*L. complicata*	Australia	X	X	X	O	O	O	O
		L. variabilis	France, Russia	O	O	O	X	O	O	O
	Sycon	***S. ciliatum***	Australia	X	X	X	O	X	O	O
		S. coactum	Australia	O	X	O	O	O	O	O
		S. raphanus	Germany	O	X	X	X	O	O	O
Demospongiae	*Amphimedon*	***A. queenslandica***	Australia	X	X	X	X	X	O	O
	Ephydatia	*E. fluviatilis*	Japan	O	X	X	O	X	O	O
		E. muelleri	Canada, USA	O	X	X	X	X	X	X
	Halisarca	*H. caerulea*	Netherlands	O	X	O	X	O	O	O
		H. dujardinii	France, Russia	O	X	X	O	O	O	O
	Spongilla	***S. lacustris***	Canada, Germany	O	X	X	X	X	X	O
	Suberites	***S. domuncula***	Germany	O	X	O	X	X	X	O
	Tethya	***T. wilhelma***	Germany	X	X	O	O	O	X	X
Hexactinellida	*Oopsacas*	*O. minuta*	France	X	X	O	O	O	O	O
	Aphrocallistes	*A. vastus*	Canada	O	X	O	O	O	O	O
Homoscleromorpha	*Oscarella*	***O. lobularis***	France	X	X	X	X	X	O	O
		O. pearsei	USA	X	X	O	X	O	O	O

For each species, the countries where the species is currently studied, as well as available genetic resources and experimental tools are given: Red O stands for the absence of the resource/technique, while green X stands for its presence. Species names in bold indicate promising developing models for which a larger set of data and techniques are already available (Adamska et al. 2007a, b, 2010; Alexander et al. 2014, 2015; Alié et al. 2015; Borisenko et al. 2016; Elliott and Leys 2010; Ellwanger and Nickel 2006; Ellwanger et al. 2007; Fernandez-Valverde and Degnan 2016; Fierro-Constaín et al. 2017; Fortunato et al. 2012, 2014, 2015, 2016; Funayama et al. 2005a, b, 2010; Gazave et al. 2008; Hall et al. 2019; Lapébie et al. 2009; Leininger et al. 2014; Ludeman et al. 2014; Mah and Leys 2017; Manuel 2001; Miller et al. 2018; Mohri et al. 2008; Müller et al. 2012; Nakanishi et al. 2015; Nichols et al. 2006, 2012; Nickel 2010; Peña et al. 2016; Richards et al. 2008; Riesgo et al. 2014; Rivera et al. 2013, 2011; Sakarya et al. 2007; Schenkelaars et al. 2016b; Schippers et al. 2018; Sebé-Pedrós et al. 2013; Sogabe et al. 2016; Ueda et al. 2016; Voigt et al. 2014; Windsor and Leys 2010; Windsor Reid et al. 2018)

– Differences between and within sponge classes justify the development and comparison of several species to determine the ancestral features of the last common ancestor of sponges.
– Different sponge species or classes are thought to have undergone gene loss events, irrespective of their position in the animal tree. Consequently, comparisons of the four non-bilaterian lineages are necessary to decipher confidently the ancestral metazoan molecular toolkit and to try to deduce the ancestral morphological features.
– Functional data on non-bilaterians in general, and in sponges in particular, are rare. This is the biggest challenge we face currently for sponge evo-devo to enter the post-genomic era.

Acknowledgements The authors acknowledge the Région Provence Alpes Côte d'Azur and the French Research ministry for the PhD funding they provided to L. Fierro Ph.Ds of L. Fierro-Constain's and the Q. Schenkelaars respectively. The authors thank Prof. Stephen Prime (University of Bristol) and Mr. Thomas Smith, two native English speakers, for providing proofreading services. The authors are grateful to Dr. Pierre Pontarotti and Dr. Anne Chenuil-Maurel for constructive comments for Sect. 7.2. The authors thank the French National Center for Scientific Research (CNRS), Aix-Marseille University and the Amidex foundation for providing funds to support fundamental research.

References

Adams EDM, Goss GG, Leys SP (2010) Freshwater sponges have functional, sealing Epithelia with high transepithelial resistance and negative transepithelial potential. PLoS One 5(11):e15040.

Adamska M (2016) Sponges as models to study emergence of complex animals. Curr Opin Genet Dev 39:21–28

Adamska M, Matus DQ, Adamski M, Green K, Rokhsar DS, Martindale MQ, Degnan BM (2007a) The evolutionary origin of hedgehog proteins. Curr Biol 17:R836–R837

Adamska M, Degnan SM, Green KM, Adamski M, Craigie A, Larroux C, Degnan BM (2007b) Wnt and TGF-beta expression in the sponge *Amphimedon queenslandica* and the origin of metazoan embryonic patterning. PLoS ONE 2:e1031

Adamska M, Larroux C, Adamski M, Green K, Lovas E, Koop D, Richards GS, Zwafink C, Degnan BM (2010) Structure and expression of conserved Wnt pathway components in the demosponge *Amphimedon queenslandica*. Evol Dev 12:494–518

Adamska M, Degnan BM, Green K, Zwafink C (2011) What sponges can tell us about the evolution of developmental processes. Zoology (Jena) 114:1–10

Adell T, Thakur AN, Müller WEG (2007) Isolation and characterization of Wnt pathway-related genes from Porifera. Cell Biol Int 31:939–949

Adl SM, Bass D, Lane CE, Lukeš J, Schoch CL, Smirnov A, Agatha S, Berney C, Brown MW, Burki F, et al (2018) Revisions to the classification, nomenclature, and diversity of Eukaryotes. J Eukaryot Microbiol

Adoutte A, Balavoine G, Lartillot N, Lespinet O, Prud'homme B, de Rosa R (2000) The new animal phylogeny: reliability and implications. Proc Natl Acad Sci USA 97:4453–4456

Aguinaldo AM, Turbeville JM, Linford LS, Rivera MC, Garey JR, Raff RA, Lake JA (1997) Evidence for a clade of nematodes, arthropods and other moulting animals. Nature 387:489–493

Alexander BE, Liebrand K, Osinga R, van der Geest HG, Admiraal W, Cleutjens JPM, Schutte B, Verheyen F, Ribes M, van Loon E et al (2014) Cell turnover and detritus production in marine sponges from tropical and temperate benthic ecosystems. PLoS ONE 9:e109486

Alexander BE, Achlatis M, Osinga R, van der Geest HG, Cleutjens JPM, Schutte B, de Goeij JM (2015) Cell kinetics during regeneration in the sponge *Halisarca caerulea*: how local is the response to tissue damage? Peer J 3:e820

Alié A, Hayashi T, Sugimura I, Manuel M, Sugano W, Mano A, Satoh N, Agata K, Funayama N (2015) The ancestral gene repertoire of animal stem cells. Proc Nat Acad Sci 201514789

Atkinson SD, Bartholomew JL, Lotan T (2018) Myxozoans: ancient metazoan parasites find a home in phylum Cnidaria. Zoology (Jena) 129:66–68

Babonis LS, Martindale MQ (2017) Phylogenetic evidence for the modular evolution of metazoan signalling pathways. Philos Trans R Soc Lond B Biol Sci 372

Belahbib H, Renard E, Santini S, Jourda C, Claverie J-M, Borchiellini C, Bivic AL (2018). New genomic data and analyses challenge the traditional vision of animal epithelium evolution. BMC Genomics 228452

Booth DS, Szmidt-Middleton H, King N (2018) Choanoflagellate transfection illuminates their cell biology and the ancestry of animal septins. Mol Biol Cell mbcE18080514

Borchiellini C, Boury-Esnault N, Vacelet J, Le Parco Y (1998) Phylogenetic analysis of the Hsp70 sequences reveals the monophyly of Metazoa and specific phylogenetic relationships between animals and fungi. Mol Biol Evol 15:647–655

Borchiellini C, Chombard C, Lafay B, Boury-Esnault N (2000) Molecular systematics of sponges (Porifera). Hydrobiologia 420:15–27

Borchiellini C, Manuel M, Alivon E, Boury-Esnault N, Vacelet J, Parco YL (2001) Sponge paraphyly and the origin of Metazoa. J Evol Biol 14:171–179

Borisenko I, Adamski M, Ereskovsky A, Adamska M (2016) Surprisingly rich repertoire of Wnt genes in the demosponge *Halisarca dujardini*. BMC Evol Biol 16:123

Boute N, Exposito J-Y, Boury-Esnault N, Vacelet J, Noro N, Miyazaki K, Yoshizato K, Garrone R (1996) Type IV collagen in sponges, the missing link in basement membrane ubiquity. Biol Cell 88:37–44

Brien P (1967) Les éponges: leur nature métazoaire-leur gastrulation-leur état colonial. Ann Soc Roy Zool Belg 97:197–235

Brosius J (2018) Exaptation at the molecular genetic level. Sci China Life Sci

Brunet T, King N (2017) The origin of animal multicellularity and cell differentiation. Dev Cell 43:124–140

Brusca RC, Brusca GJ (2003). Invertebrates (Sinauer Associates)

Carroll S (2005) Bringing Evo Devo to Life. PLOS Biol. W. W. Norton and Company, New York 350 p. ISBN (hardcover) 0-393-06016-0

Cavalier-Smith T (2017) Origin of animal multicellularity: precursors, causes, consequences—the choanoflagellate/sponge transition, neurogenesis and the Cambrian explosion. Philos Trans R Soc Lond B Biol Sci 372

Cavalier-Smith T, Allsopp MTEP, Chao EE, Boury-Esnault N, Vacelet J (1996) Sponge phylogeny, animal monophyly, and the origin of the nervous system: 18S rRNA evidence. Can J Zool 74:2031–2045

Collins AG (1998) Evaluating multiple alternative hypotheses for the origin of Bilateria: an analysis of 18S rRNA molecular evidence. Proc Natl Acad Sci USA 95:15458–15463

De Vos L, Rützler K, Boury-Esnault N, Donadey C, Vacelet J (1991) Atlas of sponge morphology. Smithsonian Institution Press, Washington. J Mar Biol Assoc UK 71(4):915–915

Degnan BM, Degnan SM, Naganuma T, Morse DE (1993) The ets multigene family is conserved throughout the Metazoa. Nucleic Acids Res 21:3479–3484

Degnan BM, Vervoort M, Larroux C, Richards GS (2009) Early evolution of metazoan transcription factors. Curr Opin Genet Dev 19:591–599

Deutsch J, Le Guyader H (1998) The neuronal zootype. An hypothesis. Comptes Rendus de l'Académie Des Sciences—Series III—Sciences de La Vie 321:713–719

Dohrmann M, Wörheide G (2017) Dating early animal evolution using phylogenomic data. Sci Rep 7

Dunn CW, Hejnol A, Matus DQ, Pang K, Browne WE, Smith SA, Seaver E, Rouse GW, Obst M, Edgecombe GD et al (2008) Broad phylogenomic sampling improves resolution of the animal tree of life. Nature 452:745–749

Dunn CW, Leys SP, Haddock SHD (2015) The hidden biology of sponges and ctenophores. Trends Ecol Evol 30:282–291

Dunning LT, Olofsson JK, Parisod C, Choudhury RR, Moreno-Villena JJ, Yang Y, Dionora J, Quick WP, Park M, Bennetzen JL, et al (2019) Lateral transfers of large DNA fragments spread functional genes among grasses. Proc Natl Acad Sci USA

Dunning Hotopp JC (2018) Grafting or pruning in the animal tree: lateral gene transfer and gene loss? BMC Genom 19:470

Eddy SR (2012) The C-value paradox, junk DNA and ENCODE. Curr Biol 22:R898–R899

Eerkes-Medrano DI, Leys SP (2006) Ultrastructure and embryonic development of a syconoid calcareous sponge. Invertebr Biol 125:177–194

Eitel M, Francis WR, Varoqueaux F, Daraspe J, Osigus H-J, Krebs S, Vargas S, Blum H, Williams GA, Schierwater B et al (2018) Comparative genomics and the nature of placozoan species. PLoS Biol 16:e2005359

Elliott GRD, Leys SP (2010) Evidence for glutamate, GABA and NO in coordinating behaviour in the sponge, *Ephydatia muelleri* (Demospongiae, Spongillidae). J Exp Biol 213:2310–2321

Ellwanger K, Nickel M (2006) Neuroactive substances specifically modulate rhythmic body contractions in the nerveless metazoon *Tethya wilhelma* (Demospongiae, Porifera). Front Zool 3:7

Ellwanger K, Eich A, Nickel M (2007) GABA and glutamate specifically induce contractions in the sponge *Tethya wilhelma*. J Comp Physiol A Neuroethol Sens Neural Behav Physiol 193:1–11

Ereskovsky A (2010) The comparative embryology of sponges, pp 209–230

Ereskovsky AV, Borisenko IE, Lapébie P, Gazave E, Tokina DB, Borchiellini C (2015) *Oscarella lobularis* (Homoscleromorpha, Porifera) regeneration: epithelial morphogenesis and metaplasia. PLoS ONE 10:e0134566

Fahey B, Degnan BM (2010) Origin of animal epithelia: insights from the sponge genome: evolution of epithelia. Evol Dev 12:601–617

Fahey B, Degnan BM (2012) Origin and evolution of laminin gene family diversity. Mol Biol Evol 29:1823–1836

Fairclough SR, Chen Z, Kramer E, Zeng Q, Young S, Robertson HM, Begovic E, Richter DJ, Russ C, Westbrook MJ et al (2013) Premetazoan genome evolution and the regulation of cell differentiation in the choanoflagellate *Salpingoeca rosetta*. Genome Biol 14:R15

Felsenstein J (1978) Cases in which parsimony or compatibility methods will be positively misleading. Syst Zool 27:401–410

Fernandez-Valverde SL, Degnan BM (2016) Bilaterian-like promoters in the highly compact *Amphimedon queenslandica* genome. Sci Rep 6:22496

Ferrier DEK (2016) The origin of the Hox/ParaHox genes, the ghost locus hypothesis and the complexity of the first animal. Brief Funct Genomics 15:333–341

Feuda R, Dohrmann M, Pett W, Philippe H, Rota-Stabelli O, Lartillot N, Wörheide G, Pisani D (2017) Improved modeling of compositional heterogeneity supports sponges as sister to all other animals. Curr Biol 0

Fidler AL, Darris CE, Chetyrkin SV, Pedchenko VK, Boudko SP, Brown KL, Gray Jerome W, Hudson JK, Rokas A, Hudson BG (2017) Collagen IV and basement membrane at the evolutionary dawn of metazoan tissues. Elife 6

Fierro-Constaín L, Schenkelaars Q, Gazave E, Haguenauer A, Rocher C, Ereskovsky A, Borchiellini C, Renard E (2017) The conservation of the germline multipotency program, from sponges to vertebrates: a stepping stone to understanding the somatic and germline origins. Genome Biol Evol 9:474–488

Fortunato S, Adamski M, Bergum B, Guder C, Jordal S, Leininger S, Zwafink C, Rapp HT, Adamska M (2012) Genome-wide analysis of the sox family in the calcareous sponge *Sycon ciliatum*: multiple genes with unique expression patterns. EvoDevo 3:14

Fortunato SAV, Adamski M, Ramos OM, Leininger S, Liu J, Ferrier DEK, Adamska M (2014) Calcisponges have a ParaHox gene and dynamic expression of dispersed NK homeobox genes. Nature 514:620–623

Fortunato SAV, Adamski M, Adamska M (2015) Comparative analyses of developmental transcription factor repertoires in sponges reveal unexpected complexity of the earliest animals. Mar Genomics 24:121–129

Fortunato SAV, Vervoort M, Adamski M, Adamska M (2016) Conservation and divergence of bHLH genes in the calcisponge *Sycon ciliatum*. EvoDevo 7

Francis WR, Eitel M, Vargas S, Adamski M, Haddock SH, Krebs S, Blum H, Erpenbeck D, Wörheide G (2017) The genome of the contractile demosponge *Tethya wilhelma* and the evolution of metazoan neural signalling pathways. BioRxiv 120998

Fu X, Adamski M, Thompson EM (2008) Altered miRNA repertoire in the simplified chordate, *Oikopleura dioica*. Mol Biol Evol 25:1067–1080

Funayama N (2018) The cellular and molecular bases of the sponge stem cell systems underlying reproduction, homeostasis and regeneration. Int J Dev Biol 62:513–525

Funayama N, Nakatsukasa M, Kuraku S, Takechi K, Dohi M, Iwabe N, Miyata T, Agata K (2005a) Isolation of Ef silicatein and Ef lectin as molecular markers sclerocytes and cells involved in innate immunity in the freshwater sponge *Ephydatia fluviatilis*. Jzoo 22:1113–1122

Funayama N, Nakatsukasa M, Hayashi T, Agata K (2005b) Isolation of the choanocyte in the fresh water sponge, *Ephydatia fluviatilis* and its lineage marker. Ef annexin Dev Growth Differ 47:243–253

Funayama N, Nakatsukasa M, Mohri K, Masuda Y, Agata K (2010) Piwi expression in archeocytes and choanocytes in demosponges: insights into the stem cell system in demosponges. Evol Dev 12:275–287

Galliot B (2012) Hydra, a fruitful model system for 270 years. Int J Dev Biol 56:411–423

Ganot P, Zoccola D, Tambutté E, Voolstra CR, Aranda M, Allemand D, Tambutté S (2015) Structural molecular components of septate junctions in cnidarians point to the origin of epithelial junctions in eukaryotes. Mol Biol Evol 32:44–62

Gazave E, Lapébie P, Renard E, Bézac C, Boury-Esnault N, Vacelet J, Pérez T, Manuel M, Borchiellini C (2008) NK homeobox genes with choanocyte-specific expression in homoscleromorph sponges. Dev Genes Evol 218:479–489

Gazave E, Lapébie P, Richards GS, Brunet F, Ereskovsky AV, Degnan BM, Borchiellini C, Vervoort M, Renard E (2009) Origin and evolution of the Notch signalling pathway: an overview from eukaryotic genomes. BMC Evol Biol 9:249

Gazave E, Lavrov DV, Cabrol J, Renard E, Rocher C, Vacelet J, Adamska M, Borchiellini C, Ereskovsky AV (2013) Systematics and molecular phylogeny of the family oscarellidae (homoscleromorpha) with description of two new oscarella species. PLoS ONE 8:e63976

Gazave E, Guillou A, Balavoine G (2014) History of a prolific family: the Hes/Hey-related genes of the annelid Platynereis. EvoDevo 5:29

Gilbert SF (2010) Developmental biology, 9th edn. Sinauer Associates Inc, Sunderland, MA

Glenner H, Hansen AJ, Sørensen MV, Ronquist F, Huelsenbeck JP, Willerslev E (2004) Bayesian inference of the metazoan phylogeny: a combined molecular and morphological approach. Curr Biol 14:1644–1649

Grau-Bové X, Torruella G, Donachie S, Suga H, Leonard G, Richards TA, Ruiz-Trillo I (2017). Dynamics of genomic innovation in the unicellular ancestry of animals. ELife 6

Hadzi J (1953) An attempt to reconstruct the system of animal classification. Syst Zool 2(4):145–154

Haeckel E (1874) Die Gastraea-Theorie, die phylogenetische Classification des Thierreichsund die Homologie der Keimblätter. Jenaische Zeitschrift für Naturwissenschaft 8:1–55

Hahn MW, Wray GA (2002) The g-value paradox. Evol Dev 4:73–75

Halanych KM (2015) The ctenophore lineage is older than sponges? That cannot be right! Or can it? J Exp Biol 218:592–597
Halanych KM, Bacheller JD, Aguinaldo AM, Liva SM, Hillis DM, Lake JA (1995) Evidence from 18S ribosomal DNA that the lophophorates are protostome animals. Science 267:1641–1643
Halanych KM, Whelan NV, Kocot KM, Kohn AB, Moroz LL (2016) Miscues misplace sponges. Proc Natl Acad Sci USA 113:E946–E947
Hall C, Rodriguez M, Garcia J, Posfai D, DuMez R, Wictor E, Quintero OA, Hill MS, Rivera AS, Hill AL (2019) Secreted frizzled related protein is a target of PaxB and plays a role in aquiferous system development in the freshwater sponge *Ephydatia muelleri*. PLoS ONE 14:e0212005
Hanelt B, Van Schyndel D, Adema CM, Lewis LA, Loker ES (1996) The phylogenetic position of *Rhopalura ophiocomae* (Orthonectida) based on 18S ribosomal DNA sequence analysis. Mol Biol Evol 13:1187–1191
Hestetun JT, Tompkins-Macdonald G, Rapp HT (2017) A review of carnivorous sponges (Porifera: Cladorhizidae) from the Boreal North Atlantic and Arctic. Zool J Linn Soc 181:1–69
Hinck AP, Mueller TD, Springer TA (2016). Structural biology and evolution of the TGF-β Family. Cold Spring Harb Perspect Biol 8
Hoffmeyer TT, Burkhardt P (2016) Choanoflagellate models—*Monosiga brevicollis* and *Salpingoeca rosetta*. Curr Opin Genet Dev 39:42–47
Holland LZ (2016) Tunicates. Curr Biol 26:R146–R152
Hyman LH (1940). The invertebrates: mollusca I. McGraw-Hill
Hyman LH (1951) The invertebrates: Platyhelminthes and Rhynchocoela, the acoelomate Bilateria. Invertebr. Platyhelminthes Rhynchocoela Acoelomate Bilateria 2
Jékely G, Paps J, Nielsen C (2015) The phylogenetic position of ctenophores and the origin (s) of nervous systems. EvoDevo 6:1
Jenner RA, Wills MA (2007) The choice of model organisms in evo-devo. Nat Rev Genet 8:311–314
Johnson TA, Milan-Lobo L, Che T, Ferwerda M, Lambu E, McIntosh NL, Li F, He L, Lorig-Roach N, Crews P et al (2017) Identification of the first marine-derived opioid receptor "balanced" agonist with a signaling profile that resembles the endorphins. ACS Chem Neurosci 8:473–485
Kalyaanamoorthy S, Minh BQ, Wong TK, von Haeseler A, Jermiin LS (2017) ModelFinder: fast model selection for accurate phylogenetic estimates. Nat Methods 14:587–589
Kim J, Kim W, Cunningham CW (1999) A new perspective on lower metazoan relationships from 18S rDNA sequences. Mol Biol Evol 16:423–427
King N, Carroll SB (2001) A receptor tyrosine kinase from choanoflagellates: molecular insights into early animal evolution. PNAS 98:15032–15037
King N, Rokas A (2017) Embracing uncertainty in reconstructing early animal evolution. Curr Biol 27:R1081–R1088
Kobayashi M, Wada H, Satoh N (1996) Early evolution of the Metazoa and phylogenetic status of diploblasts as inferred from amino acid sequence of elongation factor-1 alpha. Mol Phylogenet Evol 5:414–422
Kotpal RL (2012). Modern text book of Zoology: Invertebrates. Rastogi Publications
Kruse M, Leys SP, Müller IM, Müller WE (1998) Phylogenetic position of the Hexactinellida within the phylum Porifera based on the amino acid sequence of the protein kinase C from *Rhabdocalyptus dawsoni*. J Mol Evol 46:721–728
Lanna E (2015) Evo-devo of non-bilaterian animals. Genet Mol Biol 38:284–300
Lapébie P, Gazave E, Ereskovsky A, Derelle R, Bézac C, Renard E, Houliston E, Borchiellini C (2009) WNT/beta-catenin signalling and epithelial patterning in the homoscleromorph sponge Oscarella. PLoS ONE 4:e5823
Larroux C, Fahey B, Degnan SM, Adamski M, Rokhsar DS, Degnan BM (2007) The NK Homeobox gene cluster predates the origin of Hox genes. Curr Biol 17:706–710
Larroux C, Luke GN, Koopman P, Rokhsar DS, Shimeld SM, Degnan BM (2008) Genesis and expansion of metazoan transcription factor gene classes. Mol Biol Evol 25:980–996
Le Bivic A (2013) Evolution and cell physiology. 4. Why invent yet another protein complex to build junctions in epithelial cells? Am J Physiol Cell Physiol 305:C1193–C1201

Leger MM, Eme L, Stairs CW, Roger AJ (2018) Demystifying eukaryote lateral gene transfer. BioEssays 40:1700242 (Response to Martin 2017 https://doi.org/10.1002/bies.201700115)
Leininger S, Adamski M, Bergum B, Guder C, Liu J, Laplante M, Bråte J, Hoffmann F, Fortunato S, Jordal S, et al (2014) Developmental gene expression provides clues to relationships between sponge and Eumetazoan body plans. Nat Commun 5:ncomms4905
Leys SP (2015) Elements of a "nervous system" in sponges. J Exp Biol 218:581–591
Leys SP, Hill A (2012) The physiology and molecular biology of sponge tissues. In: Advances in marine biology. Elsevier, pp 1–56
Leys SP, Riesgo A (2012) Epithelia, an evolutionary novelty of metazoans. J Exp Zool (Mol Dev Evol) 318:438–447
Leys SP, Nichols SA, Adams EDM (2009) Epithelia and integration in sponges. Integr Comp Biol 49:167–177
Ludeman DA, Farrar N, Riesgo A, Paps J, Leys SP (2014) Evolutionary origins of sensation in metazoans: functional evidence for a new sensory organ in sponges. BMC Evol Biol 14:3
Mah JL, Leys SP (2017) Think like a sponge: the genetic signal of sensory cells in sponges. Dev Biol 431:93–100
Manuel M (2001) Origine et evolution des mecanismes moleculaires controlant la morphogenese chez les metazoaires : un nouveau modele spongiaire, *Sycon raphanus* (calcispongia, calcaronea) thesis. Paris 11
Manuel M, Le Parco Y (2000) Homeobox gene diversification in the calcareous sponge, *Sycon raphanus*. Mol Phylogenet Evol 17:97–107
Marlow H, Arendt D (2014) Evolution: ctenophore genomes and the origin of neurons. Curr Biol 24:R757–R761
Martin WF (2017) Too much eukaryote LGT. Bioessays 39
Medina M, Collins AG, Silberman JD, Sogin ML (2001) Evaluating hypotheses of basal animal phylogeny using complete sequences of large and small subunit rRNA. Proc Natl Acad Sci USA 98:9707–9712
Miller PW, Clarke DN, Weis WI, Lowe CJ, Nelson WJ (2013) The Evolutionary origin of epithelial cell-cell adhesion mechanisms. Curr Top Membr 72:267–311
Miller PW, Pokutta S, Mitchell JM, Chodaparambil JV, Clarke DN, Nelson WJ, Weis WI, Nichols SA (2018) Analysis of a vinculin homolog in a sponge (phylum Porifera) reveals that vertebrate-like cell adhesions emerged early in animal evolution. J Biol Chem 293:11674–11686
Mills DB, Francis WR, Vargas S, Larsen M, Elemans CP, Canfield DE, Wörheide G (2018) The last common ancestor of animals lacked the HIF pathway and respired in low-oxygen environments. Elife 7
Mohri K, Nakatsukasa M, Masuda Y, Agata K, Funayama N (2008) Toward understanding the morphogenesis of siliceous spicules in freshwater sponge: differential mRNA expression of spicule-type-specific silicatein genes in *Ephydatia fluviatilis*. Dev Dyn 237:3024–3039
Moroz LL (2015) Convergent evolution of neural systems in ctenophores. J Exp Biol 218:598–611
Moroz LL, Kohn AB (2015) unbiased view of synaptic and neuronal gene complement in Ctenophores: are there pan-neuronal and pan-synaptic genes across Metazoa? Integr Comp Biol 55:1028–1049
Moroz LL, Kohn AB (2016) Independent origins of neurons and synapses: insights from ctenophores. Philos Trans R Soc Lond B Biol Sci 371
Moroz LL, Kocot KM, Citarella MR, Dosung S, Norekian TP, Povolotskaya IS, Grigorenko AP, Dailey C, Berezikov E, Buckley KM et al (2014) The ctenophore genome and the evolutionary origins of neural systems. Nature 510:109–114
Morris PJ (1993) The developmental role of the extracellular matrix suggests a monophyletic origin of the kingdom animalia. Evolution 47:152–165
Mouchel-Vielh E, Rigolot C, Gibert JM, Deutsch JS (1998) Molecules and the body plan: the Hox genes of Cirripedes (Crustacea). Mol Phylogenet Evol 9:382–389
Müller WE (1995) Molecular phylogeny of Metazoa (animals): monophyletic origin. Naturwissenschaften 82:321–329

Müller WEG (1998) Molecular phylogeny of Eumetazoa: genes in sponges (Porifera) give evidence for monophyly of animals. In: Molecular evolution: evidence for monophyly of metazoa. Springer, pp 89–132

Müller WEG, Wang X, Grebenjuk VA, Korzhev M, Wiens M, Schlossmacher U, Schröder HC (2012) Common genetic denominators for Ca++-based skeleton in Metazoa: role of osteoclast-stimulating factor and of carbonic anhydrase in a calcareous sponge. PLoS ONE 7:e34617

Murray PS, Zaidel-Bar R (2014) Pre-metazoan origins and evolution of the cadherin adhesome. Biol Open 3:1183–1195

Nakanishi N, Sogabe S, Degnan BM (2014) Evolutionary origin of gastrulation: insights from sponge development. BMC Biol 12:26

Nakanishi N, Stoupin D, Degnan SM, Degnan BM (2015) Sensory flask cells in sponge larvae regulate metamorphosis via calcium signaling. Integr Comp Biol 55:1018–1027

Nedelcu AM (2019) Independent evolution of complex development in animals and plants: deep homology and lateral gene transfer. Dev Genes Evol 229:25–34

Nichols SA, Dirks W, Pearse JS, King N (2006) Early evolution of animal cell signaling and adhesion genes. Proc Natl Acad Sci USA 103:12451–12456

Nichols SA, Roberts BW, Richter DJ, Fairclough SR, King N (2012) Origin of metazoan cadherin diversity and the antiquity of the classical cadherin/β-catenin complex. Proc Natl Acad Sci USA 109:13046–13051

Nickel M (2010) Evolutionary emergence of synaptic nervous systems: what can we learn from the non-synaptic, nerveless Porifera? Invertebr Biol 129:1–16

Nosenko T, Schreiber F, Adamska M, Adamski M, Eitel M, Hammel J, Maldonado M, Müller WEG, Nickel M, Schierwater B et al (2013) Deep metazoan phylogeny: when different genes tell different stories. Mol Phylogenet Evol 67:223–233

Oakley TH (2017) Furcation and fusion: the phylogenetics of evolutionary novelty. Dev Biol 431:69–76

Oxusoff L, Préa P, Perez Y (2018) A complete logical approach to resolve the evolution and dynamics of mitochondrial genome in bilaterians. PLoS ONE 13:e0194334

Parra-Acero H, Ros-Rocher N, Perez-Posada A, Kożyczkowska A, Sánchez-Pons N, Nakata A, Suga H, Najle SR, Ruiz-Trillo I (2018) Transfection of *Capsaspora owczarzaki*, a close unicellular relative of animals. Development 145

Pastrana CC, DeBiasse MB, Ryan JF (2019) Sponges lack ParaHox genes. Genome Biol Evol

Pawlowski J, Montoya-Burgos JI, Fahrni JF, Wüest J, Zaninetti L (1996) Origin of the Mesozoa inferred from 18S rRNA gene sequences. Mol Biol Evol 13:1128–1132

Peña JF, Alié A, Richter DJ, Wang L, Funayama N, Nichols SA (2016) Conserved expression of vertebrate microvillar gene homologs in choanocytes of freshwater sponges. Evodevo 7:13

Peterson KJ, Butterfield NJ (2005) Origin of the Eumetazoa: testing ecological predictions of molecular clocks against the Proterozoic fossil record. Proc Natl Acad Sci USA 102:9547–9552

Peterson KJ, Eernisse DJ (2001) Animal phylogeny and the ancestry of bilaterians: inferences from morphology and 18S rDNA gene sequences. Evol Dev 3:170–205

Pett W, Adamski M, Adamska M, Francis WR, Eitel M, Pisani D, Wörheide G (2019) The role of homology and orthology in the phylogenomic analysis of metazoan gene content. Mol Biol Evol 36:643–649

Philippe H, Derelle R, Lopez P, Pick K, Borchiellini C, Boury-Esnault N, Vacelet J, Renard E, Houliston E, Quéinnec E et al (2009) Phylogenomics revives traditional views on deep animal relationships. Curr Biol 19:706–712

Philippe H, Brinkmann H, Lavrov DV, Littlewood DTJ, Manuel M, Wörheide G, Baurain D (2011) Resolving difficult phylogenetic questions: why more sequences are not enough. PLoS Biol 9:e1000602

Pick KS, Philippe H, Schreiber F, Erpenbeck D, Jackson DJ, Wrede P, Wiens M, Alié A, Morgenstern B, Manuel M et al (2010) Improved phylogenomic taxon sampling noticeably affects nonbilaterian relationships. Mol Biol Evol 27:1983–1987

Pisani D, Pett W, Dohrmann M, Feuda R, Rota-Stabelli O, Philippe H, Lartillot N, Wörheide G (2015) Genomic data do not support comb jellies as the sister group to all other animals. Proc Natl Acad Sci USA 112:15402–15407

Ramulu HG, Raoult D, Pontarotti P (2012) The rhizome of life: what about metazoa? Front Cell Infect Microbiol 2:50

Renard E, Vacelet J, Gazave E, Lapébie P, Borchiellini C, Ereskovsky AV (2009) Origin of the neuro-sensory system: new and expected insights from sponges. Integr Zool 4:294–308

Renard E, Eve G, Fierro-Constain L, Schenkelaars Q, Ereskovsky A, Vacelet JV, Borchiellini CB (2013). Porifera (Sponges): recent knowledge and new perspectives. In eLS, John Wiley & Sons, Ltd (Ed.). https://doi.org/10.1002/9780470015902.a0001582.pub2

Renard E, Leys SP, Wörheide G, Borchiellini C (2018) Understanding animal evolution: the added value of sponge transcriptomics and genomics. BioEssays 40:1700237

Rentzsch F, Layden M, Manuel M (2017) The cellular and molecular basis of cnidarian neurogenesis. Wiley Interdisc Rev Dev Biol 6:e257

Revilla-I-Domingo R, Schmidt C, Zifko C, Raible F (2018) Establishment of transgenesis in the demosponge *Suberites domuncula*. Genetics 210:435–443

Reynolds AS (2019) Ernst Haeckel and the philosophy of sponges. Theory Biosci

Richards GS, Degnan BM (2009) The dawn of developmental signaling in the metazoa. Cold Spring Harb Symp Quant Biol 74:81–90

Richards GS, Degnan BM (2012) The expression of delta ligands in the sponge *Amphimedon queenslandica* suggests an ancient role for Notch signaling in metazoan development. Evodevo 3:15

Richards GS, Simionato E, Perron M, Adamska M, Vervoort M, Degnan BM (2008) Sponge genes provide new insight into the evolutionary origin of the neurogenic circuit. Curr Biol 18:1156–1161

Richter DJ, Fozouni P, Eisen MB, King N (2018) Gene family innovation, conservation and loss on the animal stem lineage. Elife 7

Riesgo A, Farrar N, Windsor PJ, Giribet G, Leys SP (2014) The analysis of eight transcriptomes from all poriferan classes reveals surprising genetic complexity in sponges. Mol Biol Evol 31:1102–1120

Rivera A, Winters I, Rued A, Ding S, Posfai D, Cieniewicz B, Cameron K, Gentile L, Hill A (2013) The evolution and function of the Pax/Six regulatory network in sponges. Evol Dev 15:186–196

Rivera AS, Hammel JU, Haen KM, Danka ES, Cieniewicz B, Winters IP, Posfai D, Wörheide G, Lavrov DV, Knight SW et al (2011) RNA interference in marine and freshwater sponges: actin knockdown in *Tethya wilhelma* and *Ephydatia muelleri* by ingested dsRNA expressing bacteria. BMC Biotechnol 11:67

Russell JJ, Theriot JA, Sood P, Marshall WF, Landweber LF, Fritz-Laylin L, Polka JK, Oliferenko S, Gerbich T, Gladfelter A, et al (2017) Non-model model organisms. BMC Biol 15(1):55

Ryan JF (2014) Did the ctenophore nervous system evolve independently? Zoology (Jena) 117:225–226

Ryan JF, Chiodin M (2015) Where is my mind? How sponges and placozoans may have lost neural cell types. Philos Trans R Soc Lond B Biol Sci 370

Sakarya O, Armstrong KA, Adamska M, Adamski M, Wang I-F, Tidor B, Degnan BM, Oakley TH, Kosik KS (2007) A post-synaptic scaffold at the origin of the animal kingdom. PLoS ONE 2:e506

Schenkelaars Q, Fierro-Constain L, Renard E, Hill AL, Borchiellini C (2015) Insights into Frizzled evolution and new perspectives. Evol Dev 17:160–169

Schenkelaars Q, Fierro-Constain L, Renard E, Borchiellini C (2016a) Retracing the path of planar cell polarity. BMC Evol Biol 16:69

Schenkelaars Q, Quintero O, Hall C, Fierro-Constain L, Renard E, Borchiellini C, Hill AL (2016b) ROCK inhibition abolishes the establishment of the aquiferous system in *Ephydatia muelleri* (Porifera, Demospongiae). Dev Biol 412:298–310

Schenkelaars Q, Pratlong M, Kodjabachian L, Fierro-Constain L, Vacelet J, Le Bivic A, Renard E, Borchiellini C (2017) Animal multicellularity and polarity without Wnt signaling. Sci Rep 7:15383

Schippers KJ, Nichols SA, Wittkopp P (2018) Evidence of Signaling and adhesion roles for β-Catenin in the sponge *Ephydatia muelleri*. Mol Biol Evol 35:1407–1421

Sebé-Pedrós A, Ruiz-Trillo I (2017) Evolution and classification of the T-Box transcription factor family. Curr Top Dev Biol 122:1–26

Sebé-Pedrós A, Ariza-Cosano A, Weirauch MT, Leininger S, Yang A, Torruella G, Adamski M, Adamska M, Hughes TR, Gómez-Skarmeta JL et al (2013) Early evolution of the T-box transcription factor family. Proc Natl Acad Sci USA 110:16050–16055

Sebé-Pedrós A, Degnan BM, Ruiz-Trillo I (2017) The origin of Metazoa: a unicellular perspective. Nat Rev Genet 18:498–512

Sebé-Pedrós A, Chomsky E, Pang K, Lara-Astiaso D, Gaiti F, Mukamel Z, Amit I, Hejnol A, Degnan BM, Tanay A (2018) Early metazoan cell type diversity and the evolution of multicellular gene regulation. Nat Ecol Evol 2:1176–1188

Simion P, Philippe H, Baurain D, Jager M, Richter DJ, Di Franco A, Roure B, Satoh N, Quéinnec É, Ereskovsky A et al (2017) A large and consistent phylogenomic dataset supports sponges as the sister group to all other animals. Curr Biol 27:958–967

Simpson TL (1984) The cell biology of sponges. Springer-Verlag New York Inc.

Slack JM, Holland PW, Graham CF (1993) The zootype and the phylotypic stage. Nature 361:490–492

Sogabe S, Nakanishi N, Degnan BM (2016) The ontogeny of choanocyte chambers during metamorphosis in the demosponge *Amphimedon queenslandica*. EvoDevo 7:6. https://doi.org/10.1186/s13227-016-0042-x

Sperling EA, Pisani D, Peterson KJ (2007) Poriferan paraphyly and its implications for Precambrian palaeobiology. Geol Soc 286:355–368 London, Special Publications

Sperling EA, Peterson KJ, Pisani D (2009) Phylogenetic-signal dissection of nuclear housekeeping genes supports the paraphyly of sponges and the monophyly of Eumetazoa. Mol Biol Evol 26:2261–2274

Srivastava M, Simakov O, Chapman J, Fahey B, Gauthier MEA, Mitros T, Richards GS, Conaco C, Dacre M, Hellsten U et al (2010) The *Amphimedon queenslandica* genome and the evolution of animal complexity. Nature 466:720–726

Stolfi A, Christiaen L (2012) Genetic and genomic toolbox of the chordate Ciona intestinalis. Genetics 192:55–66

Suga H, Ono K, Miyata T (1999) Multiple TGF-beta receptor related genes in sponge and ancient gene duplications before the parazoan-eumetazoan split. FEBS Lett 453:346–350

Suga H, Chen Z, de Mendoza A, Sebé-Pedrós A, Brown MW, Kramer E, Carr M, Kerner P, Vervoort M, Sánchez-Pons N et al (2013) The capsaspora genome reveals a complex unicellular prehistory of animals. Nat Commun 4:2325

Syvanen M (1984) The evolutionary implications of mobile genetic elements. Annu Rev Genet 18:271–293

Tickle C, Urrutia AO (2017) Perspectives on the history of evo-devo and the contemporary research landscape in the genomics era. Philos Trans Roy Soc B: Biol Sci 372:20150473

Tompkins-Macdonald GJ, Gallin WJ, Sakarya O, Degnan B, Leys SP, Boland LM (2009) Expression of a poriferan potassium channel: insights into the evolution of ion channels in metazoans. J Exp Biol 212:761–767

Ueda N, Richards GS, Degnan BM, Kranz A, Adamska M, Croll RP, Degnan SM (2016) An ancient role for nitric oxide in regulating the animal pelagobenthic life cycle: evidence from a marine sponge. Sci Rep 6:37546

Uzzell T, Corbin KW (1971) Fitting discrete probability distributions to evolutionary events. Science 172:1089–1096

Voigt O, Adamski M, Sluzek K, Adamska M (2014) Calcareous sponge genomes reveal complex evolution of α-carbonic anhydrases and two key biomineralization enzymes. BMC Evol Biol 14:230

Wells GD, Tang Q-Y, Heler R, Tompkins-MacDonald GJ, Pritchard EN, Leys SP, Logothetis DE, Boland LM (2012) A unique alkaline pH-regulated and fatty acid-activated tandem pore domain potassium channel (K2P) from a marine sponge. J Exp Biol 215:2435–2444

Whelan NV, Kocot KM, Halanych KM (2015) Employing phylogenomics to resolve the relationships among Cnidarians, Ctenophores, Sponges, Placozoans, and Bilaterians. Integr Comp Biol 55:1084–1095

Whelan NV, Kocot KM, Moroz TP, Mukherjee K, Williams P, Paulay G, Moroz LL, Halanych KM (2017). Ctenophore relationships and their placement as the sister group to all other animals. Nat Ecol Evol 1(11):1737-1746. Epub 2017 Oct 9. https://doi.org/10.1038/s41559-017-0331-3

Whittaker RH (1959) On the broad classification of organisms. Q Rev Biol 34:210–226

Windsor PJ, Leys SP (2010) Wnt signaling and induction in the sponge aquiferous system: evidence for an ancient origin of the organizer. Evol Dev 12:484–493

Windsor Reid PJ, Matveev E, McClymont A, Posfai D, Hill AL, Leys SP (2018) Wnt signaling and polarity in freshwater sponges. BMC Evol Biol 18:12

Zrzavý J, Mihulka S, Kepka P, Bezděk A, Tietz D (1998) Phylogeny of the Metazoa based on morphological and 18S ribosomal DNA evidence. Cladistics 14:249–285

Chapter 8
A Geometric Morphometrics-Based Mapping Model of Leaf Shape Evolution

Yige Cao, Xuli Zhu, Rongling Wu and Lidan Sun

Abstract Understanding how leaf shape diversifies and evolves requires the genetic information that encodes morphological variation within and between species. Genetic mapping provides a unique way for the dissection of leaf shape into its underlying Mendelian genes, known as quantitative trait loci (QTLs), but this approach relies on how leaf shape is quantified. In this chapter, we describe a computational approach that integrates geometric morphometrics (GM)-based shape analysis and principal component analysis (PCA) into the statistical framework of genetic mapping. This approach can map how QTLs affect both global and detailed variation of leaf shape. We demonstrate its utility by applying it to the genetic mapping study of a woody plant—mei (*Prunus mume* Sieb. et Zucc). This approach could serve as a powerful tool to understand the genetic basis underlying the natural variation and evolution of leaf shape in plants.

8.1 Introduction

Because leaves play a pivotal role in the success of plants, studies of leaf anatomy, morphology and function have been one of the most active and important areas in plant biology for centuries (Micol 2009; Peppe et al. 2011). While considerable studies focus on leaf size and leaf physiology, their variation and correlations with yield traits (Wu et al. 1997) and life-history traits (Parkhusrt and Loucks 1972; Garnier and Laurent 1994), there has also been long recognition that leaf shape affects the photosynthetic capacity of leaves by regulating radiation absorption and gas exchange

Y. Cao · X. Zhu
Center for Computational Biology, College of Biological Sciences and Biotechnology, Beijing Forestry University, Beijing 100083, China

R. Wu · L. Sun (✉)
National Engineering Research Center for Floriculture, Beijing Forestry University, Beijing 100083, China
e-mail: sunlidan@bjfu.edu.cn

R. Wu
Center for Statistical Genetics, The Pennsylvania State University, Hershey, PA 17033, USA

P. Pontarotti (ed.), *Evolution, Origin of Life, Concepts and Methods*,
https://doi.org/10.1007/978-3-030-30363-1_8

(Givnish 1979; Nicotra et al. 2011) and can be an indicator for species differences and divergences and plant evolution (Peppe et al. 2011; Nicotra et al. 2011; Royer et al. 2009). Molecular experiments have identified several genes, such as CLAVATA1, CLAVATA3, WUSCHEL, KNOTTED1, and PHANTASTICA that are involved in leaf initiation at the shoot apical meristem, triggering a profound effect on leaf shape and size through mediating hormone gradients and transport (Kessler and Sinha 2004; White 2006; Barkoulas et al. 2007). In a recent literature review, Dkhar and Pareek (2014) synthesized several underlying factors that contribute to the formation process of leaf shape; e.g., PIN1 and KNOX1 as indicators of leaf initiation; HD-ZIPIII, KANADI, and YABBY related to leaf outgrowth; ANGUSTIFOLIA3 and GROWTH-REGULATING FACTOR5 mediating leaf expansion and maturation; and APUM23 as a newly identified player to specify leaf polarity. Many studies that dissect how these factors interact with environmental factors to determine leaf shape are under way (Tsukaya 2005; Royer et al. 2009; Peppe et al. 2011).

Despite these developments, however, a detailed understanding of the genetic architecture of leaf shape from genetic mapping is still lacking. One important reason is that current morphological analyses are either descriptive for leaf shape that is classified into distinct groups, such as round, long, or ovate, or measure shape variation based on length–width ratios and angles (Rohlf and Marcus 1993). These approaches, even as well as some more advanced approaches, cannot quantitatively assess numerous subtle but collectively important structures of leaf morphology, such as serrated or lobed leaf margins. Langlade et al. (2005) presented one of the first studies for mapping quantitative trait loci (QTLs) for leaf shape by using a geometric morphometric (GM) model. They identified 19 representative points for a leaf in *Antirrhinum* to capture the allometry of leaf shape and further applying this information to study interspecific divergence and evolution. More recently, Fu et al. (2010) have developed a binary model to map shape QTLs through modeling black and white shape data. A radius-centroid-contour analysis approach, implemented with a procrustes analysis for aligning shape with different poses, scales, and rotations has been developed by Fu et al. (2013) to describe morphological shape and map its underlying QTLs. The advantage of this approach, shape mapping, lies in its precise estimation of genetic effects on global and detailed aspects of shape variation. The use of shape recognition and shape registration based on the polar coordinates of anatomical landmarks from digitized data is not only tractable but also highly accurate for the measure of any shape. Bo et al. (2014) described an elliptic Fourier-based outline analysis approach to approximate the object's outline, showing its ability to characterize the control patterns of specific QTLs on the origin, properties, and functions of shape variation.

In this chapter, we provide a description of how the shape mapping model is used to map specific QTLs that are associated with leaf shape. We used an ornamental woody plant, mei (*Prunus mume* Sieb. et Zucc.), as the study material. Apart from its economic value, mei has been thought to play a pivotal role in systematic studies of the Rosaceae. Since its genome was sequenced (Zhang et al. 2012), mei has been increasingly used as an important material to study the comparative genomics of Rosaceae and map key genes that control growth and ornamental traits. Based on the

reference genome of mei, Sun et al. (2013a) characterized hundreds of thousands of SNPs between cultivars Fenban and Kouzi Yudie, among which a couple of thousands were genotyped for the F_1 hybrids of these two cultivars. These SNPs were added to the SSR linkage map, previously reported in Sun et al. (2013b), leading to the construction of a first high-density genetic map for mei. Several important QTLs that contribute to stem growth and growth-component traits in the young seedlings of mei have been identified and mapped to different regions of the mei genome using standard mapping approaches (2014).

8.2 Shape Mapping

8.2.1 Population Establishment

Sun et al. (2013a) reported a full-sib mapping population of 190 mei hybrids derived from the intraspecific cross between two cultivars, *P. mume* "Fenban" (BJFU1210120013) as the female parent and *P. mume* "Kouzi Yudie" (BJFU1210120022) as the male parent. This population was genotyped for 1484 SNP markers and InDels and grown in the Xiao Tangshan Horticultural Trial, Beijing, China. For a full-sib family derived from two heterozygous parents, a marker may be segregating according to either F_2-like intercross or backcross-like testcross (Wu et al. 2002; Lu et al. 2004). The markers genotyped include 271 intercross markers and 1213 testcross markers. Of these testcross markers, 620 and 593 are segregating due to the heterozygosity of parents Fenban and Kouzi Yudie, respectively. JoinMap version 4 (van Ooijen 2006) was used to construct a genetic linkage map from SSR and SNP markers. The map is composed of eight linkage groups corresponding to mei's eight haploid chromosomes. The total length of the map is 670 cm, with an average marker distance of 0.5 cm.

8.2.2 Image Analysis of Leaf Shape

All F_1 hybrids were cut off after one year of growth in the field. Next spring, new shoots sprout from one-year rooting systems. During the end of July, the fast-growing season of mei, photographs were taken for three representative leaves from each tree. These chosen leaves were located at the main stem with leaf plastochron index of 10–12. From 600×900 pixels from a leaf digital image, three matrices for red, green, and black colors were obtained to discern the object and background, from which we generated a smaller binary matrix on the leaf's contour. Using the procedure for shape alignment described in Fu et al. (2013), we obtained a vector of 360 coordinates on the boundary, which are equally angled to the centroid of the leaf. In the end, 360 representative points, expressed as a radius-centroid-contour (RCC)

curve, can represent the contour shape of a leaf. We took the mean of RCC curves over three leaves sampled from the same seedlings as a measure of leaf shape for this particular genotype. Principal component (PC) analysis of the aligned image data on mei leaves is incorporated to capture information about global and local variability in leaf shape.

8.2.3 QTL Mapping

The map density used in this study is dense enough to directly associate marker genotypes with PC values as phenotypic traits to detect significant QTLs. By formulating the null hypothesis (in which there is no QTL) and alternative hypothesis (in which there is a QTL) and then calculating the log-ratio of likelihoods (LR) between the two hypotheses, we scanned the linkage map to obtain an LR profile over genome position. We performed permutation tests to determine the genome-wide threshold. Through transforming the PC information back to image domains, an intuitive visualization on how QTLs affect shape variation is obtained.

The QTLs detected from intercross and testcross markers provide different amounts of genetic information. For the intercross QTLs, both additive effect and dominant effects can be estimated. Tong et al. (2011) provided a procedure for estimating the genetic effects of a multiallelic intercross QTL, which was used to map botanical traits in mei (Sun et al. 2014). The multiallelic intercross QTL allows us to estimate the allelic (additive) effects each contributed by a different parent and the dominant effect due to the interaction between alleles from the two different parents. For the testcross QTL, only the additive genetic effect can be estimated but from the parent that is heterozygous, we can determine the direction of the additive effect.

8.2.4 Results

From leaf image data, we extracted 360 coordinates along leaf boundaries using Fu et al.'s (2013) RCC approach, by which to build the shape of each leaf from the mapping population (Fig. 8.1a). After shape alignment, these coordinates were normalized as 360 spatially correlated RCC values for describing the shape of leaf contour (Fig. 8.1b). Such high-dimensional image data of leaf shape were then reduced by principal component analysis (PCA) into their low-dimensional representations. From PCA, we identified five major orthogonal axes, termed PCs, which together explained 85.1% of the variation among the samples. According to the percentages of variance, they explained, these five PCs are ordered as PC1, 42.3%; PC2, 21.7%; PC3, 11.1%; PC4, 5.7%; and PC5, 4.2%. By fitting them to raw data through the RCC curve, each of the PCs was found to describe a different aspect of leaf shape. PC1 tends to explain global variation of leaf shape, whereas PC2–PC5 are responsible for local variation.

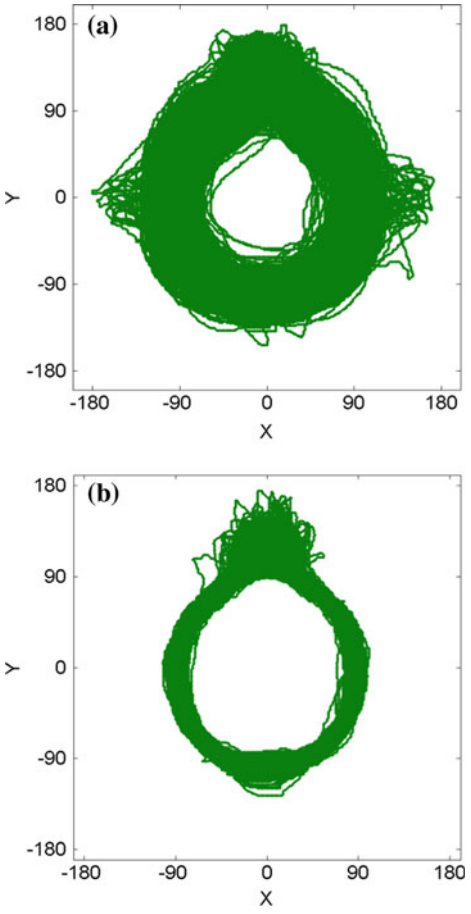


Fig. 8.1 Linking 360 coordinates on the leaf outlines for leaves of all sampled trees from the mapping population. In **a**, raw leaf shapes, showing variation in scale, position, and orientation. In **b**, this variation is removed from the objects through shape alignment

By associating individual markers with each PC, we obtained a plot of LR values over the genome. After the genome-wise critical thresholds were determined from permutation tests, we found identified 29, 2, and 8 significant QTLs associated with PC1, PC2, and PC5, respectively (Fig. 8.2). The locations and segregating types of these significant QTLs are given in Table 8.1, along with the proportions of the total phenotypic variance explained by each QTL. In general, PC1-associated QTLs tend to affect global leaf shape and PC2- and PC3-associated QTLs are responsible for local leaf shape. On average, global QTLs explain a larger proportion of the phenotypic variance (11.7%) than local QTLs (9.0% for PC2 and 6.0% for PC5) (Table 8.1). These two types of QTLs reside on different genomic locations; all global QTLs are located on linkage group 6 (Fig. 8.2a), whereas a majority of local QTLs are located on linkage group 2 (Fig. 8.2b and c). Only a local QTL was detected to locate on linkage group 6, but it is distant by over 15 cm from the nearest global QTL on the same group.

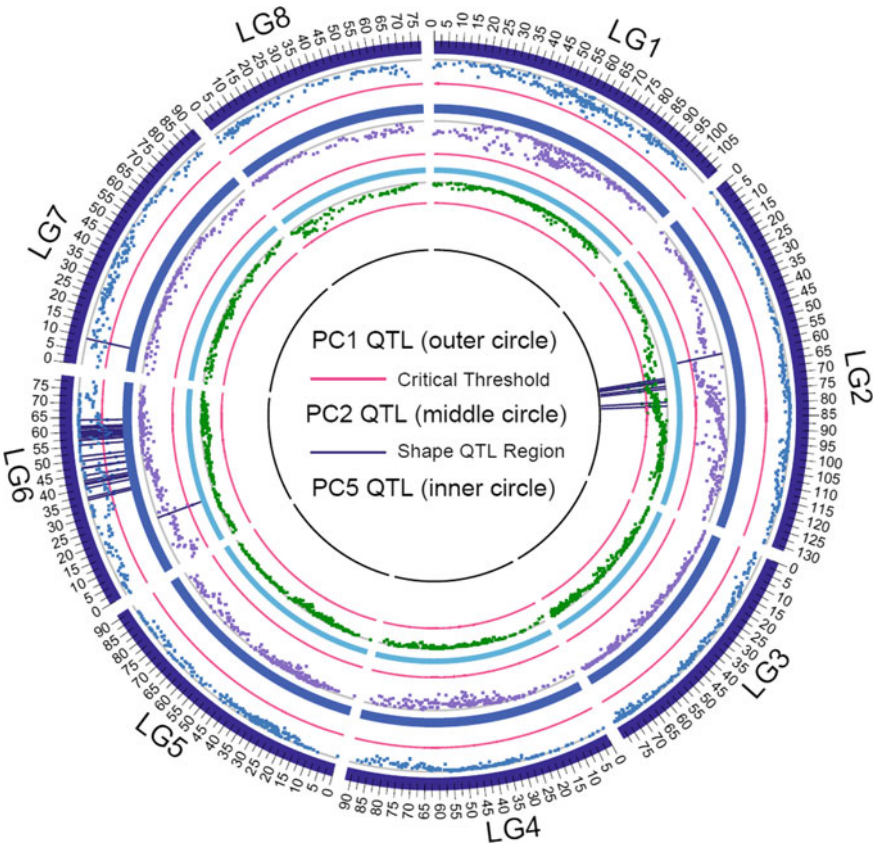


Fig. 8.2 The genetic map locations of QTLs that control PC1 (outer circle), PC2 (middle circle), and PC5 (inner circle) throughout eight linkage groups (LG) of the mei genome. Significant QTLs were claimed on the basis of the genome-wise critical threshold at the 5% significance level determined from permutation tests, which are highlighted by solid lines

Of the 29 global QTLs, 27 are the testcross type that are segregating due to the female parent Fenban, none is the testcross type due to parent Kouzi Yudie, and the remaining two are the intercross type due to both parents (Table 8.1). This suggests that the two cultivars used as crossing parents contribute unequally to phenotypic variation in the global shape of a mei leaf, with the most contribution by Fenban. Among the ten QTLs accounting for the local variation of leaf shape by PC2 and PC5, five belong to the testcross type derived from parent Kouzi Yudie, none is the testcross type due to parent Fenban, and five are the intercross due to both parents. This appears to suggest that cultivar Kouzi Yudie, has a higher likelihood to contribute to local variation in mei leaf shape.

Table 8.1 The information about 29 QTLs for PC1, 2 QTLs for PC2, and 8 QTLs for PC5 detected to affect leaf shape and their hypothetical functions in mei

No.	PCs	Linkage group	Marker ID	Marker type	Marker position	Genetic distance	R^2	BLAST annotation	GO ID
1	PC1	LG6	fheter_102659	Testcross	99698	36.803	0.108	Citrate transmembrane transporter activity; citrate transport; integral to membrane; transmembrane transport	GO:0015137 GO:0015746 GO:0016021 GO:0055085
2	PC1	LG6	fheter_102588	Testcross	35051	37.141	0.107	Protein binding	GO:0005515
3	PC1	LG6	hk_hk974	Intercross	22148	40.329	0.082	NA	
4	PC1	LG6	fheter_107052	Testcross	345215	41.991	0.102	Metabolic process; Hydrolase activity	GO:0008152; GO:0016787
5	PC1	LG6	fheter_109025	Testcross	203131	42.444	0.109	NA	
6	PC1	LG6	fheter_101454	Testcross	94868	43.315	0.110	NA	
7	PC1	LG6	fheter_109209	Testcross	68036	44.566	0.140	NA	
8	PC1	LG6	fheter_104977	Testcross	12007	47.397	0.119	NA	
9	PC1	LG6	fheter_107978	Testcross	224583	48.15	0.142	dolichyl-diphosphooligosaccharide-protein glycotransferase activity; endoplasmic reticulum membrane; protein N-linked glycosylation via asparagine	GO:0004579; GO:0005789; GO:0018279;
10	PC1	LG6	fheter_100576	Testcross	312844	49.845	0.132	NA	

(continued)

Table 8.1 (continued)

No.	PCs	Linkage group	Marker ID	Marker type	Marker position	Genetic distance	R^2	BLAST annotation	GO ID
11	PC1	LG6	fheter_105328	Testcross	956937	51.493	0.118	tRNA binding; nucleotide binding; aminoacyl-tRNA ligase activity; phenylalanine-tRNA ligase activity; ATP binding; cytoplasm; translation; tRNA aminoacylation for protein translation; phenylalanyl-tRNA aminoacylation; tRNA aminoacylation	GO:0000049; GO:0000166; GO:0004812; GO:0004826; GO:0005524; GO:0005737; GO:0006412; GO:0006418; GO:0006432; GO:0043039;
12	PC1	LG6	fheter_107288	Testcross	397840	55.461	0.117	DNA binding; protein binding; zinc ion binding	GO:0003677; GO:0005515; GO:0008270;
13	PC1	LG6	fheter_107551	Testcross	106038	57.238	0.107	farnesyl-diphosphate farnesyltransferase activity; lipid biosynthetic process; biosynthetic process; integral to membrane; transferase activity; transferase activity, transferring alkyl or aryl (other than methyl) groups	GO:0004310; GO:0008610; GO:0009058; GO:0016021; GO:0016740; GO:0016765;
14	PC1	LG6	fheter_108638	Testcross	328544	57.555	0.115	NA	
15	PC1	LG6	fheter_99076	Testcross	68965	58.327	0.115	RNA binding	GO:0003723
16	PC1	LG6	fheter_109978	Testcross	184584	58.751	0.110	NA	
17	PC1	LG6	fheter_102299	Testcross	8093	58.993	0.118	NA	

(continued)

Table 8.1 (continued)

No.	PCs	Linkage group	Marker ID	Marker type	Marker position	Genetic distance	R^2	BLAST annotation	GO ID
18	PC1	LG6	fheter_109114	Testcross	129133	59.157	0.130	NA	
19	PC1	LG6	fheter_99697	Testcross	64826	59.434	0.114	NA	
20	PC1	LG6	fheter_98857	Testcross	49766	59.848	0.134	Nucleotide binding; ATP binding; membrane; ATPase activity; nucleoside-triphosphatase activity	GO:0000166; GO:0005524; GO:0016020; GO:0016887; GO:0017111
21	PC1	LG6	fheter_98306	Testcross	14000	60.035	0.143	NA	
22	PC1	LG6	fheter_103257	Testcross	110753	60.152	0.132	NA	
23	PC1	LG6	fheter_106303	Testcross	143815	60.591	0.119	GTP cyclohydrolase II activity; 3,4-dihydroxy-2-butanone-4-phosphate synthase activity; riboflavin biosynthetic process	GO:0003935; GO:0008686; GO:0009231;
24	PC1	LG6	fheter_110074	Testcross	34563	61.24	0.126	NA	
25	PC1	LG6	fheter_109983	Testcross	33410	61.648	0.113	NA	
26	PC1	LG6	fheter_105228	Testcross	78543	61.843	0.119	transmembrane receptor activity; ATP binding; apoptosis; defense response; signal transduction; intrinsic to membrane; innate immune response	GO:0004888; GO:0005524; GO:0006915; GO:0006952; GO:0007165; GO:0031224; GO:0045087;
27	PC1	LG6	fheter_102485	Testcross	135252	62.125	0.107	NA	
28	PC1	LG6	fheter_104442	Testcross	84547	64.446	0.120	NA	

(continued)

Table 8.1 (continued)

No.	PCs	Linkage group	Marker ID	Marker type	Marker position	Genetic distance	R^2	BLAST annotation	GO ID
29	PC1	LG7	hk_hk271	Intercross	511516	9.705	0.079	oxidoreductase activity, acting on CH–OH group of donors; long-chain-alcohol oxidase activity; flavin adenine dinucleotide binding; oxidation–reduction process	GO:0016614; GO:0046577; GO:0050660; GO:0055114;
Mean R^2							0.117		
30	PC2	LG2	mheter_109820	Testcross	19601	59.938	0.106	Binding; protein binding; nuclear pore; nucleocytoplasmic transport; structural constituent of nuclear pore	GO:0005488; GO:0005515; GO:0005643; GO:0006913; GO:0017056;
31	PC2	LG6	hk_hk2360	Intercross	226576	22.265	0.092	NA	
Mean R^2							0.099		
32	PC5	LG2	mheter_109155	Testcross	28621	66.155	0.056	Protein domain specific binding	GO:0019904;
33	PC5	LG2	mheter_109867	Testcross	1290780	66.575	0.041	Intracellular; zinc ion binding	GO:0005622; GO:0008270;
34	PC5	LG2	mheter_111831	Testcross	1351344	66.924	0.047	NA	

(continued)

Table 8.1 (continued)

No.	PCs	Linkage group	Marker ID	Marker type	Marker position	Genetic distance	R^2	BLAST annotation	GO ID
35	PC5	LG2	hk_hk789	Intercross	971400	67.24	0.078	Exocyst; vesicle docking involved in exocytosis	GO:0000145; GO:0006904
36	PC5	LG2	mheter_107786	Testcross	127438	69.533	0.032	NA	
37	PC5	LG2	hk_hk1742	Intercross	14970	70.673	0.079	NA	
38	PC5	LG2	hk_hk552	Intercross	43976	77.835	0.078	NA	
39	PC5	LG2	hk_hk3135	Intercross	126153	79.822	0.069	NA	
Mean R^2							0.060		

R^2: Proportion of the total phenotypic variance explained by a single QTL

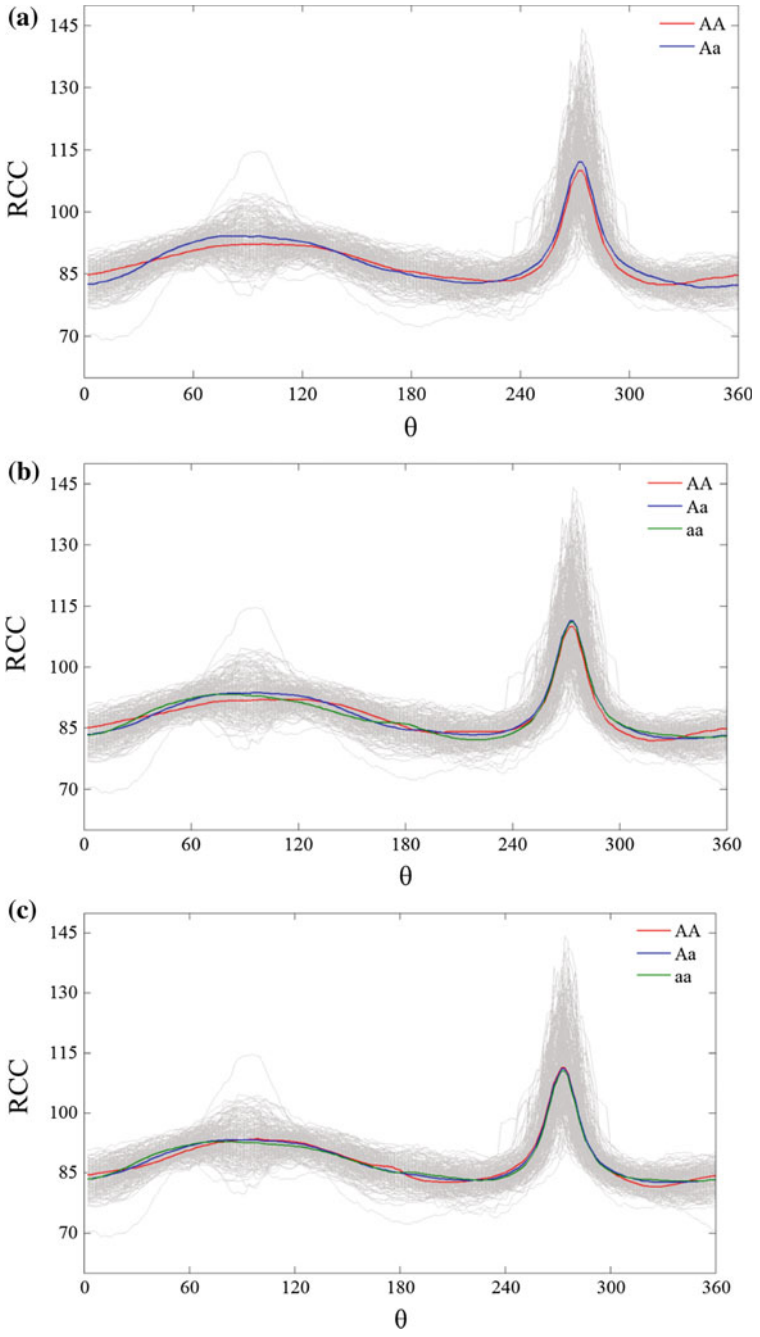


Fig. 8.3 RCC curves of different genotypes at a QTL detected. RCC curves of leaf shape as a function of radial angle θ at the centroid, curve **a** explained by the PC1, at the QTL detected by marker fheter_107978; curve **b** explained by the PC2, at the QTL detected by marker hk_hk2360; and curve **c** explained by the PC5, at the QTL detected by marker hk_hk789

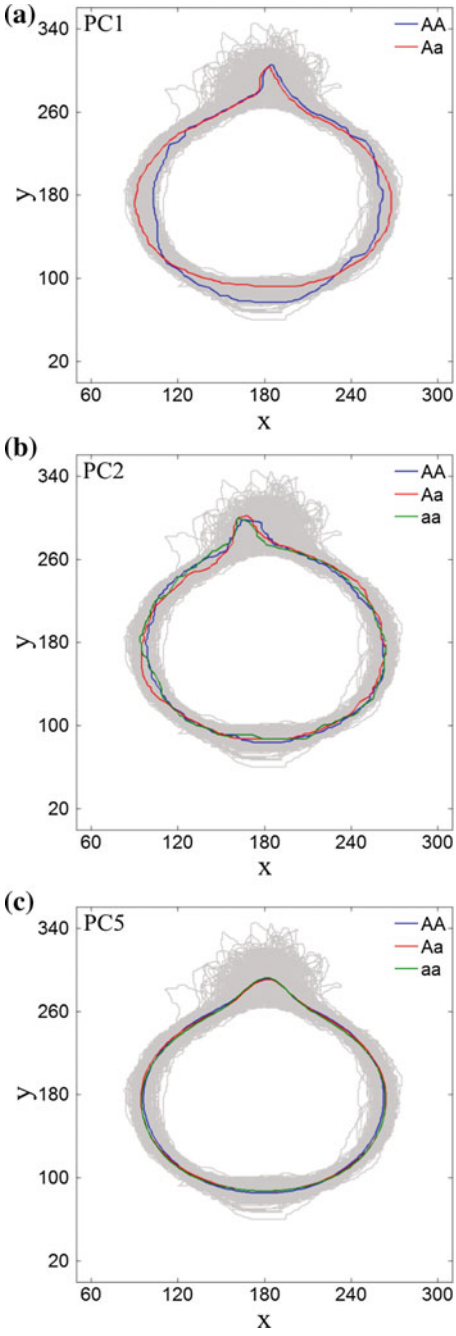


Fig. 8.4 Visualized variation in leaf shape in mei explained by a QTL detected. The different genotype at different QTLs control the leaf shape is shown. Figure **a** at the QTL detected by marker fheter_107978; Figure **b** at the QTL detected by marker hk_hk2360; and Figure **c** at the QTL detected by marker hk_hk789

To demonstrate how a QTL affects leaf shape, we plot RCC curves of PC values for each genotype at a QTL detected as a function of radial angle (θ) at the centroid. Figure 8.3 gives three such examples each for a QTL affecting a different PCs. The global QTL detected by marker fheter_107978 is a testcross type at which there are two genotypes denoted as *AA* and *Aa*. The RCC curves of these two genotypes differ in a wide range of θ, especially from $\theta = 0$ to 120° (Fig. 8.3a), confirming that this QTL governs global variation of leaf shape. As like in Fu et al. (2013), we transformed the RCC curves of PC1 back to leaf shape for each genotype at this QTL to visualize the detail of genetic control over leaf shape (Figs. 8.4a). It is found that two genotypes have dramatically different leaf shapes, with *Aa* being more rounded but *AA* being more elongate, despite of their similar size in terms of leaf area.

For a PC2-associated local QTL, detected by marker hk_hk2360, and a PC5-assocated local QTL, detected by marker hk_hk789, both of which are an intercross type, we found that three genotypes at each differ slightly in a few particular regions of RCC profile (Fig. 8.3b and c). By transforming the RCC curves back to leaf shape, overall differences in leaf shape among different genotypes can be visualized. The former QTL basically affects variation in the left part of leaf shape (Fig. 8.4b), whereas three genotypes of the latter QTL appear to be visually similar in the entire contour of leaf shape (Fig. 8.4c). This result suggests that, although PC2 and PC5 determine local leaf variation, PC5 can discern more subtle variation than PC2. In the other word, while marked changes of leaf shape are determined by QTLs for PC1 and QTLs for PC2 and PC5 modify the detailed structure of leaf shape, of which PC5-associated QTLs assist PC2-assocaited QTLs in this modification. It is shown that shape mapping has a powerful capacity of not only charting an overall picture of leaf shape variation explained by individual QTLs, but also providing detailed information about which part of leaf shape a QTL determines.

Many of the leaf shape QTLs identified are within the candidate genes that regulate particular biological functions (Table 8.1). Two global-shape QTLs detected by marker fheter_102659 and fheter_105228 are related to transmembrane transport activity. Many other QTLs for global leaf shape are associated with genes for hydrolase activity, dolichyl- diphosphooligosaccharide-protein glycotransferase activity, tRNA binding, ATP binding, and farnesyl-diphosphate farnesyltransferase activity, etc. It is interesting to note that four local-shape QTLs also map to genes with roles in protein and ion binding and exocytosis. Several QTLs with which no related genes have been identified merit further investigation.

8.3 Discussion

A detailed understanding of an organism's anatomical and shape feature can help to enhance the mechanistic relationships between organ structure and function. One of the most conspicuous aspects of these relationships in plants is leaf shape as well as its impact on plant production and fitness. Because of this, many studies have begun to study the molecular mechanisms of leaf shape by identifying many

genes and pathways that regulate variation in leaf shape (e.g., Kessler and Sinha 2004; Dkhar and Pareek 2014). Earlier quantitative geneticists used the ratio of leaf length to width as a surrogate to measure leaf shape and found that this trait is under strong genetic control (Wu et al. 1997). Langlade et al. (2005) reported the first study for genetic mapping of shape variation, in which they dissected and quantified phenotypic variation in leaf shape through a geometric morphometric (GM) approach.

To precisely chart the genetic architecture of morphological shape, Wu and group have for the first time integrated statistical shape analysis into a framework of QTL mapping to develop a new mapping approach, called shape mapping (Fu et al. 2010, 2013; Wang et al. 2012; Bo et al. 2014; Sun et al. 2014). Shape mapping has proven of great power to illustrate both global and local variation of an organ with arbitrary shape and explain each of these two types of variation by specific genes. Subsequent principal component analysis followed by image reading and shape registration allows different aspects of a shape to be characterized. Thus, the control patterns of specific QTLs on the origin, properties, and functions of leaf shape can be quantified. Furthermore, shape mapping, as a new tool for mapping complex traits, can readily find its widespread application in practice, because phenotyping by taking photographs is one of the simplest and most precise approaches for phenotypic measurement.

As an attempt to use an advanced computational model to study the genetic control of botanical traits in a full-sib family of an ornamental woody plant, mei, we have identified and mapped over three dozens of significant QTLs for leaf shape. More than two-thirds of the QTLs detected affect global variation of leaf shape, whereas less than one-third are related to local variation. These two types of QTLs are located on different chromosomes. Also, it is notable that a majority of global-shape QTLs are segregating due to the heterozygosity of the female parent Fenban which has a more rounded leaf blade, whereas almost all local-shape QTLs are derived from the male parent Kouzi Yudie with a longer tip. There are also a few QTLs that are segregating due to the heterozygosity of both parents. More contributions to overall and local shape by alleles from Fenban and Kouzi Yudie, respectively, are consistent with the differences of these two parents in key aspects of leaf shape. By GO analysis, we have identified candidate genes of different functions associated with most shape QTLs detected, well explaining, and validating the biological relevance of shape mapping.

In recent years, high-through phenotyping techniques have been developed to not only monitor the dynamic change of growth and development, but also detail the subtle variation of morphology and anatomy (Yang et al. 2014). As one major component of these techniques, digital image processing can provide a complete description of the characteristics of leaf morphology and development. Shape mapping, derived by integrating shape-based image analysis into genetic association and mapping studies, holds a great promise to shed light on the genetic architecture of complex living leaves. However, there is much room for shape mapping to be improved as a more useful tool. First, leaf shape, as a biological trait, undergoes a developmental change

across a time, space, and signal; for example, leaves often change their shape dramatically from one stage of growth to next, from a higher position to lower position within the same plant, and from one environment to others (Royer et al. 2009). Such phenomena of shape ontogeny, heteroblasty and heterophylly (Costa et al. 2012; Sun et al. 2014), and phenotypic plasticity (Royer et al. 2009) should be integrated into shape mapping to better unravel the genetic and developmental machineries of leaf growth and plant fitness.

Second, the current computational model for shape mapping focuses on the identification of QTLs for 2D shape variation. However, leaf shape, ranging from flat to curved or ruffled, can emerge in a 3D surface. More recently, Rolland-Lagan et al. (2014) have developed a model to quantify tissue deformation and surface shape changes, which can be implemented into shape mapping to better study the genetic basis for the interplay between leaf growth and shape. Third, clonal analysis techniques have been developed to analyze and interpret the geometric shape of a leaf by studying the relationship between gene expression pattern and leaf shape (Nath et al. 2003). Using a combination of scaling, stability, and asymptotic analysis, Liang and Mahadevan (2009) can quantify leaf shape and investigate the conditions that cause leaf morphological differences. The integration of all these technologies can empower shape mapping to chart a complete picture of the genetic control mechanisms underlying complex traits.

Acknowledgements This work is supported by grants National Natural Science Foundation of China (Grant No.31401900) and Special Fund for Forest Scientific Research in the Public Welfare (201404102).

References

Bo WH, Wang Z, Xu F, Fu GF, Sui YH, Wu WM, Zhu XL, Yin DN, Yan Q, Wu RL (2014) Shape mapping: Genetic mapping meets geometric morphometrics. Brief Bioinform 15:571–581

Barkoulas M, Galinha C, Grigg SP, Tsiantis M (2007) From genes to shape: regulatory interactions in leaf development. Curr Opin Plant Biol 10:660–666

Costa MM, Yang S, Critchley J, Feng X, Wilson Y, Langlade N, Copsey L, Hudson A (2012) The genetic basis for natural variation in heteroblasty in Antirrhinum. New Phytol 196:1251–1259

Dkhar J, Pareek A (2014) What determines a leaf's shape? EvoDevo 5:47

Fu GF, Berg A, Das K, Li JH, Li R, Wu RL (2010) A statistical model for mapping morphological shape. Theor Biol Med Model 7:28

Fu GF, Bo WB, Pang XM, Wang Z, Chen L, Song YP, Zhang ZY, Li J, Wu RL (2013) Mapping shape QTLs using a radius-centroid-contour model. Heredity 110:511–519

Garnier E, Laurent G (1994) Leaf anatomy, specific mass and water content in congeneric annual and perennial grass species. New Phytol 128:725–736

Givnish T (1979) On the adaptive significance of leaf form. In: Solbrig OT, Jain S, Johnson GB, Raven PH (eds) Topics in plant population Biology. Columbia University Press, New York

Kessler S, Sinha N (2004) Shaping up: the genetic control of leaf shape. Curr Opin Plant Biol 7:65–72

Langlade NB, Feng X, Dransfield T, Copsey L, Hanna AI, Thébaud C, Bangham A, Hudson A, Coen E (2005) Evolution through genetically controlled allometry space. Proc Natl Acad Sci U S A 102:10221–10226

Liang H, Mahadevan L (2009) The shape of a long leaf. Proc Natl Acad Sci U S A 106:22049–22054

Lu Q, Cui YH, Wu RL (2004) A multilocus likelihood approach to joint modeling of linkage, parental diplotype and gene order in a full-sib family. BMC Genet 5:20

Micol JL (2009) Leaf development: time to turn over a new leaf? Curr Opin Plant Biol 12:9–16

Nath U, Crawford BCW, Carpenter R, Coen E (2003) Genetic control of surface curvature. Science 299:1404–1407

Nicotra AB, Leigh A, Boyce K, Jones CS, Niklas KJ, Royer DL, Tsukaya H (2011) The evolution and functional significance of leaf shape in the angiosperms. Funct Plant Biol 38:535–552

Parkhurst DF, Loucks OL (1972) Optimal leaf size in relation to environment. J Ecol 60:505–537

Peppe DJ, Royer DL, Cariglino B, Oliver SY et al (2011) Sensitivity of leaf size and shape to climate: global patterns and paleoclimatic applications. New Phytol 190:724–739

Rohlf FJ, Marcus LF (1993) A revolution in morphometrics. Trends Ecol Evol 8:129–132

Rolland-Lagan AG, Remmler L, Girard-Bock C (2014) Quantifying shape changes and tissue deformation in leaf development. Plant Physiol 165:496–505

Royer DL, Meyerson LA, Robertson KM, Adams JM (2009) Phenotypic plasticity of leaf shape along a temperature gradient in *Acer rubrum*. PLoS ONE 4:e7653

Sun L, Yang W, Zhang Q, Cheng T, Pan H, Xu Z, Zhang J, Chen C (2013a) Genome-wide characterization and linkage mapping of simple sequence repeats in mei (*Prunus mume* Sieb. et Zucc.). PLoS ONE 8: e59562

Sun L, Zhang Q, Xu Z, Yang W, Guo Y, Lu J, Pan H, Chneg T, Cai M (2013b) Genome-wide DNA polymorphisms in two cultivars of mei (*Prunus mume* sieb. et zucc.). BMC Genet 14: 98

Sun LD, Wang YQ, Yan XL, Cheng TR, Ma KF, Yang WR, Pan HT, Zhu XL, Zheng CF, Wang J, Wu RL, Zhang QX (2014) Genetic control of juvenile growth and botanical architecture in an ornamental woody plant, *Prunus mume* Sieb. et Zucc., as revealed by a high-density linkage map. BMC Genet 15: S1

Tong CF, Wang Z, Zhang B, Shi JS, Wu RL (2011) 3FunMap: full-sib family functional mapping of dynamic traits. Bioinformatics 27:2006–2008

Tsukaya H (2005) Leaf shape: genetic controls and environmental factors. Inl J Dev Biol 49:547–555

van Ooijen JW (2006) JoinMap 4.0, software for the calculation of genetic linkage maps. Kyazama B.V, Wageningen, The Netherlands

Wang NT, Wang YQ, Wang Z, Han H, Wu RL (2012) Mapping body shape genes through shape mapping. J Biomet Biostat 3:e121

White DWR (2006) PEAPOD regulates lamina size and curvature in Arabidopsis. Proc Natl Acad Sci U S A 103:13238–13243

Wu R, Bradshaw HD, Stettler RF (1997) Molecular genetics of growth and development in Populus. V. mapping quantitative trait loci affecting leaf variation. Am J Bot 84:143–153

Wu RL, Ma CX, Painter I, Zeng ZB (2002) Simultaneous maximum likelihood estimation of linkage and linkage phases in outcrossing species. Theor Pop Biol 61:349–363

Yang W, Guo Z, Huang C, Duan L, Chen G, Jiang N, Fang W, Feng H, Xie W, Lian X, Wang G, Luo Q, Zhang Q, Liu Q, Xiong L (2014) Combining high-throughput phenotyping and genome-wide association studies to reveal natural genetic variation in rice. Nat Comm 5:5087

Zhang Q, Chen W, Sun L, Zhao F, Huang B et al (2012) The genome of *Prunus mume*. Nat Comm 3:1318

Chapter 9
Allostery and Structural Dynamics in Protein Evolution

P. Campitelli and S. B. Ozkan

Abstract The sequence-structure-function paradigm for proteins has been well accepted in the scientific community. However, with the emergence of the ensemble view of proteins, it has become clear that this fails to incorporate the importance of protein dynamics, which we now understand govern the underlying function of proteins. Here, we introduce two tools—the dynamic flexibility index (DFI) and the dynamic coupling index (DCI) which can quantify structural flexibility and dynamic coupling at a site-specific, single amino acid level. We show that it is possible to relate evolutionary conservation to amino acid flexibility and that disease-associated protein variants coincide with mutations at rigid positions within a protein. In combination with experimental data, we also capture important changes in structural conformation that coincide with the evolution of protein function through changes in structural dynamics. Finally, we discuss how nature can modulate change through allosteric mutations which alter the internal interaction network of proteins, and how changes in allosteric regulation can result in disease phenotypes.

9.1 Protein Structural Dynamics Govern Function

The discovery of the protein structure proved to be one of the major scientific achievements in biological sciences and marked a milestone in the field, spawning the sequence-structure-function paradigm. That is, how can the one-dimensional set of information from the amino acid sequence give rise to a unique three-dimensional structure which, in turn, determines the function of a given protein. Major breakthroughs followed in experimental biophysics, including the ability to capture high-resolution images of protein structures with techniques using x-ray crystallography or NMR spectroscopy. Advances were also made in computational and theoretical biophysics as well as bioinformatics, giving rise to the field of structural prediction; the determination of a protein's three-dimensional structure using only the

P. Campitelli · S. B. Ozkan (✉)
Department of Physics and Center for Biological Physics, Arizona State University, Tempe, AZ 85287, USA
e-mail: Banu.Ozkan@asu.edu

P. Pontarotti (ed.), *Evolution, Origin of Life, Concepts and Methods*,
https://doi.org/10.1007/978-3-030-30363-1_9

amino acid sequence and development of structure-based databases such as CATH (Knudsen and Wiuf 2010) to annotate function. This drive even resulted in a unique biannual competition in structural prediction called the Critical Assessment of Structural Prediction (CASP) with the purpose of assessing current structural prediction methods and techniques. However, it quickly became apparent that it was not so simple as assigning a single three-dimensional structure to an amino acid sequence. Some proteins which bind to ligands could take on multiple, distinct conformations depending on the binding event, sometimes resulting in conformations very different from one another. Additionally, with advances in sequencing techniques, a large database of protein structures was created containing many homologs; proteins which had different amino acid sequences but shared the same three-dimensional structure, or even specific domains of proteins present across protein families.

These discoveries helped give rise to the ensemble picture of proteins. That is, proteins are not simple static objects but tend to sample many conformational states even in the absence of ligand binding. Additionally, point mutation studies, where a single amino acid would be mutated from the wild-type structure, showed that it was possible for a protein to exhibit dramatic changes in function even in the absence of any significant structural changes. Through a vast body of work and contributions of many members in various fields throughout the scientific community, it became clear that the important functions of proteins were not solely a result of the structure, but of *structural dynamics*, and how the dynamical fluctuations of a protein governed biological function (Bhabha et al. 2011; Eisenmesser et al. 2002; 2005; Gerek and Ozkan 2011; Jackson et al. 2009; Liberles et al. 2012; Liu et al. 2010; Maguid et al. 2006, 2008; Tang and Altman 2011; Tokuriki and Tawfik 2009; Wang et al. 2004; Zheng et al. 2006, 2007).

9.2 Variation in Sequence Can Modulate Function by Keeping the Three-Dimensional Fold Conserved

The information contained in evolutionary records of proteins has been used extensively in an effort to engineer proteins with a desired function. Variations in amino acid sequences can alter ligand recognition, binding rates, and other biophysical, thermodynamic, and kinetic properties of homologous enzymes while still maintaining similar three-dimensional folds. Differing functions between structural homologs gave rise to a view of protein evolution which proceeds through conformational dynamics and functional promiscuity, where a relationship may exist between active site flexibility and amino acid evolvability (Tokuriki and Tawfik 2009). Comparative research on enzymes sharing close common ancestors with similar structures provided further evidence of evolution proceeding through dynamics. One such study of human and *Escherichia coli* enzyme dihydrofolate reductase showed that while the

two proteins are structurally conserved, both the timescale and motion of enzymatic activity have diverged, indicating two enzymes possess fundamentally different conformational dynamics (Bhabha et al. 2013).

More recently, an emerging concept suggests that sequence evolution may proceed in such a manner that alters the accessibility of conformational states of a given fold through changes in "soft" modes of motion, largely governed by entropic effects (Haliloglu and Bahar 2015). Computational work showed a direct correlation between coevolving residues and mobility or local dynamics of recognition sites, suggesting that this may in fact be a principle of design necessary for substrate interactions (Liu and Bahar 2012). Additionally, this study indicated that it is possible to identify sites whose evolutionary conservation is associated directly with dynamics, including the relationship between hinge sites and conserved residues.

Experimental techniques utilizing directed evolution along with the resurrection of ancestral proteins in vivo have also been instrumental in advancing knowledge of how protein dynamics impact function, and how sequence variations give rise to dynamical changes (Campbell et al. 2018). One such experiment was performed by applying directed evolution to generate nine enzymes which represented different structures along an evolutionary trajectory, as well as three designed variants. Over this trajectory, the protein backbone underwent very little reorganization, while the interaction network of hydrogen bonds, salt bridges, and hydrophobic interactions changed significantly and beta-factors of several regions were altered indicating that, in fact, evolution was proceeding through changes in structural dynamics (Campbell et al. 2016). Particularly, the mutations developed throughout the evolutionary trajectory controlled the accessibility of various conformational substates.

An understanding of sequence variations, evolutionary information, and their relationship to protein dynamics and protein function leads to questions about applications toward rationally designed enzymes with a desired function. Many difficulties arise in such attempts at sequence design, particularly because the function of a protein is a property of many amino acids within a given sequence (Swint-Kruse 2016). One of the biggest challenges to this effort remains in capturing the complex interaction network within a protein in a manner that allows for subtle alterations or fine-tuning in a predictable manner.

9.3 A Position-Specific Method to Compute Each Position's Contribution: Dynamic Flexibility Index (DFI) and Dynamic Coupling Index (DCI)

It is indeed now well accepted that protein dynamics govern function and therefore play an important role in evolution. A major reason for the lack of methods incorporating protein dynamics into evolution and, particularly, the sequence-structure-dynamics-function paradigm, despite the growing realization of its importance, has

been the absence of amino acid site-specific measures that can statistically quantify the contribution and impact of each position on the structural dynamics of the protein.

To that end, we introduce two techniques that are able to capture these dynamics in such a way, the dynamic flexibility index (DFI) (Eq. 3.3) and the dynamic coupling index (DCI) (Eq. 6.1). These are combinations of the elastic network model (ENM) and linear response theory (LRT) where the protein is modeled as a network of nodes (here, individual amino acids) connected with Hookean springs.

To construct DFI, we first model the protein as an elastic network where the interaction between two amino acids close in space due to their three-dimensional structure is represented with a simple harmonic function (i.e., they are connected with a Hookean spring). Then, a random Brownian kick in the form a force perturbation is applied to an individual position which generates a perturbation response propagating through the rest of the structure, causing other positions to respond to this perturbation through a network of interactions. Using LRT, we can calculate the fluctuation response ΔR (Eq. 3.1) of each position and create response vector that measures the magnitude and direction (x, y, and z) of displacement of every residue from its mean. To first order, this mimics the effects of in vivo interactions of a protein as, for example, a ligand-binding event will apply a force to residues in the binding pocket of a receptor protein. In our perturbation residue scanning (PRS) approach, this is averaged over multiple unit force directions to simulate an isotropic perturbation.

$$[\Delta R]_{3N\times 1} = [H]^{-1}_{3N\times 3N}[F]_{3N\times 1} \tag{3.1}$$

H is the Hessian, a $3N \times 3N$ matrix which can be constructed from 3D atomic coordinate information where it is composed of the second derivatives of the harmonic potential with respect to the components of the positions vectors of length N. The Hessian matrix can be extracted directly from molecular dynamics simulations as the inverse of the covariance matrix. This method allows one to implicitly capture specific physiochemical properties and more accurate residue–residue interactions via atomistic force fields and subsequent all-atom simulation data.

Repeating this process, each position in the structure is perturbed sequentially to generate a perturbation response matrix A

$$A_{N\times N} = \begin{bmatrix} \left|\Delta R^1\right|_1 & \cdots & \left|\Delta R^N\right|_1 \\ \vdots & \ddots & \vdots \\ \left|\Delta R^1\right|_N & \cdots & \left|\Delta R^N\right|_N \end{bmatrix} \tag{3.2}$$

where $\left|\Delta R^j\right|_i = \sqrt{(\Delta R)^2}$ is the magnitude of fluctuation response at position i due to the perturbations at position j. The DFI value of position i is then the displacement response of position i relative to the net displacement response of the entire protein, calculated by sequentially perturbing each position in the structure (Eq. 3.3).

$$\mathrm{DFI}_i = \frac{\sum_{j=1}^{N} \left|\Delta R^j\right|_i}{\sum_{i=1}^{N} \sum_{j=1}^{N} \left|\Delta R^j\right|_i} \tag{3.3}$$

DFI quantifies the resilience of a position to perturbations exerted at other parts of the chain. The method consists of applying a random Brownian kick as a mechanical perturbation to a given single residue in the chain and then computing the fluctuation response profile of all other residues to this perturbation. Repeating this random perturbation to each of the positions in the chain one at a time, we are able to compute the normalized response profile (i.e., DFI value) for every residue in the protein. Thus, DFI is a relative value, being higher or lower than the average change that is observed at any position. The residue positions with very low DFI indicate dynamic stability. While these residues do not exhibit large fluctuations upon external force perturbations, they play a key role to transfer these perturbations throughout the chain in a cascade fashion. Thus, they will often be the hinge parts of the protein that control the motion, similar to joints in a skeleton. On the other hand, sites with very high DFI are prone to perturbations to the amino acid chain. They are structurally flexible sites and, therefore, they can play an important role in biochemical function such as being anchoring sites during binding or signaling (Fig. 9.1a).

9.4 Dynamics Flexibility of Individual Positions Plays a Major Role in Protein Function and Evolution

We applied DFI analysis to the human proteome by retrieving experimentally derived monomeric crystal structures of 102 proteins from the Protein Data Bank (PDB) (Nevin Gerek et al. 2013). Particularly, when we focused on the subset of the dataset containing 1912 residues that interact with ligands or small compounds (79 proteins), 104 residues involved in catalytic activities (33 proteins), and 73 residues that are annotated to have both binding and catalytic activities (32 proteins), we observed that the residues with catalytic functions generally show lower dynamic flexibility, with over 50% of the catalytic residues showing %DFI (DFI ranked as a percentile) $\leq$20% (i.e., a hinge point) (Nevin Gerek et al. 2013). Since catalytic sites are often conserved and exhibit low flexibility necessary for enzymatic reaction, they should subsequently also have low DFI values. The binding residues exhibit a relatively higher degree of flexibility than catalytic residues, consistent with the fact that they need to accommodate binding-induced conformational change.

Since these sites of functional importance are generally associated with low DFI values, it follows that there should be some additional association of structural dynamics and therefore DFI with protein evolution. While earlier studies have been conducted showing residues acting as hinges during collective global protein motion are generally more evolutionarily conserved than other positions for specific protein families or a subset of enzymes (Liu and Bahar 2012), in a large-scale, proteome-wide analysis conducted over 39,813 residues from 100 different human proteins, it

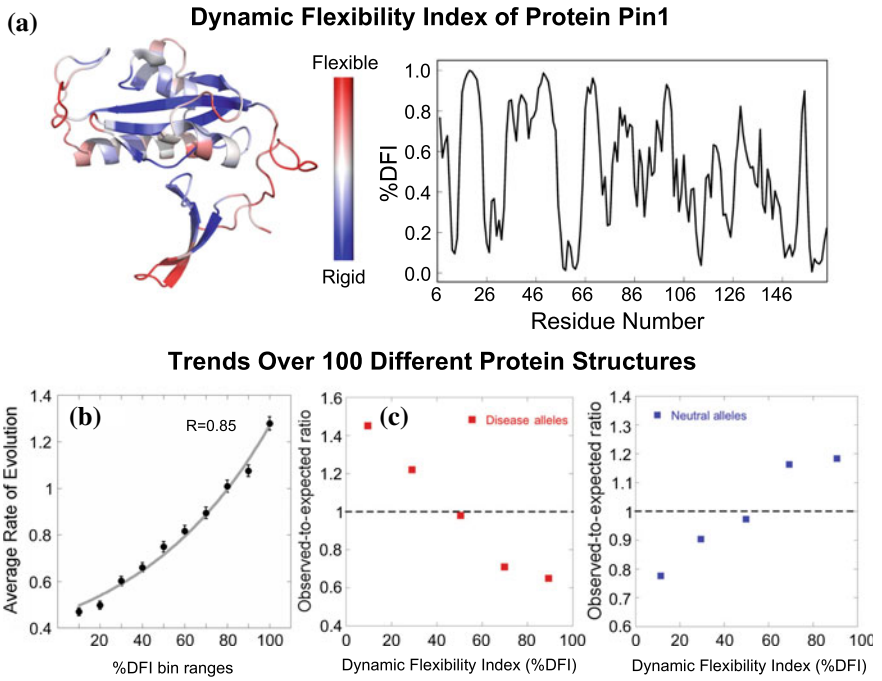


Fig. 9.1 (**a**, left) DFI values mapped onto protein Pin1 (PDB ID 1PIN) where it is colored within a spectrum of red (flexible) to blue (rigid). (**a**, right) DFI values ranked as a percentile, plotted against residue index. Positions with %DFI values under 0.2 are considered hinge regions and are often functionally important sites. **b** Evolutionary rates calculated over 100 sequences from the human proteome versus %DFI values. Amino acid flexibility and evolutionary conservation are highly correlated, with more rigid positions often being more conserved. **c** The observed-to-expected ratios of disease and neutral alleles plotted against their %DFI values calculated over the proteins from (b). Disease alleles are observed much more frequently in hinge sites with low flexibility, while the reverse relationship is true for neutral alleles

was shown that %DFI strongly correlates with position-specific rates of evolutionary change obtained from multispecies sequence analysis ($R = 0.85$, Fig. 9.1b). Here, the position-specific evolutionary rates serve as a proxy for functional importance. Thus, dynamically more important positions will be under stronger natural selection over time, which will permit fewer amino acid substitutions at those positions.

Disease-associated non-synonymous nucleotide variants (nSNVs) in exomes (via their protein products) remain the part of our genome with the best potential for understanding how sequence relates to function through known phenotypic impact and represent our best first chance to bridge genomics and evolution with biophysical aspects of protein structures. By capturing important information related to structural dynamics of protein variants, DFI can be used to help identify positions where mutations can be damaging or disease-causing. In a study conducted using 792 disease-associated variants and 788 neutral-associated variants, it was found that mutations associated with disease generally occurred at positions with low %DFI,

whereas neutral-associated variants exhibited opposite behavior (Nevin Gerek et al. 2013) (Fig. 9.1c). This analysis also suggests that modulation of structural dynamics leads to changes in protein function.

9.5 Nature Fine-Tunes Protein Structural Dynamics to Create a New Function

The relationship between evolution and the dynamics required for protein function is an ongoing topic of study, with recent evidence suggesting that these dynamic features are a result of natural selection. One method of furthering our current understanding of protein dynamics is to take a look backward; advancements in experimental techniques have made possible the resurrection of ancestral proteins and analysis of proteins from billions of years ago has proven crucial in the study of evolution and protein dynamics.

One issue currently at the forefront of human health concerns is that of antibiotic resistance. With the development of the "miracle drug" penicillin and subsequent other forms of antibiotic pharmaceuticals quickly came the emergence of antibiotic-resistant bacteria, followed by an ever-increasing cost of developing new drugs to combat this resistance. Beta-lactamases, enzymes produced by bacteria that provide antibiotic resistances to a wide variety of compounds such as penicillins, carbapenems, and cephalosporins, are at the center of much ongoing research. Mutations are major contributors to the emergence of antibiotic resistance because they can generate new functions (Culyba et al. 2015; Tomaras and Dunman 2015; Wilke et al. 2005). They can, for instance, lead to the overexpression of antibiotic efflux pumps and to the generation of modified enzymes capable of efficiently degrading new antibiotics. Indeed, the emergence of resistance against successive generations of beta-lactam antibiotics is linked to the evolution of mutant beta-lactamases capable of degrading them (Culyba et al. 2015).

In a recent study, several ancestors of the modern extant TEM-1 beta-lactamase were resurrected in a laboratory and although these ancestors had sequence identities ranging from 53% to 79% similarity with modern beta-lactamase, they all shared the similar canonical folded structure (Zou et al. 2015). Moreover, many of these ancestral enzymes were found to be more stable than the current TEM-1 and able to degrade a variety of antibiotics as compared to TEM-1, which is considered a penicillin specialist.

In the absence of structural change, it follows logically that, over several billion years, these proteins evolved in terms of dynamics, and these dynamical changes gave rise to TEM-1 specialization.

An analysis of the structural dynamics of TEM-1 and the three ancestors indicated significant changes in flexibility at sites surrounding the catalytic pocket in the progression from ancestral generalist to modern specialist; while some areas became more flexible, other areas rigidifed. By mapping the DFI values onto the structures,

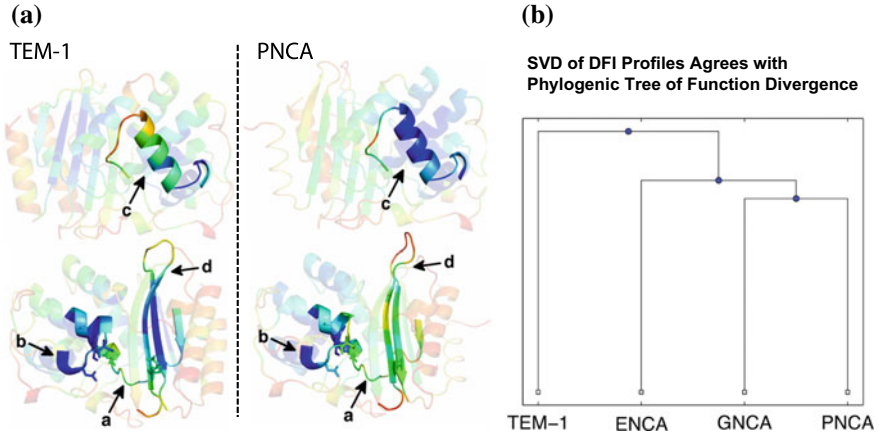


Fig. 9.2 **a** Modern TEM-1 and ancestral PNCA (last common ancestor of Gram-negative and Gram-positive bacteria) structures colored by DFI values where red tones indicate flexible sites while blue tones represent more rigid regions. The flexibility of several regions marked a-d differs notably between the two proteins. **b** Cladogram of singular value decomposition distances of DFI values from TEM-1 extant and ancestors. The structural dynamics of these enzymes clusters in agreement with the timeline of their functional divergences

the changes are readily apparent (Fig. 9.2a). In fact, it was shown that not only did sequence evolution give rise to changes in dynamics which affected enzymatic promiscuity, but a clustering of principle components taken from a singular value decomposition of the DFI matrices for TEM-1 and its ancestors showed that the changes in dynamics actually clustered with their functional divergence over time (Fig. 9.2b).

Associating the evolution of protein dynamics to specific thermodynamic properties is another important study that utilizes ancestral proteins. One fascinating class of proteins are thioredoxins, found present in all known organisms extending from Archaea to mammalian life, and are critical for performing oxidation-reduction reactions. Similar to TEM-1 beta-lactamase, thioredoxins share a high structural similarity with their ancestral forms but exhibit significantly different functional activity and stability. And indeed, their dynamics measured by variance in DFI profile distributions correlates with both thermal stability and evolutionary time (Modi et al. 2018) (Fig. 9.3). Particularly interesting here is that, at a global level, the population of low-flexibility (hinges) and high-flexibility sites both increase over evolutionary time. This position-specific flexibility variation associated with evolution indicates that careful analysis of the evolution of dynamics within a given protein's ancestral line could allow a mechanistic explanation for new phenotypic emergences.

A study conducted on a particular family of green-fluorescent proteins (GFPs) was carried out on a resurrected ancestral line that exhibited a particular characteristic: when exposed to blue or ultraviolet light, these proteins undergo an irreversible transition from green to red fluorescence. These proteins are of high interest in the

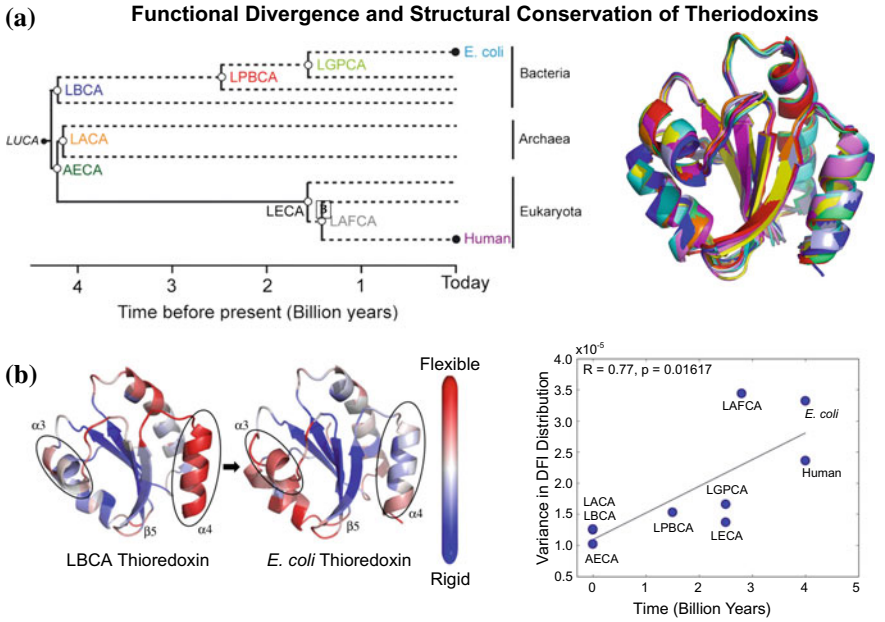


Fig. 9.3 (**a**, left) Phylogenic tree showing human and *E. coli* thioredoxin ancestry back to their shared ancestor. (**a**, right) Structural overlay of thioredoxin and ancestors. While exhibiting significant sequence variation, the three-dimensional folds are conserved. **b** Flexibility measured as %DFI comparing modern *E.coli* thioredoxin to its distant last common bacterial ancestor, LBCA. There are notable flexibility changes in major secondary structures including two helices. (**c**) Variance in DFI distributions of ancestral and modern-day THRX proteins correlates strongly ($R = 0.77$) with evolutionary time

scientific community in large part due to their usefulness in super-resolution fluorescent microscopy. A series of ancestral proteins was created in a laboratory, ranging from a common green ancestor (ALL) to a least evolved ancestor (LEA) which contains the minimum number of mutations needed to produce red fluorescence (Kim et al. 2015). Comparing the dynamics of a modern mutant (ALL-Q62H) to LEA by mapping DFI values onto the structure, one can see striking differences in flexibility about the external region of the protein's beta barrels (Fig. 9.4a) which, again, occur in the absence of any significant backbone rearrangement between the two proteins.

Comparing the changes in DFI values from LEA to the modern extant ALL-Q62H makes the mechanistic difference readily apparent, observing a clear separation of regions of increased and decreased flexibility (Fig. 9.4b). This indicates that a hinge-shift mechanism has taken place, where a rigid knob-like component has shifted diagonally across the beta-barrel in response to the historical mutations correlating with photochromism and provides a direct connection between natural protein evolution and mechanistic changes emergent from dynamical properties.

In addition to the differences in dynamics, what is perhaps even more striking is that the sites of mutations themselves exhibit fairly small changes in DFI, including

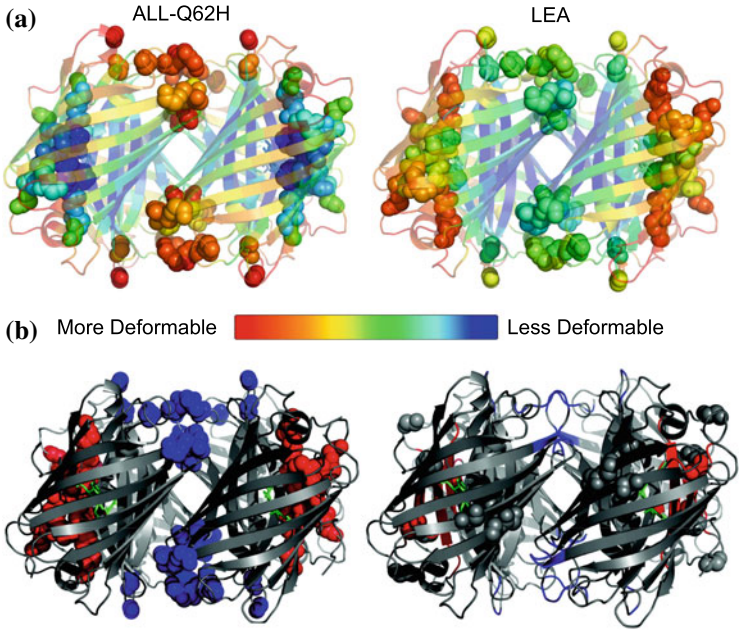


Fig. 9.4 **a** Structures of GDP common ancestor variant ALL-Q62H and least evolved ancestor (LEA) colored by %DFI values. Residues with a significantly different flexibility are represented as space-filling spheres using main and side-chain atoms. **b** The top 15% of ΔDFI (difference of DFI values between LEA and ALL-Q62H) values mapped onto the ALL-Q62H structure (right), where red residues are more flexible and blue residues are more rigid in LEA as compared to the modern extant. The formation of new hinge sites regulates changes in function between the two enzymes. The sites of sequence variation are marked as gray spheres (left). Note that the highest changes in flexibility occur at regions that were not directly mutated

the chromophore itself, in which the Q62H mutation is included. This indicates that the changes in amino acid sequence over the course of evolution were usually not occurring at sites directly involved with photochromatic activity (i.e., at functionally critical catalytic sites). In fact, as shown previously, many of the sites in a protein critically important for function are generally also sites of evolutionary conservation.

9.6 Nature Uses Allosteric Mutations to Modulate Function

How, then, does evolution proceed in a way that gives rise to new emergent properties but also in a manner that does not completely disrupt the proper function of a protein? The answer lies in allostery, the regulation of one part of the protein by regions distal to it, located elsewhere in the structure. Famously coined “the second secret of life” (the first, of course, being DNA) by Jacques Monod, allosteric regulation was first and foremost understood as interaction between two binding pockets of a given protein. In

the traditional view of allostery, a ligand would bind at a site other than the protein's active site, which would then induce some change in the catalytic rate or overall enzymatic activity or binding affinity at the active site, often through conformational changes either at the active site or at other regions of the protein that impacted the active site.

However, it became readily apparent that not all allosteric regulation was so straightforward. Discoveries were made that indicated allosteric regulation upon ligand binding occurred in the absence of any significant structural rearrangement, suggesting that the effect may proceed through changing conformational dynamics. Furthermore, this picture of allosteric regulation was unable to explain effects associated with point mutation studies which showed that, occasionally, single mutations at positions in the protein not directly involved with binding or part of an active site dramatically altered the protein's ability to function.

At its core, allostery is the governing of protein dynamics by regulation of interactions between distal positions within a given protein and disruption of these interactions can have a profound impact on protein behavior. In an effort to fully capture these effects, we here introduce the dynamic coupling index (DCI).

In a manner similar to DFI, DCI captures the strength of displacement response of a given position i upon perturbation to a single functionally important position (or subset of positions) j, relative to the average fluctuation response of position i when all of the positions within a structure are perturbed.

$$\mathrm{DCI}_i = \frac{\sum_j^{N_{\text{functional}}} \left|\Delta R^j\right|_i / N_{\text{functional}}}{\sum_{j=1}^{N} \left|\Delta R^j\right|_i / N} \tag{6.1}$$

As such, DCI can be considered a measure of the dynamic coupling between residue i and j upon perturbation to residue j.

Single-point mutations having a dramatic effect on the function of a protein in the absence of structural changes or even dynamical changes at the point of mutation or the directly surrounding region is a strong indicator of the presence of allostery. In fact, outside of the traditional view of allostery, it seems that this type of regulation through alteration of dynamics between distally located regions of a protein is prevalent in perhaps *all* proteins. This requires an understanding of a protein as a molecular machine at its most fundamental level, that of a signal processor. Allostery, then, is simply the regulation of the networking throughout the entire structure, connecting every region of the protein to other regions. The propagation of signals from one region to another is integral to protein function, enzymatic activity, and, subsequently, allosteric interaction.

Indeed, the comparison of the amino acid substitutions found in all known clinical TEM-1 isolates with those found in vitro or in vivo experimental evolution of TEM-1 has revealed that the majority of mutations which contribute to resistance are distally located (Kumar et al. 2015; Larrimore et al. 2017; Modi et al. 2018). However, irrespective of the separation distance between the mutational sites and the active

sites, most of the resistance driving mutations studied were dynamically coupled with the active sites as shown by overall higher DCI scores (Fig. 9.5).

Sites which exhibit strong dynamic allosteric residue coupling (DARC spots) can contribute to protein function, regardless of the separation distance. It appears that nature can take advantage of allosteric regulation by altering protein dynamics which then give rise to changes in dynamic coupling and thus allosteric networking through DARC spot mutations. Applied to the previous study on ancestral thioredoxins, this is shown explicitly when comparing the coupling to catalytic site residues of an ancestor with the extant enzyme (Fig. 9.6). Specifically, one of the alpha helices ($\alpha 3$) exhibits much weaker coupling to these catalytic sites. This helix had also previously been shown to impact the overall stability of the structure when its formation was disrupted.

As discussed previously, changes in conformational dynamics of proteins can often lead to diseases, even in cases of only a single point mutation due to nSNVs (i.e., missense variations). An interesting test case is that of human ferritin protein, which stores and releases iron in a controlled manner. Diseased variants of ferritin have been linked with a wide range of conditions including anemia, cataract syndrome, Parkinson's disease, Huntington's disease, and many other neurodegenerative diseases. Composed of 24 subunits, single point mutations in the light chain subunit (FTL) can be either neutral or disease-causing, even when these mutations are very close in sequence space. A study conducted on variants associated with cataract syndrome showed that the DFI profiles of variants associated with disease lead to an increase in flexibility of two particular regions of the protein, the C-terminal end and a regulatory loop denoted L1 (Kumar et al. 2015). Interestingly, many of these mutations were not located in either of those regions. Two exceptions were the D42Y and R40G mutations, both very close to loop L1; however, R40G is a neutral mutation while D42Y is disease-associated.

To fully understand how these different mutations were affecting the functional behavior of the protein, the dynamic coupling was measured between the functionally

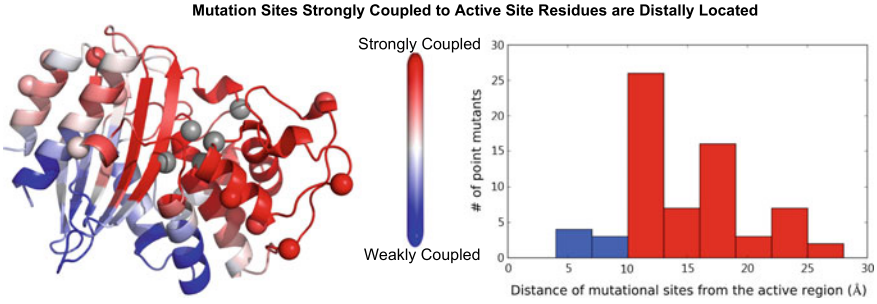


Fig. 9.5 Coupling to active site residues measured by DCI, where red sites are strongly coupled and blue sites are weakly coupled. All of the mutation sites strongly coupled to the active site residues are located distally, show in the histogram distribution of distances from the active site

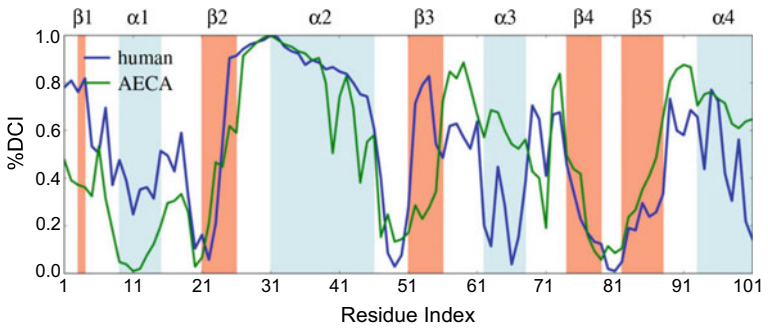


Fig. 9.6 Comparison of dynamic coupling of catalytic site residues of human ancestral thioredoxin AECA with the extant enzyme, measured as %DCI. The coupling diverges significantly, particularly to alpha-helix α3, which had been previously shown to impact the overall stability of the structure when its formation was disrupted

important sites (here, L1 and the C-terminus) and the rest of the structure. The difference in the coupling profiles between the wild-type, average neutral variants, and average disease variants was striking and suggests that these mutations are disrupting important allosteric regulation involving these two functionally critical regions (Fig. 9.7).

An excellent example highlighting this description is Pin1, a protein with a WW domain and a peptidyl prolyl isomerase (PPIase) domain whose up-regulation is associated with various cancers and down-regulation is associated with Alzheimer's disease. The WW domain can bind to the same ligands as the PPIase domain, but only the PPIase domain is involved in enzymatic activity. However, when a ligand is bound to the WW domain, the rate of catalysis at the PPIase domain increases dramatically. Additionally, a ligand binding to the WW domain does not cause any large-scale global structural rearrangement or changes to the PPIase binding pocket

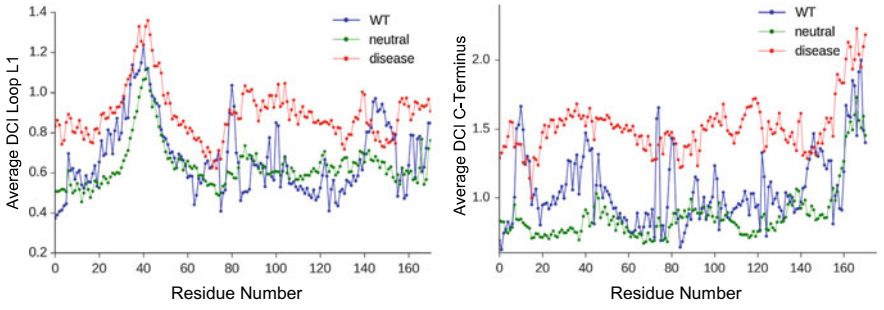


Fig. 9.7 Dynamic coupling profiles to functionally important regions loop L1 (left) and C-terminus (right) of the wild-type, average neutral variants and average disease variants. Disease-associated variants exhibit overall stronger coupling to these regions across the entire protein structure, which disrupts important allosteric regulation

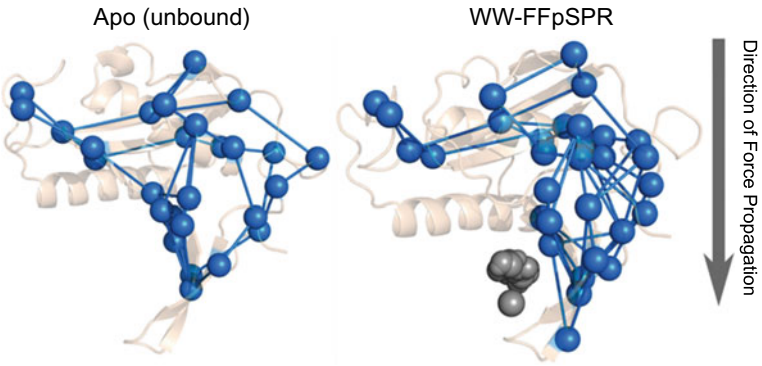


Fig. 9.8 Comparison of shortest distance force propagation pathways between the apo (unbound) form of Pin1 and the WW-FFpSPR bound structures, traveling from PPIase domain binding sites to the WW domain. While the structures do not have significant deviations in backbone orientation, they exhibit largely different allosteric pathways, where the bound form has a more robust communication network between the two domains

itself. It was proposed that allosteric regulation through a WW domain binding event proceeded by enhancing the communication pathways between the two domains. This was investigated through the construction of pathways of force propagation originating at the PPIase domain and traveling to the WW domain.

Here, a residue is perturbed isotropically, calculating the displacement response vector of the resultant fluctuations of nearby residues with a sequence separation of three or greater ($j \geq i + 3$). Residues are considered linked to the perturbed residue if they respond with a directionality of 0.98 or greater than that of the perturbed residue, as evaluated by the cosine of the angle between response vectors. Creating shortest distance pathways between the two domains shows directly the changes in internal networking when a ligand is bound at the WW domain (Campitelli et al. 2018). In short, the allosteric regulation is a result of additional pathways of communication as well as the modulation of which residues participate in the interaction network between the two domains (Fig. 9.8).

9.7 Conclusion

In summary, protein evolution proceeds through modulation of structural dynamics, where mutations at distal regions alter the dynamics of functionally critical conserved sites. We are in the era of rapid development of next-generation methods for whole-genome, whole-exome, and targeted sequencing that has generated an unprecedented amount of data. In order to decipher the information contained in sequence variations and its relationship with function, we need to unite the genetic code, which constitutes "the first secret of life," and allostery, "the second secret of life". This could unravel the physical principles behind the evolutionary process and potentially

reveal a generalized disease mechanism and allow for discovery of novel drugs, as well as the blueprints for innovative personalized treatment methods.

References

Bhabha G, Lee J, Ekiert DC, Gam J, Wilson IA, Dyson HJ, Benkovic SJ, Wright PE (2011) A dynamic knockout reveals that conformational fluctuations influence the chemical step of enzyme catalysis. Science 332:234–238. https://doi.org/10.1126/science.1198542

Bhabha G, Ekiert DC, Jennewein M, Zmasek CM, Tuttle LM, Kroon G, Dyson HJ, Godzik A, Wilson IA, Wright PE (2013) Divergent evolution of protein conformational dynamics in dihydrofolate reductase. Nat Struct Mol Biol 20:1243–1249. https://doi.org/10.1038/nsmb.2676

Campbell E, Kaltenbach M, Correy GJ, Carr PD, Porebski BT, Livingstone EK, Afriat-Jurnou L, Buckle AM, Weik M, Hollfelder F, Tokuriki N, Jackson CJ (2016) The role of protein dynamics in the evolution of new enzyme function. Nat Chem Biol 12:944 EP. https://doi.org/10.1038/nchembio.2175

Campbell EC, Correy GJ, Mabbitt PD, Buckle AM, Tokuriki N, Jackson CJ (2018) Laboratory evolution of protein conformational dynamics. Curr Opin Struct Biol 50:49–57. https://doi.org/10.1016/j.sbi.2017.09.005

Campitelli P, Guo J, Zhou H-X, Ozkan SB (2018) Hinge-shift mechanism modulates allosteric regulations in human pin1. J Phys Chem B 122:5623–5629. https://doi.org/10.1021/acs.jpcb.7b11971

Culyba MJ, Mo CY, Kohli RM (2015) Targets for combating the evolution of acquired antibiotic resistance. Biochemistry 54:3573–3582. https://doi.org/10.1021/acs.biochem.5b00109

Eisenmesser EZ, Millet O, Labeikovsky W, Korzhnev DM, Wolf-Watz M, Bosco DA, Skalicky JJ, Kay LE, Kern D (2005) Intrinsic dynamics of an enzyme underlies catalysis. Nature 438:117–121. https://doi.org/10.1038/nature04105

Eisenmesser EZ, Bosco DA, Akke M, Kern D (2002) Enzyme dynamics during catalysis. Science 295:1520–1523. https://doi.org/10.1126/science.1066176

Gerek ZN, Ozkan SB (2011) Change in allosteric network affects binding affinities of PDZ domains: analysis through perturbation response scanning. PLoS Comput Biol, vol 7. https://doi.org/10.1371/journal.pcbi.1002154

Haliloglu T, Bahar I (2015) Adaptability of protein structures to enable functional interactions and evolutionary implications. Curr Opin Struct Biol 35:17–23. https://doi.org/10.1016/j.sbi.2015.07.007

Jackson CJ, Foo J-L, Tokuriki N, Afriat L, Carr PD, Kim H-K, Schenk G, Tawfik DS, Ollis DL (2009) Conformational sampling, catalysis, and evolution of the bacterial phosphotriesterase. Proc Natl Acad Sci USA 106:21631–21636. https://doi.org/10.1073/pnas.0907548106

Kim H, Zou T, Modi C, Dörner K, Grunkemeyer TJ, Chen L, Fromme R, Matz MV, Ozkan SB, Wachter RM (2015) A hinge migration mechanism unlocks the evolution of green-to-red photoconversion in GFP-like proteins. Structure 23:34–43. https://doi.org/10.1016/j.str.2014.11.011

Knudsen M, Wiuf C (2010) The CATH database. Hum Genomics 4:207–212. https://doi.org/10.1186/1479-7364-4-3-207

Kumar A, Glembo TJ, Ozkan SB (2015) The role of conformational dynamics and allostery in the disease development of human ferritin. Biophys J 109:1273–1281. https://doi.org/10.1016/j.bpj.2015.06.060

Larrimore KE, Kazan IC, Kannan L, Kendle RP, Jamal T, Barcus M, Bolia A, Brimijoin S, Zhan C-G, Ozkan SB, Mor TS (2017) Plant-expressed cocaine hydrolase variants of butyrylcholinesterase exhibit altered allosteric effects of cholinesterase activity and increased inhibitor sensitivity. Scientific Reports 7:10419. https://doi.org/10.1038/s41598-017-10571-z

Liberles DA, Teichmann SA, Bahar I, Bastolla U, Bloom J, Bornberg-Bauer E, Colwell LJ, de Koning APJ, Dokholyan NV, Echave J, Elofsson A, Gerloff DL, Goldstein RA, Grahnen JA, Holder MT, Lakner C, Lartillot N, Lovell SC, Naylor G, Perica T, Pollock DD, Pupko T, Regan L, Roger A, Rubinstein N, Shakhnovich E, Sjölander K, Sunyaev S, Teufel AI, Thorne JL, Thornton JW, Weinreich DM, Whelan S (2012) The interface of protein structure, protein biophysics, and molecular evolution. Protein Sci 21:769–785. https://doi.org/10.1002/pro.2071

Liu Y, Bahar I (2012) Sequence evolution correlates with structural dynamics. Mol Biol Evol 29:2253–2263. https://doi.org/10.1093/molbev/mss097

Liu Y, Gierasch LM, Bahar I (2010) Role of Hsp70 ATPase domain intrinsic dynamics and sequence evolution in enabling its functional interactions with NEFs. PLoS Comput Biol, vol 6. https://doi.org/10.1371/journal.pcbi.1000931

Maguid S, Fernández-Alberti S, Parisi G, Echave J (2006) Evolutionary conservation of protein backbone flexibility. J Mol Evol 63:448–457. https://doi.org/10.1007/s00239-005-0209-x

Maguid S, Fernandez-Alberti S, Echave J (2008) Evolutionary conservation of protein vibrational dynamics. Gene 422:7–13. https://doi.org/10.1016/j.gene.2008.06.002

Modi T, Huihui J, Ghosh K, Ozkan SB (2018) Ancient thioredoxins evolved to modern-day stability-function requirement by altering native state ensemble. Philos Trans R Soc Lond, B, Biol Sci 373. https://doi.org/10.1098/rstb.2017.0184

Nevin Gerek Z, Kumar S, Banu Ozkan S (2013) Structural dynamics flexibility informs function and evolution at a proteome scale. Evol Appl 6:423–433. https://doi.org/10.1111/eva.12052

Swint-Kruse L (2016) Using evolution to guide protein engineering: the devil is in the details. Biophys J 111:10–18. https://doi.org/10.1016/j.bpj.2016.05.030

Tang GW, Altman RB (2011) Remote thioredoxin recognition using evolutionary conservation and structural dynamics. Structure 19:461–470. https://doi.org/10.1016/j.str.2011.02.007

Tokuriki N, Tawfik DS (2009) Protein dynamism and evolvability. Science 324:203–207. https://doi.org/10.1126/science.1169375

Tomaras AP, Dunman PM (2015) In the midst of the antimicrobial discovery conundrum: an overview. Curr Opin Microbiol 27:103–107. https://doi.org/10.1016/j.mib.2015.08.005

Wang Y, Rader AJ, Bahar I, Jernigan RL (2004) Global ribosome motions revealed with elastic network model. J Struct Biol 147:302–314. https://doi.org/10.1016/j.jsb.2004.01.005

Wilke MS, Lovering AL, Strynadka NCJ (2005) Beta-lactam antibiotic resistance: a current structural perspective. Curr Opin Microbiol 8:525–533. https://doi.org/10.1016/j.mib.2005.08.016

Zheng W, Brooks BR, d. Thirumalai (2006) Low-frequency normal modes that describe allosteric transitions in biological nanomachines are robust to sequence variations. Proc Natl Acad Sci USA 103:7664–7669. https://doi.org/10.1073/pnas.0510426103

Zheng W, Brooks BR, d. Thirumalai (2007) Allosteric transitions in the chaperonin GroEL are captured by a dominant normal mode that is most robust to sequence variations. Biophys J 93:2289–2299. https://doi.org/10.1529/biophysj.107.105270

Zou T, Risso VA, Gavira JA, Sanchez-Ruiz JM, Ozkan SB (2015) Evolution of conformational dynamics determines the conversion of a promiscuous generalist into a specialist enzyme. Mol Biol Evol 32:132–143. https://doi.org/10.1093/molbev/msu281

Chapter 10
Coevolutionary Approaches to the Science of Language

Nicholas Evans

> If this is right, then all the achievements of human culture—language, art, religion, ethics, science itself—are themselves artifacts (of artifacts of artifacts …) of the same fundamental process that developed the bacteria, the mammals, and Homo Sapiens Dennett (1995:144).

Abstract Since the famous exchange of letters between Darwin and Schleicher, the parallels between evolutionary processes in the biological and linguistic spheres have been evident. In this paper, I present a coevolutionary approach to language evolution, both in the early phase during which hominins evolved language and in subsequent phases during which humans evolved many thousands of languages whose vastly differing structures serve as a basic resource for understanding the operation of evolutionary processes on languages and cultures. The key elements in this coevolutionary approach are (a) the adoption of a gradualist approach to initial language evolution and (b) the recognition of a large number of selectors (systemic, modality, demographic, usage patterns, biogenetic, epidemiological, sociocultural) which are unevenly distributed across speaker populations and which may nudge emerging languages structures into quite different parts of the design space. Not only does the coevolutionary approach presented here bring the methods of studying linguistic evolution closer to those used in biology, it places the phenomenon of diversity and variability—diversity at the level of differences between languages, and variability between how individuals use them—into the same central role that these occupy in evolutionary biology.

N. Evans (✉)
ARC Centre of Excellence for the Dynamics of Language (CoEDL), School of Culture, History and Language, College of Asia, and the Pacific, Australian National University, Canberra, Australia
e-mail: nicholas.evans@anu.edu.au

P. Pontarotti (ed.), *Evolution, Origin of Life, Concepts and Methods*,
https://doi.org/10.1007/978-3-030-30363-1_10

10.1 Evolution in Biology and Linguistics: An Enduring Relationship

Biology and linguistics are old allies in developing evolutionary approaches (Atkinson and Gray 2005). This famous quote from Charles Darwin occurs in *The Descent of Man*

> The formation of different languages and of distinct species, and the proofs that both have been developed through a gradual process, are curiously the same. (Charles Darwin 1871:59)

The interest in dialogue between evolutionary biology and the methods of historical–comparative linguists was mirrored by the linguist Schleicher (1869), who published an open letter after reading the German translation of the *Origin of Species*, including this passage:

> The kinship of the different languages may consequently serve, so to speak, as a paradigmatic illustration of the origin of species, for those fields of inquiry which lack, for the present at least, any similar opportunities of observation. (Schleicher 1869, pp. 44–45)

Darwin, alerted to Schleicher's publication when he read a review by the Cambridge philologist Max Müller of an English translation of it, then used language as a source of metaphors for how evolution proceeds in the biological realm:

> We see variability in every tongue, and new words are continually cropping up; but as there is a limit to the powers of the memory, single words, like whole languages, gradually become extinct. As Max Müller has well remarked: "A struggle for life is constantly going on amongst the words and grammatical forms in each language. The better, the shorter, the easier forms are constantly gaining the upper hand". (Darwin 1871, 60; referring to Müller's article in *Nature*, Jan. 6, 1870, 257)

This is just a small snippet from the very productive dialogue that occurred between Darwin and historical linguists during the formative years of the theory of evolution, but it indicates how useful it is to view evolutionary theory not just as something that applies to the biological domain, but also as a model that can profitably be applied to systems representing human cultural institutions.

However, this natural interfertility between biology and linguistics has not proceeded uninterrupted. In particular during the half-century from the late 1950s, marked by the dominance of Chomskyan linguistics, the reigning conceptions of language and how it may have evolved were not congenial to evolutionary approaches.

For Chomsky and his colleagues, as evidenced in numerous publications, the essence of language is defined by a single characteristic—recursion, the ability to embed linguistic expressions inside one another, as in the recursively embedded 'that-complements' in the sentence: [Chomsky argued [that philosophers believed [that biologists proceed by induction]]]. Recursion, by allowing language to telescope out expressions of indefinite complexity, provides the means by which language is 'generative', allowing the infinite use of finite means, taken as a defining feature of language for many linguists since the German linguist and philosopher Wilhelm von Humboldt in the nineteenth century.

While important for the development of generative syntax, this move distanced the explanation of language genesis from the sorts of evolutionary accounts given for crucial biological developments in evolutionary theory, by making its precursors non-adaptive and by proposing that it arose in a single saltationist leap. The evolution of language, on this view, was not driven by any selective factors or any communicative role for the development of recursion:

> It surely cannot be assumed that every trait is specifically selected. In the case of such systems as language or wings it is not even easy to imagine a course of selection that might have given rise to them. A rudimentary wing, for example, is not "useful" for motion but is more of an impediment. (Chomsky 1988)

More recently, the field has seen a rapprochement between evolutionary narratives and the study of how language may have evolved (Croft 2008; Dediu et al. 2013). A number of factors have contributed to this renewed synergy, among them a much broader conception of how many elements need to be brought together for language to evolve, a retreat from essentialist views of a single 'universal grammar' to a conception that moves linguistic diversity from noise to signal (Evans and Levinson 2009), and the development of coevolutionary approaches to language evolution. This article reviews the benefits that these approaches are bringing to our understanding of language, and points forward to some of the major tasks that face evolutionary accounts both of the deep origins of language and of the striking diversification that is found in modern tongues.

10.2 Coevolution and Language

There is a central difference between the genetic transmission of biological traits and the transmission of language, regarding the selective processes that apply between the form it takes in each generation: whereas the code of life is transmitted directly from one generation to the next through DNA, the transmission of language must be seen as a *twin-track* process, reflecting its double life as

(a) *a social institution*—an agreed body of conventions about what words are, how they can be combined, what is 'grammatical', what is 'slang' and what is simply outlandish or never heard. This resides in what linguists call 'the speech community', which we can think of intuitively as the set of people sharing (more or less) the same conventions for the use of language. For example, taking the word spelled *sensible*, members of the Anglophone speech community agree that it means, roughly, 'practical, reflecting good sense or judgement' while members of the Francophone speech community agree that it means what in English would be rendered as 'perceptible' but also 'sensitive' and 'appreciable, detectable'. The two speech communities also differ significantly in how they would pronounce the word: respectively [sensəbl] and [sãsibl]).
(b) *an individual phenomenon*—a complex network of mental representations coupled with practised perceptual pathways and complex articulatory movements,

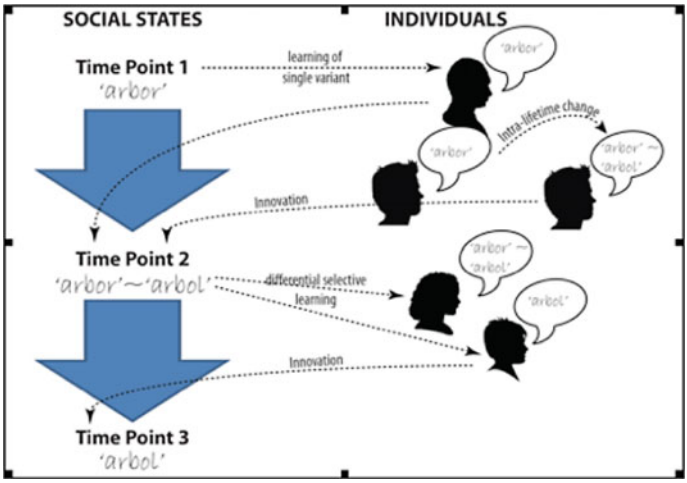


Fig. 10.1 Twin-track evolution and the evolution of Spanish *árbol* from Latin *arbor*

which each individual builds over their lifetime through interactions with those speaking the language around them.

Changes in transmission can take place at either level. At the social level, for example, decisions[1] about what should or should not be transmitted, or what is emblematic of a particular group, act as institutional filters or biases that accept or reject particular variants. At the individual level, learning of forms may be affected by sampling bias (e.g. those they interact with are outliers in terms of the overall social group, or different individuals learn the meaning of a word, say *peruse,*[2] in different contexts) or by their own streamlinings or changes (Fig. 10.1).

For example, saying the same sound twice in rapid succession is more difficult than saying two different sounds, which often leads to 'dissimilation', e.g. avoiding saying the *r* sound twice in the word *arbor* 'tree' and replacing the second instance with *l*, giving *arbol*. This is precisely the change that turned Latin *arbor* into Spanish *árbol*. To understand how it occurred, we need to understand first how it arises in individuals producing particular utterances (through processes like dissimilation),

[1] Here I abstract away from a host of fascinating questions regarding the relative role of conscious and unconscious factors, overt and covert prestige (the former being the 'prescribed norm' associated with education, the latter associated with, e.g., working-class toughness or authenticity), and the positioning and changing prestige of different groups with respect to one another (e.g. aristocrats and the merchant class at the time of Shakespeare, or of different immigrant groups with respect to non-immigrant groups in large cities). For more on these factors, consult such sociolinguistics textbooks as Meyerhoff (2015).

[2] For some speakers, this word means 'read slowly and carefully, over a long period' while for others it means 'read casually, when there's a bit of time'. This is likely to reflect different inductions from a context like 'peruse this at your leisure', where the formulation can be taken to mean either 'this is such a substantial task that you'll need a lot of reading time to do it' or 'do this as part of your leisure activities, so not particularly seriously'.

and second how, at some point at the social level, the speech community adopts this variant as correct (and possibly a badge of distinction from other language) rather than incorrect. Generally, this change in social convention sums over what is happening with a substantial number of individuals, except in rare cases such as when a single high-prestige individual serves as a model.

Any language is the product of many thousands of changes of this type—each sound, each word, each grammatical form or syntactic rule is the product of such a process (even if some changes are linked together). And it often takes centuries, sometimes millennia, for such a change to take place. For example, the future tense suffix in a French word like *chanterai* 'I will sing' derives from the *-eo* at the end of the Vulgar Latin phrase *cantare habeo*, in a development whose first inkling can already be glimpsed in the second century A.D. in the writings of Tertullian (Pinkster 1985).

Two important differences between language evolution and biological evolution follow. Firstly, the impact of changes at individual level, which can subsequently be incorporated into social norms, adds a Lamarckian element to linguistic evolution: changes through the course of an individual's lifetime can indeed be transmitted. Secondly, the existence of two loci at which the phenomenon of language exists—social and individual—sets the scene for us to adopt *coevolutionary approaches* to language change that pay attention to the mutually interacting effects of change on each of these tracks.

The notion of coevolution, the interaction of two sets of entities each of which exerts selective pressure on the other, already appears (without the term) in Darwin's (1859) discussions of the evolutionary interactions between flowering plants and insects in the *Origin of Species*. A beautiful example, for which the full evidence took 130 years to unfold, is the coevolution of the long nectaries of Darwin's orchid (*Angraecum sequipedale*), on the basis of which Darwin hypothesised the existence of a moth with an outsized proboscis. Alfred Russel Wallace (1867), in his paper Creation by Law, gives a coevolutionary account of how this symbiosis would have evolved through selective pressures on both sides, and the article even led the artist Thomas Wood to an illustration of an imaginary hawkmoth probing the foot-long nectary with its outsized proboscis. The relevant moth subspecies (*Xanthopan morganii praedicta*) was not discovered until 1907, and it was not until 1992 that an individual was actually observed feeding and transferring pollen. The actual term coevolution was first coined by Ehrlich and Raven (1964).

More recently, several other fields have adopted coevolutionary accounts. In physical anthropology, for example, biological anthropologists have begun to explore interactions between the specifics of culture and population genetics (e.g. pastoralism and associated drinking of cow's milk with evolved lactose tolerance) or between the geographic setting where a population resides and their physical characteristics (e.g. oxygen-transport in the blood of Tibetans dwelling at high altitudes). See Durham (1991) for an overview.

Such coevolutionary approaches have obvious relevance to many questions regarding the evolution of language (Evans 2016), and we briefly frame their relevance here.

The centrality of language to human life, culture and thought, means that several types of coevolutionary relations can be identified. First, as already mentioned, there are the relations between the social and individual manifestations of language. The 'social' itself conceals a multitude of sub-effects, from demography of the speech community (e.g. it has been argued that certain types of linguistic complexity are most likely to develop in small, inward-looking speech communities (Trudgill 2011)) to social structure (e.g. stratified vs egalitarian, or divided in moieties) to the impact of habitual ways of talking, or cultural context, upon the evolution of particular grammatical or semantic structures (Evans 2003).

Then, because languages are used by embodied human beings, a whole range of biological factors enter the picture: articulators (e.g. the exact shape of the mouth, the lowering of the larynx), breath control, perceptual apparatus and brain configuration, to name just a few. Environmental factors such as altitude may influence the ease with which certain sounds such as ejectives are articulated (Everett 2013), while tree cover and precipitation may affect the amount of 'consonant-heaviness' by effecting the fidelity of high-frequency acoustic transmission (Maddieson and Coupé 2015). The dietary transition to refined food, reducing the amount of heavy chewing required, may reset the overbite through the lifetime, then influencing the likelihood that a language will have labiodental fricatives (like f in English, where the lower lip touches the teeth) as opposed to 'bilabial fricatives' in which the two lips are brought close together (Blasi et al. 2019).

A key ingredient of these coevolutionary processes is the amplifying effect of iterated learning across hundreds of generations, as language is passed back and forth from the individual to the speech community, which in turn supplies the exemplars for the next generation of learners. Computational simulations demonstrate how multiple iteration (Kirby et al. 2007; Dediu 2011) can turn small selector biases into quite different evolutionary outcomes, and we now have clear models of how processing and learning bottlenecks can nudge the evolution of linguistic systems in particular directions (Christiansen and Chater 2016).

The society and culture in which a particular language evolves, including the biological makeup of its speaker population, is thus likely to exert some influence on what languages emerge from the ongoing evolutionary process. We will return to this issue in Sect. 10.4, but before doing so it is useful to return to the earliest stages in language evolution, so as to set out the broad elements that needed to develop before something like modern languages could appear.

10.3 The Initial Evolution of Language

If we are to reject the saltationist views advanced by Chomsky (1988, 2007) and his colleagues (e.g. Chomsky and Berwick 2016), we need to invoke a number of principles for early language evolution: gradualism (Sect. 10.3.1), the cumulative effects of iteration (Sect. 10.3.2), the embedding of language evolution in human interaction (Sect. 10.3.3), the importance of horizontal transmission in pooling innovations

across populations (Sect. 10.3.4) and the likelihood that different partial precursors to language emerged in different places ('polysemigenesis') rather than a complete modern language emerging in one fell swoop Sect. 10.3.5).

10.3.1 Gradualism

Languages, and the cognitive abilities underlying their use, are packages of a large number of elements at various levels that would have evolved in many gradual steps—perhaps going back to the common ancestor we share with Neanderthals (Dediu and Levinson 2018).

This applies equally to

(a) MAJOR ARCHITECTURAL PRINCIPLES, such as compositionality, dual patterning, recursion (see above) and arbitrariness.

Compositionality is the mechanism by which the meaning of complex signs like *fearlessness* can be inferred from that of their constituents (*fear* + *less* + *ness*), plus rules of composition. *Dual patterning* allows meaningful words (like *ban* or *bat*) to composed from meaningless elements (b, a, n, t) in much the same way as amino acids can be composed from nucleotides. *Arbitrariness* refers to the fact that meanings of words do not bear any direct 'iconic' relationship to the sounds or hand-signs that denote them—there is no inherent reason why the meaning 'fish' should be expressed by *fish*, *ryba* (Russian), *yakuri* (Kayardild), *wagib* (Nen) or any of the thousands of words used to express this meaning in the world's languages, and the same holds for the vast majority of words outside a few onomatopoeics like *woof woof* (and even there not all languages agree—in Albanian dogs go *ham! ham!*).

Note that each of these principles is logically independent. One can, for example, apply compositionality to form complex signs whose components are non-arbitrary signs: if you were searching two-species cages in a zoo, and I tell you one is the *meow-cockadoodledoo* cage and the other the *meow-quack* cage, you would know which cage was which. Languages, like Pirahã, claimed to lack recursion (Everett 2005), have the other three properties, and so on for other combinations of these major principles. This shows that language could evolve gradually even with respect to these fundamental properties, evolving first one then the other.

(b) DISTRIBUTION OVER CHANNELS (eye-hand vs ear-mouth) AND DIFFERENTIAL AFFORDANCES. Not all languages are spoken—some are signed (and there are even tactile languages for the blind-deaf)—demonstrating that the same meanings and functions can be realised by different channels (e.g. saying 'no' vs shaking one's head). In spoken languages, there is often a close-knit distribution between more than one channel, for example integrating a pointing gesture with the use of a demonstrative like *this*. In the earliest evolution of language, it is likely that some elements evolved in the eye-hand medium (e.g. pointing for attentional coordination) and others in the mouth-ear medium (e.g. sounds

imitating bird calls, but also intonation indicating speaker attitude). Since we know there are modern sign languages with full expressivity, it would have been logically possible for a full language to evolve in the eye-hand medium first, then selecting for brains able to handle its structural complexities, while the articulatory system for speech 'caught up' rather than needing to be available from the beginning of language.

(c) GRAMMATICAL CATEGORIES AND STRUCTURES—tense, mood, negation, number. Some of these are more effable than others (number is easier than tense or mood, for example), making it likely that different categories evolved at different times, easiest first.

10.3.2 Iteration and Cumulativity Effects

Taking the two-track view of language transmission outlined in Sect. 10.1 has enabled a new view of how several features of languages evolve. Linguistic variants compete for transmission—to be remembered and analysed in the moment by children, to be learned and to be passed on to their children. Christiansen and Chater (2016) have modelled the general impact of these iterated bottlenecks (the 'processing' and 'learning' bottlenecks) on what linguistic forms get selected for transmission, and such 'design features' of language as compositionality have been shown to arise from natural selection in the process of iterative transmission: as Smith and Kirby (2012) put it, 'compositionality can be explained as a cultural adaptation *by language* to the problem of transmission through a learning bottleneck'.

10.3.3 Embedding in Face-to-Face Interaction

Language is fundamentally adapted to face-to-face dialogue, and Levinson (2006, 2016) has argued that it is the unique propensity for face-to-face interaction, in closely linked turns, which provided the matrix within which linguistic systems evolved. As can be ascertained by watching simple but more or less successful interactions between those not sharing a common language (young children, foreigners), a great deal can be communicated thanks to the human ability to attribute communicative intentions to their partners. This is how parents, or carers of old people, are able to extract sensitive speech-intentions from their truncated utterances.

It is thus plausible that this uniquely human way of interacting evolved first, then providing a context in which participants could enrich signs (sounds, gestures) with meaning whose contextual enrichment is then generalised into other contexts and users, leading to a gradual evolution in the richness of the linguistic system (cf. Keller 1998). Observed structures arise, through time, by summing the outcomes of many individual communicative interactions (Haspelmath 1999). The human propensity

for cultural transmission (Tomasello 2008) allows innovations to be pooled across individuals and through historical time, 'ratcheting up' the level of complexity in the system. The scenario sketched in a nutshell here provides a framework by which gradual enrichments in linguistic systems can occur through time, and be transmitted as ever-more-sophisticated cultural products, arising in the first place in situations (face-to-face interaction) which can occur with little or no need for initial linguistic structuring.

10.3.4 Horizontal Transmission

Like viruses and bacteria, but unlike higher organisms, languages allow a great deal of horizontal transmission (in the preceding sentence *virus*, *bacteria*, *organism*, *language*, *horizontal* and *transmission* are all words that have come into English not from its ancestral language proto-Germanic, but from Classical Greek, Latin and French in various combinations). This means that a full understanding of human language evolution needs to draw on tools that examine the development of characteristics both through regular mechanisms of phylogenetic descent and through horizontal transmission across lineages—often labelled by linguists, rather misleadingly, as 'borrowing'.

The susceptibility of each language to enrichment through the adoption of words and structures from others means that, as a result of what is likely to have been primal multilingualism, innovations arising in particular languages (such as any of those mentioned in Sects. 10.3.1 and 10.3.2) could have been pooled across small intermarrying populations, through multilingual contact (Evans 2018). Since elements of the total package are partly independent of each other, it is possible that they were 'invented' separately in different groups speaking different early languages, and gradually integrated into a compelling package through language contact (in addition to 'independent' emergence in different groups). This is rather similar to the socially distributed emergence of modern 'farming packages', or of complex systems of hardware and software like the Internet, which bring elements innovated by different groups together at different places and times. Not only does this scenario gel with a more gradualist approach to language evolution, it allows us to investigate the evolution of individual phenomena (say, tense systems, or relative clause structures) and see how they get integrated with other linguistic elements.

10.3.5 Polysemigenesis

A classic but sterile debate in the evolution of language literature is the opposition between 'monogenesis' (language evolved once, in one place/population) versus

'polygenesis' (it evolved more than once, in several places/populations). An alternative, suggested by the gradualist, integrationist, scenario above, is 'polysemigenesis'—particular elements or suites thereof could have evolved independently in different places, gradually being integrated through horizontal transmission into the 'modern language package'. This is the position I will assume in the rest of this chapter. It provides an account that gives some purchase to standard models of evolution by allowing the gradual assembly of a complex phenomenon. But it is also demographically more plausible in that, however small the population of our ancestors was at the tightest bottleneck, by extrapolation from existing hunter-gatherer populations which rarely exceed 1000 speakers per language, we can say that even when human or early populations were at their smallest there would have been enough people to support hundreds of tribe or clan groups, and as many (semi-)languages.

10.4 Evolution Still at Work: Why Are Modern Languages so Diverse

A further consequence of the evolutionary scenario set out in Sect. 10.3 is that it links the evolutionary mechanisms found at the first emergence of human language to those still driving language evolution today. The world's languages are incredibly diverse (Evans and Levinson 2009), not only making the Chomskyan notion of a single 'universal grammar' look increasingly untenable but also furnishing rich data for studying the mechanisms of language evolution (Evans 2013).

Accounting for how this diversity evolves is one of the main challenges for linguistics.

To do so, we must draw in a search for a wide range of selectional mechanisms, a question that works best if we employ coevolutionary models with many different types of selector—physiological, acoustic, cognitive, sociocultural, demographic, system-integrational (i.e. how subsystems can work with each other compatibly) and so forth.

10.4.1 Linguistic Diversity: An Illustration

To briefly illustrate the extent of this diversity, consider how the concept 'know' is expressed across a range of languages (Fig. 10.2):

Here we see differences at every level of linguistic structure.

First, at the level of *sounds*, we see such differences as tone (Chinese only), pitch accent (Japanese only), palatalisation of consonants (Russian only) and retroflex *zh* (Mandarin only). Five independent histories are represented across these six words—only English *know* and Russian *znat'* are homologies ('cognate' in linguistic terms),

English	French	Russian	Japanese	Mandarin	Tamil	Kayardild	Dalabon
know	savoir	Знать znat'	知る shirú	知道 zhīdao	தெரி teri	mungurru	bengkan

Fig. 10.2 Expressing 'know' in eight languages (English, French and Russian represent three branches of the Indo-European family; Japanese, Mandarin and Tamil each represent a distinct family (Japanese is an isolate without proven relatives, Mandarin is Sino-Tibetan and Tamil is Dravidian), while Kayardild and Dalabon represent two branches of the Australian family.)

deriving from a common ancestral word (and note that the 'silent' *k* of the English *know* preserves evidence of a sound that used to be there).[3]

Second, at the level of the writing systems used to represent them, we have quite different organisational logics: the Latin alphabet in English and French (recently adapted to the writing of the Aboriginal languages Kayardild and Dalabon), the distantly related Cyrillic alphabet in Russian, an abugida in Tamil (in which symbols like த/ta/have an inherent vowel, which can then be overridden by adding another symbol like the loop to its left which changes the/a/into an/e/), a character system in Chinese, and a hybrid system in Japanese which mixes characters with syllable symbols (る is pronounced *ru* without reference to meaning). Note also that the grapheme 知 in Japanese reflects horizontal transfer from Chinese: it represents the meaning 'know' but with a different pronunciation.

Third, there are many differences in the grammar of these words. The English, Mandarin and Kayardild words can be used without change in a normal sentence (*I know*; *wǒ zhīdao*, *ngada mungurru* all simply combine the word for 'I' with the word for 'know'). In Russian or French, the verb needs to undergo substantial change through inflection (*je sais*; *ja znaju*), and in Dalabon, it is possible to build morphologically complex words off the basic form (*djehlngwungurrubengkan* 'Then you and I, being in odd-numbered relations to each other, recognise whose spirit it is'). There are also differences in the part-of-speech—Kayardild *mungurru* is an adjective instead of a verb, much like the English adjective *knowledgeable*—but it really means 'know', and unlike the English word it is a basic rather than a derived word. Additionally, there are differences in 'argument structure'—in English, Russian, etc., 'know' is a transitive verb, taking the knower as the subject, while in Tamil the structure is more like 'To Kumar this place knows'.

Fourth, although I have chosen the best available translation equivalents, they don't always mean quite the same thing. This may reflect (i) 'contrast sets' distinguishing different types of knowledge, e.g. French *savoir* 'know (a fact)' *connaître* 'know (a person, situation)'; Russian *znat'* 'know (a fact), *umet'* 'know (a procedure)', (ii) differences in the temporal organisation of knowledge: Japanese *shiru* is better translated as 'come to know', with the resultant state of knowing shown by verbal inflection and (iii) different entailments about factuality—Dalabon *bengkan* is more

[3] 'The frequent presence of rudiments, both in languages and in species, is still more remarkable… In the spelling … of words, letters often remain as the rudiments of ancient forms of pronunciation' (Darwin 1871:60).

like ‘think, have in mind’ and factivity can be cancelled by following it with another word meaning something like ‘believedly’.

This little vignette illustrates how much languages can vary just in one word (and we’ve only used eight of the world’s 7000 languages). Add up the tens of thousands of words in their vocabularies, the thousands of sounds drawn up differentially across the world’s languages, and the differences at every level in how syntax is organised and you begin to get some idea of how variable human languages can be—in ways that directly parallel the diversity of biological species that has driven the development of evolutionary biology.

10.4.2 Diversity and Evolutionary Approaches

The great diversity of contemporary languages calls for a body of explanatory theory that accounts for how such extraordinarily different solutions to the need for functioning human communicative systems have arisen.[4] It may come as a surprise to evolutionary biologists, used to invoking adaptation as a prime factor driving diversification, that explanations of this type have long remained on the margins in linguistics.

Instead, there has been a focus on the interplay of ‘internal’ factors—namely the inherent instabilities and numerous interactions that occur between the elements of complex systems, with variability generated by trade-offs between the adaptive advantages of particular subcomponents. A simple example is the trade-off between maintaining ‘faithfulness’ of form (saving learners and users from variant forms) and ‘ease of articulation’, smoothing out the muscular gestures producing the sound. If we illustrate with negative prefixes to words starting with the ‘alveolar set’ (t/d/n) versus the ‘bilabial set’ (p/b/m), we can see that German favours faithfulness (*undeutlich* ‘indistinct’, *unbelebt* ‘unpopular’) while Italian favours ease of articulation (*indistinto* ‘indistinct’ but *imbattibile* ‘unbeatable’); English compromises by favouring the first solution for words of Germanic origin (*unbelievable*) but the second for words of Latin or Italian origin (*imperceptible*, *immortal*).

As has long been known in engineering theory, ‘the number of equally efficient designs for an artifact generally is proportional to both the number and the complexity of the tasks tha[t] an artifact must perform’ (Niklas 1994, 6772). Given that language is the human multipurpose communication technology par excellence (cf. Dor 2015),

[4] After several decades during which the notion of ‘Universal Grammar’ was vigorous promoted, there remain linguists who see linguistic diversity as a marginal phenomenon: ‘Without proceeding, it seems to me no longer absurd to speculate that there may be a single internal language, efficiently yielding the infinite array of expressions that provide a language of thought. Variety and complexity of language would then be reduced to the lexicon, which is also the locus of parametric variation, and to the ancillary mappings involved in externalisation, which might turn out to be best possible solutions to relating organs with independent origins and properties’ (Chomsky 2007:25). For those taking this view, there is no need to explain diversification, since they regard it as a minor phenomenon. For arguments against this position, see Evans and Levinson (2009).

this line of argument has undoubted plausibility and variants of it have appeared in many forms, from 'competing motivations' (Du Bois 1987) within the discourse-functional tradition to differential weightings of various factors within Optimality Theory (Anttila and Cho 1998). Similar insights have of course come from studies of simulated biological evolution, where 'morphological diversification became easier on complex as opposed to simple fitness landscapes. Likewise, it is biologically reasonable to suppose that the morphological diversity manifested by extant species occupying similar or identical habitats vouchsafes that very different phenotypes can have equivalent capacities for growth, survival, or reproductive success' (Niklas 2004, 65).

However, no biologist would claim that all species diversity comes from such causes—as the preceding quote indicates, ecological niches vary widely, bringing in 'external' factors as a second[5] and arguably even more powerful cause of diversification in evolution.

In the history of linguistic ideas, there has been a long neglect of such 'external' factors, going back to Franz Boas's early twentieth-century rejection of any suggestion that genetic differences may underlie linguistic ones, based on the 'obvious' fact that a child of any racial background can acquire total mastery of any language provided they are exposed to it from early life. Perhaps more surprisingly, even Benjamin Lee Whorf, famous for his views of how language influences thought, rejected any suggestion that culture may impact on language, moving linguistics away from the emphasis on cultural influences that had characterised the Romantic period from Giambattista Vico in Italy to Johann Gottfried Herder and then Wilhelm von Humboldt in Germany.

Among the external factors that have been marginalised or rejected are genetic, epidemiological, climatic/geographic, demographic, sociolinguistic, and cultural factors. Instead, there has been a tendency to attribute explanations for diversity to properties of the human cognitive, perceptual and articulatory systems (all presumed to be invariant across human populations). Coevolutionary approaches, sketched in the next section, are finally linking studies of language diversity back to the study of human cultural, demographic, epidemiological and environmental diversity in a way that re-engages with fundamental questions about human history and the diversity of cultural responses to the myriad of circumstances in which speaking humans have evolved their linguistic instruments. With computational simulations demonstrating how easily repeated iteration (Kirby et al. 2007; Dediu 2011) can turn small selector biases into quite different evolutionary outcomes, and clear models of how the processing and learning bottlenecks can lead the evolution of linguistic systems in particular directions (Christiansen and Chater 2016), we are no longer in thrall to Boas's observation about equal human potential to learn any language: small biases

[5]There are others, of course, such as historical contingency: next evolutionary steps are constrained by present structures. Just as preexistent anatomy determines what elements will evolve into a wing, preexisting linguistic structures determine much of what elements will evolve into a future tense marker, an/s/sound, or a complementiser, even when the broad selectional pressures are comparable.

can nudge language evolution in different directions without, in general, creating significantly observable differences in learnability across different populations.

In the next two sections, we discuss one coevolutionary case study each from the realms of phonetics and its relation to population genetics (Sect. 10.4.3)—'language diversity seeded by genetic variation' (Levinson and Dediu 2013:228), and from how culture, through a range of communicative practices, can shape a language's semantico-grammatical structures (Sect. 10.4.4).

10.4.3 Phonetics and Population Genetics

Thousands of distinguishable speech sounds are found across the world's languages, but most languages work with a selection of fifty or less, and even the most generous analysis of the most exuberant language, Taa (!Xóõ) of Namibia and Botswana, only counts 87 consonants, 20 vowels and two tones. For any human language, therefore, and likewise up the phylogenetic tree to the families they belong to, we face the question: why this subset of sounds rather than others. For example, why have some languages evolved tone, some click sounds and some retroflex sounds?

Early proposals (Vendryès 1902; Brosnahan 1961) that some of the diversity in languages' sound systems may reflect interpopulation differences in vocal tract were dismissed on the Boasian grounds that people of any race, if raised in a language from birth, can learn to speak it without accent. However, as mentioned above the amplifying effect of hundreds of generations of slight biases means the two positions are not incompatible. Small interpopulation differences in anatomy, amplified by iteration over many generations, may exercise differential selection pressure influencing which speech sounds emerge, but the child's adaptability to its environment still allows it to learn whichever language it grows up in.

A first salvo in the reinstatement of population-genetic factors as a possible explanation was fired in Dediu and Ladd's (2007) study of weak genetic biases and the evolution of tone. Examining the population frequency of two alleles associated with pitch perception, ASPM and microcephalin, across a database containing 983 alleles and 26 linguistic features in 49 populations, they found correlations that go beyond the

> usual explanatory factors represented by geography and history. The relationship between genetic and linguistic diversity in this case may be causal: certain alleles can bias language acquisition or processing and thereby influence the trajectory of language change through iterated cultural transmission. (Dediu and Ladd 2007)

More recently, the relationship between palate shape and the evolution of clicks has been examined (Moisik and Dediu 2017) using biomechanical simulations of click-production against different palate shapes. The palate shapes found in !Xóõ populations, with no prominence on its alveolar ridge, facilitates the production and learning of click sounds (Fig. 10.3, left), where European-style palates, with prominences on their alveolar ridges (Fig. 10.3, right), inhibit the production and

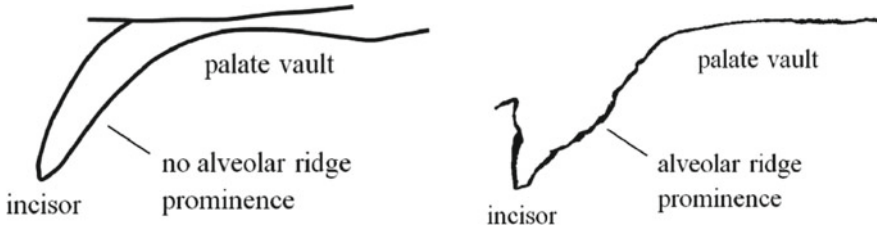


Fig. 10.3 Typical palate shapes in the !Xóõ (left) and European (right) populations (Moisik and Dediu 2017)

learning of click sounds. It is no coincidence, this study suggests, that click languages are found among populations with !Xóõ-style palate shapes and not European ones, rather than vice versa—though we await worldwide studies where we can match anatomical population facts against most of the world's languages.

Studies like these suggest that, at the level of phonetic diversity, at least some of the variability found across languages may have an underlying physiological, and ultimately a genetic, basis: 'the naturalness of sound change may not be universal, but rather, to a lesser or greater extent (depending on the sounds in question), vary from one community to the next' (Dediu and Moisik 2019). Beyond genetics, many other factors may also shape the external fitness landscape of phonetic evolution, including, as mentioned above, altitude (Everett 2013) and diet (Blasi et al. 2019).

10.4.4 Grammatico-Semantics and Culture: The Evolution of Numerals

Numeral systems lie at the intersection of grammar (being generative systems exhibiting compositional combinations of signs) and lexical meaning (forming words listed in dictionaries, and cutting up the space of meaning into different bundles in different languages). The vast majority of the world's languages organise the upward extension through powers of numerals which are based on our digits, in the etymologically original sense—fingers of one hand (base 5, e.g. many Austronesian languages) or both (base 10, English, Chinese, etc.), sometimes augmented by the toes (base 20, Mayan), or using the fingers of one hand to count iterations to twelve on the other hand, by which the thumb points in turn to each bone of the other four fingers (base 60, Sumerian). All humans have the same anatomical apparatus in their hands and feet, so the choice between these systems is rather arbitrary. On the other hand, systems using powers outside these are rare, and call for special explanation.

One unusual 'senary' system, using powers of six, is found in just one part of the world—Southern New Guinea (Donohue 2008; Evans 2009; Hammarström 2009). Illustrating with the Nen language, there are unitary power terms of six up to the fifth power—*pus* '6', *pérta* '6^2, 36', *taromba* '6^3, 216', *damno* '6^4, 1296' and *weremaka*

'6^5, 7776'. Other numbers are expressed as multiples of successively smaller powers (with *kp* 'fruit; unit' used for multiples of 1): 75, for example, is *sombes pérta nambis kp* '2 (×) 6^2 (+) 3 × 1'. Given that the hands of Nen speakers, like those of English or Chinese speakers, have five fingers, how could such a system have evolved?

Now it is clear that these senary systems evolved from yam-counting procedures, which is the dominant context in which they are used across the region. Yam-counting, at the end of each harvest, is a major public event, validating the worth of each household as hard-working, competent cultivators—each family should store at least a *damno* (1296) of yams after a harvest, to get them through to the next year. It is carried out with the same level of public scrutiny to be found among scrutineers in a Florida election, appropriately ritualised with the beating of a drum and someone calling out the numbers as they are assembled. The counting procedure shows how the base-six system could arise: a pair of men, each carrying three yams, deposits them together in groups of six, accompanied by calls of 'one six; two six(es); three six(es)' etc.[6] Once six sixes have been assembled in this fashion to give a *pérta* (36), a single counter yam is removed to form a meta-pile that counts units of 36. In this manner, the yam-counting procedure forms a sort of embodied algorithm that makes powers of six a natural set of units to use.

Here, then, we have an example of a culturally specific fitness landscape—the dominance of yams as the units to be counted, and the particular ritual used for counting them with its groupings into six and counter yams for iterating the enumeration of powers of six—shaping the direction in which numerals evolve in a particular group of languages. In the spirit of Vico, Herder and Humboldt, then, we can see culture as a further type of fitness landscape against which languages evolve and diversify (see Evans (2003) and the chapters in Enfield (2002) for many further examples). As Spitzer (1947:2) put it, 'of all linguistic branches, it is in semantics that the changes due to cultural development can best be seen at work, for 'meaning' is the best barometer of cultural climate'.

10.5 Conclusions

The long history of evolutionary approaches to language, dating back to the very beginning of Darwin's formulation of the theory of evolution, is now flourishing again with renewed vigour, thanks to many factors that include the reconceptualisation of linguistic variability from noise to signal, the growing availability of materials on an ever-growing proportion of the world's linguistic diversity and the application of computational methods from evolutionary biology to the study of language evolution, across a range of phenomena and scales.

[6]See the video made by Christian Döhler with the related Komnzo language for an illustration: https://vimeo.com/54887315, as well as Williams (1936, 225–7) for an early ethnographic description.

In this chapter, I have focussed on three key questions—how we can account for the original evolution of language, how we can model the ongoing processes of linguistic diversification and how we can bring in coevolutionary models to relate the exuberant differences among the world's languages to the different fitness landscapes they are spoken in, whether we characterise these in biological, ecological, cultural or demographic terms. But there are many other questions where evolutionary approaches are being applied, such as the building of phylogenetic models of relatedness among the world's seven thousand languages—which will help us understand the deep history of the populations that speak them—and the 'micro-to-macro' problem of scaling up from examinations of individual and contextual differences in contemporary speech communities to the architectonic processes that have given us language families of great genealogical depth.

Language is far from being the only phenomenon outside biology that lends itself to evolutionary processes. Many other cultural manifestations, from kinship to systems of political organisation to music to cat's cradle games, are amenable to evolutionary approaches. The coming decades are likely to see not just increasing applications of evolutionary models to these particular phenomena, but also a widening of evolutionary theory to the point where we have an articulated approach that includes both general principles (applying to evolution in all domains) and specific principles associated with particular phenomena (e.g. the biological realm, the linguistic realm). In this process of theoretical extension, language is a particularly suitable testing ground because of the large and diverse number of languages, their structural complexity and the computational tractability of many of the problems faced in explaining their change and diversification.

Acknowledgements The ideas here were presented at the 22nd Evolutionary Biology Meeting at Marseilles in September 2018, and I thank Pierre Pontarotti for his kind invitation to attend this most stimulating conference, as well as to the audience members for their questions. I would also like to thank Damián Blasi, Lindell Bromham, Bill Croft, Dan Dediu, Mark Ellison, Russell Gray, Steve Levinson, Ron Planer and Kim Sterelny for discussions bearing on the contents of this paper, Aung Si for drawing Fig. 10.1, Susan Ford for assistance with editing and the Australian Research Council for support of the work reported on here, in particular through grants FL130100111 'The Wellsprings of Linguistic Diversity' and CE140100041 'Dynamics of Language'.

References

Anttila A, Cho YY (1998) Variation and change in optimality theory. Lingua 104:31–56

Atkinson Q, Gray R (2005) Curious parallels and curious connections: phylogenetic thinking in biology and historical linguistics. Syst Biol 54(4):513–526

Blasi DE, Moran S, Moisik SR, Widmer P, Dediu D, Bickel B (2019) Human sound systems are shaped by post-neolithic changes in bite configuration. Science 363(6432)

Brosnahan LF (1961) The sounds of language: an inquiry into the role of genetic factors in the development of sound systems. W. Heffer & Sons, Cambridge

Chomsky N (1988) Language and problems of knowledge: the Managua lectures. MIT Press, Cambridge

Chomsky N (2007) Of minds and language. Biolinguistics 1:9–27
Chomsky N, Berwick RC (2016) Why only us: language and evolution. MIT Press, Cambridge
Christiansen MH, Chater N (2016) Creating language: integrating evolution, acquisition, and processing. MIT Press, Cambridge
Croft W (2008) Evolutionary linguistics. Annu Rev Anthropol 37(1):219–234
Darwin C (1859) The origin of species. John Murray, London
Darwin C (1871) The descent of man, and selection in relation to sex. John Murray, London
Dediu D (2011) Are languages really independent from genes? If not, what would a genetic bias affecting language diversity look like? Hum Biol 83:279–296
Dediu D, Ladd DR (2007) Linguistic tone is related to the population frequency of the adaptive haplogroups of two brain size genes., ASPM and microcephalin. Proc Natl Acad Sci USA 104:10944–10949
Dediu D, Levinson SC (2018) Neanderthal language revisited: not only us. Curr Opin Behav Sci 21:49–55. https://doi.org/10.1016/j.cobeha.2018.01.001
Dediu D, Moisik S (2019) Pushes and pulls from below: anatomical variation, articulation and sound change. Glossa (A Journal of General Linguistics) 4(1):7.1–33. https://doi.org/10.5334/gjgl.646
Dediu D, Cysouw M, Levinson S, Baronchelli A, Christiansen MH, Croft W, Evans N, Garrod S, Gray R, Kandler A, Lieven E (2013) Cultural evolution of language. In: Richerson PJ, Christiansen MH (eds) Cultural evolution: society, technology, language, and religion. MIT Press, Cambridge, MA, pp 303–332
Dennett D (1995) Darwin's dangerous idea. Penguin, London
Donohue M (2008) Complexities with restricted numeral systems. Ling Typol 12:423–429
Dor D (2015) The instruction of the imagination. Oxford University Press, Oxford
Du Bois J (1987) The discourse basis of ergativity. Language 55:59–138
Durham W (1991) Coevolution: genes, culture and human diversity. Stanford University Press, Stanford
Ehrlich P, Raven PH (1964) Butterflies and plants: a study in coevolution. Evolution 18:586–608
Enfield NJ (2002) Ethnosyntax. Oxford University Press, Oxford
Evans N (2003) Context, culture and structuration. Annu Rev Anthropol 32:13–40
Evans N (2009) Two *pus* one makes thirteen: senary numerals in the Morehead–Maro region. Ling Typol 13(2):319–333
Evans N (2013) The diversity of languages as a resource for studying cultural evolution. In: Richerson PJ, Christiansen M (eds) Cultural evolution: society, technology, language, and religion. Strüngmann Forum Reports, vol 12. MIT Press, Cambridge, MA, pp 233–268
Evans N (2016) Typology and coevolutionary linguistics. Ling Typol 20(3):505–520
Evans N (2018) Did language evolve in multilingual settings? Biol Philos. https://doi.org/10.1007/s10539-018-9609-3
Evans N, Levinson SC (2009) The myth of language universals. Behav Brain Sci 32(429–448):472–492
Everett C (2013) Evidence for direct geographic influences on linguistic sounds: the case of ejectives. PLoS ONE 8:e65265
Everett DL (2005) Cultural constraints on grammar and cognition in Pirahã. Another look at the design features of human language. Curr Anthropol 46:621–646
Hammarström H (2009) Whence the Kanum base-6 numeral system? Ling Typol 13:305–319
Haspelmath M (1999) Optimality and diachronic adaptation. Z Sprachwiss 19(2):180–205
Keller R (1998) A theory of linguistic signs. Oxford University Press, Oxford
Kirby S, Dowman M, Griffiths TL (2007) Innateness and culture in the evolution of language. Proc Natl Acad Sci USA 104:5241–5245
Levinson SC (2006) On the human "interaction engine". In: Levinson SC, Enfield NJ (eds) Roots of human sociality: culture, cognition and interaction. Berg, Oxford New York, pp 39–69

Levinson SC (2016) Turn-taking in human communication, origins, and implications for language processing. Trends Cogn Sci 20(1):6–14. https://doi.org/10.1016/j.tics.2015.10.010
Levinson SC, Dediu D (2013) The interplay of genetic and cultural factors in ongoing language evolution. In: Richerson PJ, Christiansen M (eds) Cultural evolution: society, technology, language, and religion. Strüngmann Forum Reports, vol 12. MIT Press, Cambridge, MA, pp 219–232
Maddieson I, Coupé C (2015) Human language diversity and the acoustic adaptation hypothesis. Proc Mtgs Acoust 25:060005. https://doi.org/10.1121/2.0000198
Meyerhoff M (2015) Introducing sociolinguistics, 2nd edn. Routledge, London
Moisik S, Dediu D (2017) Anatomical biasing and clicks: evidence from biomechanical modeling. J Lang Evol 2(1):37–51
Niklas KJ (1994) Morphological evolution through complex domains of fitness. Proc Natl Acad Sci USA 91:6772–6779
Niklas KJ (2004) Computer models of early land plant evolution. Annu Rev Earth Planet Sci 32:47–66
Pinkster H (1985) The development of future tense auxiliaries in Latin. Glotta 63(3/4):186–208. http://www.jstor.org/stable/40266718
Schleicher A (1869) Darwinism tested by the science of language, translated by AVW Bikkers. John Camden Hotten, London
Smith K, Kirby S (2012) Compositionality and linguistic evolution. In: Hinzen W, Machery E, Werning M (eds) The Oxford handbook of compositionality. Oxford University Press, Oxford
Spitzer L (1947) Essays in historical semantics. Russell & Russell, New York
Tomasello M (2008) Origins of human communication. MIT Press, Cambridge
Trudgill P (2011) Sociolinguistic typology. Oxford University Press, Oxford
Vendryès J (1902) Some thoughts on sound laws. In: Keller AR (ed) A reader in historical and comparative linguistics. Holt, Rinehart & Winston, New York, pp 109–120
Wallace AR (1867) Creation by law. Q J Sci 4:471–488
Williams FE (1936) Papuans of the trans-fly. The Clarendon Press, Oxford

Part III
Genome and Phenotype Evolution

Chapter 11
The Parallel Lives of Human Y Chromosome Lineages Across the Strait of Gibraltar

Carla García-Fernández and Francesc Calafell

Abstract The patrilineal transmission of the Y chromosome and the fact that diversity in it accumulates along a strict genealogy imply that, by observing the current Y chromosome diversity in men, inferences can be made about the male-mediated history of humans (exactly like the female-mediated history is traced by mitochondrial DNA). By resequencing the non-recombining portion of the Y chromosome, it has been recently recognized that many branches of the Y chromosome genealogy (the so-called *haplogroups*) have expanded recently in bursts often tied to lifestyle changes or technological innovations. We have analysed two such bursting haplogroups: R1b-DF27 and E-M183. R1b-DF27 is prevalent in the Iberian Peninsula (40–70%), while its frequency drops to <15% north of the Pyrenees. We have estimated its age at ~4000 years ago, in line with ancient DNA discoveries and pointing to the population upheaval of the Bronze Age as a plausible agent in its origin and dispersion. Similarly, across the Strait of Gibraltar, E-M183 (equivalent to E-M81) dominates the Y chromosome landscape in NW Africa (up to 70%) and is much rarer elsewhere. However, its origins are more recent, at ~2000 years ago, and more difficult to pin down to a particular historical event. In conclusion, either in Iberia or the Maghreb, most men share a common ancestor that lived just a few hundred generations ago.

11.1 Introduction

The Y chromosome is a genomic oddity: it is the squat half of the only mismatched couple in the karyotype, rather smaller than its X partner, and containing fewer genes. As we shall briefly discuss below, these features arose from its function: the presence of the Y chromosome in humans, as in eutherian mammals and other animals, steers development towards the production of a male embryo. The main masculinity switch is a gene called *SRY*. The human Y chromosome only recombines

C. García-Fernández · F. Calafell (✉)
Departament de Ciències, Experimentals I de La Salut, Institute of Evolutionary Biology (CSIC-UPF), Universitat Pompeu Fabra, Barcelona, Spain
e-mail: francesc.calafell@upf.edu

P. Pontarotti (ed.), *Evolution, Origin of Life, Concepts and Methods*,
https://doi.org/10.1007/978-3-030-30363-1_11

with the X chromosome at two small regions at the ends of the chromosome, called pseudoautosomal regions 1 and 2 (PAR1 and PAR2). The non-recombining region of the Y chromosome (NRY) covers most of its ~23 euchromatic Mb; since it does not recombine with the X chromosome, it is male specific (and thus, it is also denoted MSY) and evolves independently of the X chromosome.

11.2 Structural Complexity

The Y chromosome 58 Mb of sequence are made up of five different types of regions: the pseudoautosomal parts, which recombine with the X chromosome; a large heterochromatic block; the X-degenerate region that derives from the common autosomal ancestor with the X chromosome; the more recently X-transposed regions, that are unique to humans and were transposed 3–4 million years ago; and the highly repetitive ampliconic regions, the largest sequence class of the Y chromosome, covering up to one-third of its total length. These ampliconic regions contain eight massive, segmentally duplicated palindromes, all of which share >99.9% identity between their two copies, with the largest one spanning over 2.9 Mb.

The complex structure of the Y chromosome clashes with the limitations of classic sequencing technology, making the production of data the bottleneck of this field. As a result, most of the resequencing studies have been based on a set of unique regions within the male-specific part of the Y chromosome (MSY). These 8.97 Mb of sequences defined by Wei et al. (2013) are expected to be devoid of repetitive or low-complexity regions, and thus allow for high-confidence mapping and variant detection even with short reads. The rest of the Y chromosome remains poorly characterized, possibly hiding key elements relevant for understanding human male evolution.

Currently, most popular massive parallel sequencing strategies rely on short-read sequencing, which fares poorly in structurally complex genome regions. Thus, advances in Y chromosome sequencing require overcoming this shortcoming; ideally, a definitive and complete reference for the Y chromosome should be constructed. An example of this can be found in the work by Kuderna et al. (2019), who reconstructed a human Y chromosome of African ancestry by flow-cell sorting combined with Oxford Nanopore sequencing, a technology that allows for sequencing of megabase-long reads. By obtaining larger reads from the sequencing process, they managed to achieve unprecedented quality in terms of contiguity and sequence class representation.

11.3 Y Chromosome Polymorphisms

Within the Y chromosome, different types of polymorphisms may capture different events at different time depths. Currently, Y chromosome polymorphism is studied through short tandem repeats (STRs) and single-nucleotide polymorphisms (SNPs).

STRs present a high mutation rate (Kayser et al. 1997; Gusmao et al. 2005; Burgarella and Navascués 2011); a few meioses are often sufficient to produce detectable differences. They have been widely used in forensic and genealogical studies (Redd et al. 2002; Calafell and Larmuseau 2017). In population genetics, Y-STRs often bear the mark of recent demographic events, such as recent expansions (Bosch et al. 2000; Solé-Morata et al. 2017a, b). Their relatively easy identification and analysis has favoured their use, but the robustness and accuracy of the results we can obtain are limited, especially when dealing with ancient events.

The lack of recombination along the MSY, combined with the low SNP mutation rate, makes it easy to produce a maximum-parsimony tree of SNP variation in the Y chromosome. The deepest branches of the tree are called haplogroups and exhibit a clear geographical structure (van Oven et al. 2014). Genotyping previously defined SNPs is easy and fast. It can be performed with different array platforms that provide sets of predefined positions along the entire genome, or most commonly with a custom set of probes for each informative position. However, these methods can only work with the known variation, which has usually been ascertained in a restricted set of populations, often of European descent. Moreover, the methods that are commonly used to date population events or to infer effect population sizes and their evolution require full sequence ascertainment. Thus, although genotyping SNPs and determining haplogroups has been quite a fruitful tool, e.g. in detecting and measuring male-mediated gene flow, it fails to extract most of the historical information contained in the Y chromosome.

The complete (or at least the so-called *callable*) Y chromosome sequence can be obtained as a by-product of **whole genome sequencing**. The analysis of NRY sequence combined with the autosomal and mitochondrial chromosomes provides both a complete picture of population history as well details on sex-specific differences. But this comes with a higher cost of sequencing, and a distribution of the sequencing reads along the entire genome, with the subsequent lower coverage for the Y chromosome. Moreover, when working with degraded DNA as in ancient or forensic samples, the amount and quality of the DNA are too low for whole genome sequencing. Nevertheless, once whole Y chromosome sequences are retrieved, they can be used to perform calibrated analysis and infer past events with accuracy.

The sequencing process can be optimized to produce whole NRY sequences at a higher coverage, lower cost and from degraded templates. New procedures are being developed for physically capturing the Y chromosome, such as **flow cytometry, laser microdissection** and **target enrichment** (Alvarez-Cubero et al. 2018). Target enrichment of the 8.9 Mb of the Y chromosome that contains non-repetitive regions and is suitable for analysis (i.e. the *callable* region) has also been tested in ancient DNA (Cruz-Dávalos et al. 2018), with encouraging yields.

The high resolution of the haplotype tree of the male lineages built with whole sequences has produced great insight in the human past demographic events (Poznik et al. 2016). Thus, the exponential growth of the existing datasets has promoted the appearance of specialized software for haplogroup calling such as Yleaf (Ralf et al. 2018) or Yhaplo (Poznik 2016). General phylogenetic analysis software like RAxML (Stamatakis 2014) can obviously be applied to complete Y chromosome

sequences. The availability of complete sequences also allows the introduction to Bayesian evolutionary analysis that produces rooted, time-measured phylogenies (Bouckaert et al. 2014). Inferences on effective population size in time and dating demographic events such as male admixtures or expansions are also possible.

11.4 The Male Side of History

For most of its length, thus, the Y chromosome is patrilinearly transmitted *en bloc*. Then, it is subjected to the demographic specificities of men compared to women, particularly in relation to migration and effective population size. Short-distance migration rates have traditionally been higher in women than in men, since in many societies marriages were patrilocal (i.e. the bride moved to the groom's village). Also, historical inference suggests women tend more easily to cross social class, caste or ethnic barriers (Wooding et al. 2004; Verdu et al. 2009). However, long-distance migrations, usually associated to the colonization of already populated territories, are preferentially associated with males; for instance, the NRY European component in the urban, admixed populations of Latin America, is much higher than its mtDNA counterpart (see Mendizábal et al. (2008) for Cuba, among many other examples).

Especially since the Neolithic, effective population size in men has been smaller than that of women (Jobling et al. 2014), given that the accumulation of wealth by few men increases their chances to procreate (and the survival of their children), while many other men had fewer opportunities to reproduce. In extreme cases, polygyny has been practised in a number of societies (Dupanloup et al. 2003). Additionally, violence took a higher toll on men, reducing further their numbers (Kohler et al. 2017).

The different male and female human demographies have resulted in different levels of polymorphism and differentiation for the NRY and the mtDNA. The fraction of variation within human populations for Y chromosome SNPs (as opposed to variation between populations) was estimated globally at 35.5% (Seielstad et al. 1998), versus 80–85% for the autosomes and mtDNA (Barbujani et al. 1997; Romualdi et al. 2002). Also, polymorphism at the NRY seems to be dominated by relatively few lineages that expanded in bursts linked to technological and lifestyle changes (Poznik et al. 2016). These authors sequenced 10.3 Mb of the NRY in 1244 Y chromosomes sampled from 26 populations to a median haploid coverage of 4.3× . They found 60,555 biallelic SNPs, from which they constructed a phylogeny by maximum parsimony. In that phylogeny, the authors detected 20 nodes belonging to 8 haplogroups and found in five major geographical areas that fit a model of sudden expansion followed by a moderate-rate phase. All of these expansions occurred in a relatively recent time frame in human prehistory and could be linked to particular events. First, in the Americas, they observed expansion of haplogroup Q1a-M3 at ∼15,000 years ago (15 kya), the time of the initial most substantial colonization of the Western Hemisphere (Raghavan et al. 2014). In sub-Saharan Africa, two independent haplogroup E1b-M180 lineages expanded ∼5 kya, in a period before the

numerical and geographical expansions of Bantu speakers, in whom E1b-M180 now predominates (de Filippo et al. 2012). Expansions in South Asia about 4 kya were linked to the historical migration of Indo-European speakers from the western steppe into the Indian subcontinent. East Asia stands out from the rest of the Old World for its paucity of sudden expansions, perhaps reflecting a larger starting population or the coexistence of multiple prehistoric cultures wherein one lineage rarely dominated.

In western Europe, different related lineages within R1b-L11 expanded ~4.8–5.9 kya, most markedly in two episodes around 4.8 and 5.5 kya. The earlier of these times, 5.5 kya, is associated with the origin of the Bronze Age Yamnaya culture. The Yamnaya have been linked by a DNA evidence to a massive migration from the Eurasian Steppe, which may have replaced much of the previous European population (Haak et al. 2015; Allentoft et al. 2015). The later time, 4.8 kya, coincides with the origins of the Corded Ware (Battle Axe) culture in eastern Europe and the Bell–Beaker culture in western Europe.

In this chapter, we will present two such bursts that were both much more localized in scope: R1b-DF27 in the Iberian Peninsula and E-M183 in NW Africa. And while the former expanded in a time frame coincidental with the late Bronze Age, the latter, at ~2 kya, is even more recent.

11.5 R1b-DF27 in Iberia … and Out of It

The most frequent Y chromosome haplogroup in W Europe is R1b-M269, with frequencies ranging from 41% (Germany) to 83% (Ireland) (Myres et al. 2011). Direct dating from NRY sequence variation puts the origin of R1b-M269 in the Early Bronze Age, ~4500 years ago (ya) (Batini et al. 2015; Poznik et al. 2016), consistent with the growing ancient DNA record, where a surge in R1b-M269 is indeed seen at that time (Haak et al. 2015; Allentoft et al. 2015). The lack of structure of STR variation within R1b-M269 (Solé-Morata et al. 2014a; Larmuseau et al. 2014) also suggests an explosive growth.

The most important branches of R1b-M269 are R1b-U106, particularly frequent in the Low Countries and NW Germany (Myres et al. 2011; Cruciani et al. 2011), and R1b-S116 (also known as R1b-P312), which is common throughout W Europe (Myres et al. 2011). The latter trifurcates in turn into U152 (frequent in N Italy and Switzerland (Cruciani et al. 2011)), L21 (also known as M529, abundant in the British Isles (Busby et al. 2012)) and DF27 (Fig. 11.1). DF27 was first discovered by citizen scientists (Rocca et al. 2012) and, although among the burgeoning amateur genetic genealogy it is known to be frequent in Iberian populations and their overseas offshoots, few academic publications have been devoted to it. It was found in the 1000 Genome Project populations at a frequency of 49% in Iberians, 6% in Tuscans and 7% in British, and it was absent elsewhere except for admixed populations in the Americas: Colombia (40%), Puerto Rico (36%), Mexico (10%), Perú (8%), African-Americans (4%) and Afro-Caribbeans (2%) (Rocca et al. 2012; Auton et al. 2015). It was first genotyped specifically in a few Iberian populations, Brittany and Ireland as

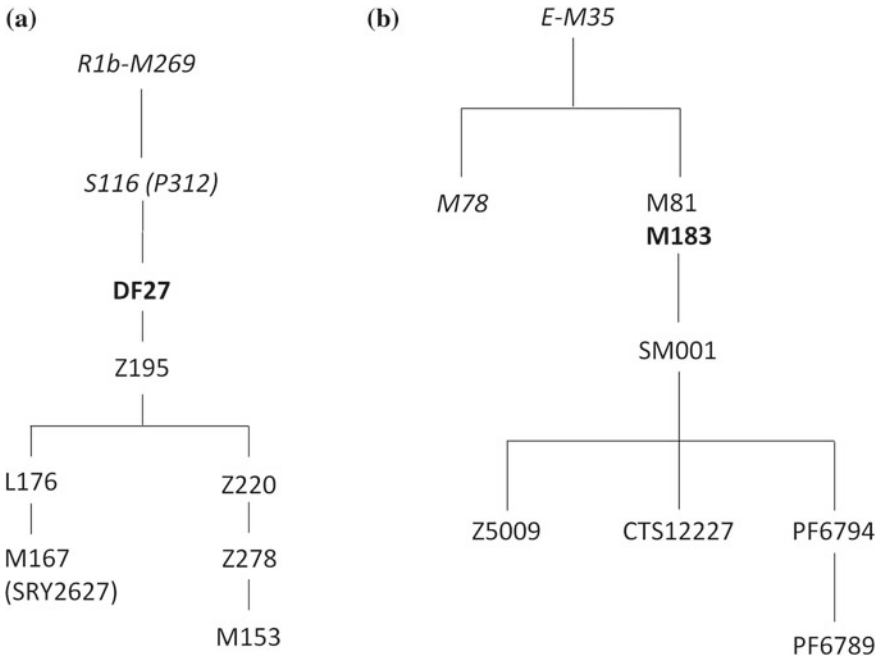


Fig. 11.1 Basic phylogenies of the haplotypes defined by different SNPs in the haplogroups R1b-DF27 (**a**) and E-M183 (**b**). Note that, in the case of E-M183, the Y chromosomes of the samples were completely resequenced, and a much larger number of SNPs were found in them

part of a study on R1b-S116 (Valverde et al. 2016), which indeed confirmed that R1b-DF27 is present at frequencies >40% in Spain and Portugal. Finally, we extended sampling to cover most of Spain and France, to refine its characterization and to describe its geographical extent and historical origin and dynamics (Solé-Morata et al. 2017b).

For this study, over one thousand individuals carrying DF27 were typed for six additional SNPs and 17 Y-STRs. DF27 itself was found at frequencies 30–50% in Iberia (with a mean of 42%), with the notable exception of native Basques, where it reached 74%. In France, it dropped to a range of 6–20% and a mean of 11%. Elsewhere, it was 15% in Britain (but <1% in Ireland) and 8% in Tuscany. Most (50–100%, with a proportion that dropped from east to west) DF27 Y chromosomes were also derived for Z195; thus, the highest frequencies of Z195 (29–41%) were reached both in the Basque Country and in E Iberia (Catalonia, Valencia), and it becomes as rare in Portugal as it is in France (Fig. 11.2).

We estimated with the weighted ρ method (Solé-Morata et al. 2017b) an age of 4190 ± 140 ya for the whole of DF27. This figure is remarkably similar both to the estimate (4128 ± 71 ya) that can be produced from whole Y chromosome sequence variability in the 88 DF27-derived individuals present overall in the 1000 Genomes Project dataset, and to the age estimated from 201 individuals in our dataset for which

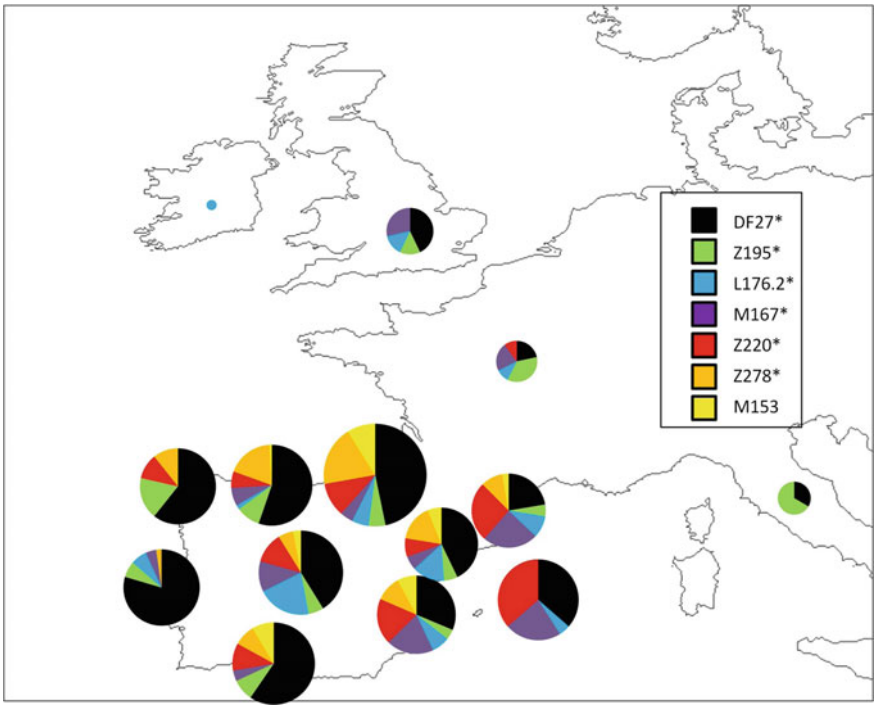


Fig. 11.2 Frequency of subhaplogroups within R1b-DF27. Pie areas are proportional to the overall frequency of R1b-DF27 in that population. As per the usual convention in human population genetics, an asterisk after a SNP name indicates a paragroup, i.e. a haplotype with the derived allele at that SNP but ancestral for all other SNPs found below the same branch. For instance, Z278* chromosomes carry the derived allele for Z278 but are ancestral for M153 (see Fig. 11.1a)

21 non-duplicated Y-STRs from the PowerPlex Y23 System were available (Purps et al. 2014) (3880 ± 165). Z195 seems to have appeared almost simultaneously within DF27, since its estimated age is actually older (4570 ± 140 ya). Of the two branches stemming from Z195, L176.2 seems to be slightly younger than Z220 (2960 ± 230 ya vs. 3320 ± 200 ya), although the confidence intervals slightly overlap. M167 is clearly younger, at 2600 ± 250 ya, a similar age to that of Z278 (2740 ± 270 ya). Finally, M153 is estimated to have appeared just 1930 ± 470 ya.

Haplogroup ages can also be estimated within each population, although they should be interpreted with caution. For the whole of DF27, the highest estimate was in Aragon (4530 ± 700 ya), and the lowest in France (3430 ± 520 ya); it was 3930 ± 310 ya in Basques.

We tested the dynamics of R1b-DF27 by means of Approximate Bayesian Computing (ABC). In particular, we compared two simple models: constant population size versus growth since time Tstart. Both the rejection and the regression method undoubtedly favoured the growth model, with associated posterior probabilities that were never lower than 0.99. The median value for Tstart has been estimated at 103

generations, with a 95% highest probability density (HPD) range of 50–287 generations; effective population size increased from 131 (95% HPD: 100–370) to 72,811 (95% HPD: 52,522–95,334). Considering patrilineal generation times of 30–35 years (Solé-Morata et al. 2015), our results indicate that R1b-DF27 started its expansion ~3000–3500 ya, shortly after its Time to Most Recent Common Ancestor (TMRCA).

With independent samples (our samples vs. the 1000 Genome Project dataset) and independent methods (variation in 15 STRs vs. whole Y chromosome sequences), the age of R1b-DF27 is firmly grounded around 4000–4500 ya, which coincides with the population upheaval in W. Europe at the transition between the Neolithic and the Bronze Age (Haak et al. 2015; Allentoft et al. 2015). Before this period, R1b-M269 was rare in the ancient DNA record, and during it the current frequencies were rapidly reached (Haak et al. 2015; Allentoft et al. 2015; Fu et al. 2016). It is also one of the haplogroups (along with its daughter clades, R1b-U106 and R1b-S116) with a sequence structure that shows signs of a population explosion or burst (Poznik et al. 2016). STR diversity in our dataset is much more compatible with population growth than with stationarity, as shown by the ABC results. Although the median time for the start of the expansion is older in R1b-DF27 than in other haplogroups and could suggest the action of a different demographic process, all HPD intervals broadly overlap, and thus, a common demographic history may have affected the whole of the Y chromosome diversity in Iberia. The HPD intervals encompass a broad time frame and could reflect the post-Neolithic population expansions from the Bronze Age to the Roman Empire (McEvedy and Jones 1978).

While it seems clear when R1b-DF27 appeared in the population, where it originated may be more difficult to pinpoint. If we extrapolated directly from haplogroup frequencies, then R1b-DF27 would have originated in the Basque Country; however, for R1b-DF27 and most of its subhaplogroups, internal diversity measures and age estimates are lower in Basques than in any other population. Then, the high frequencies of R1b-DF27 among Basques could be better explained by drift rather than by a local origin (except for the case of M153), which could also have decreased the internal diversity of R1b-DF27 among Basques.

R1b-DF27 can be used to trace male-mediated migrations out of Iberia. In particular, it has been used (Larmuseau et al. 2018) to dispel a myth about the genetic legacy of Spain in Flanders. War atrocities committed by the Spanish army in the Low Countries during the sixteenth century are so ingrained in the collective memory of Belgian and Dutch societies that they generally assume a signature of this history to be present in their genetic ancestry. Historians claim this assumption is a consequence of the so-called Black Legend and negative propaganda portraying and remembering Spanish soldiers as extreme sexual aggressors. In the sixteenth century, a rebellion started in the Low Countries, partly ignited by the political and religious hegemony of their overlord, Philip II, also King of Spain, and partly inspired by the new conflict between Catholics and Protestants (Parker and González de León 1999). Tensions increased further when the King sent an army of 10,000 soldiers under the command of the Duke of Alba from Spain to his most northern provinces in 1566 (Parker 1990a). Particularly in the period 1572–1578, several cities in the Low Countries fell victim to the so-called Spanish Furies, with mutinous and underpaid troops

plundering, raping and murdering on a large scale (Parker 1973; Soen 2008). Philip II regained control over most of the rebelling regions in the south, which left most of present-day Belgium under the reign of Spain until the Peace of Utrecht in 1713 (Parker 1990b). The northern provinces continued their resistance that eventually led to the Dutch independence, recognized by the Spanish King in 1648. During the revolt, the royal army consisting of mainly Spanish soldiers received the reputation of being a band of murderers, robbers and rapists. Historians have long pointed out that opponents to the royal regime in the Low Countries adopted the earlier existing 'Black Legend' to describe the Spanish as extremely barbarous and cruel (Swart 1975; Airey 2012). Nowadays, occasional dark hair and skin colour or pronounced eyebrows found in Belgian or Dutch people are popularly attributed to the sixteenth-century rapes. We compared 116 samples that were genealogically associated with one of the four cities in current Flanders where a Spanish Fury occurred, to other Flemish ($N = 971$) and Dutch ($N = 250$) samples. We found no significant differences in the frequencies of R1b-Z195 and R1b-SRY2627 among these three populations (Larmuseau et al. 2018), and these markers were much more frequent in Spain than in the Low Countries. This may be interpreted as exonerating the Spanish troops from sexual violence, although rape does not always lead to conception, women were often murdered afterwards during the Spanish Furies (Airey 2012), and if 'Spanish' children were born after all, it is also very likely that they might have been stigmatised and less well cared for. In any case, we can rule out any significant Spanish genetic contribution to the current genetic landscape of the Low Countries.

11.6 Across the Strait of Gibraltar: E-M183

E-M81 is a dominant Y-lineage in NW Africa, with frequencies over 50% in Morocco and Algeria, while it is much rarer in NE Africa and the Sahel elsewhere, and it is found in S Europe in areas with historical gene flow from N Africa, such as the Iberian Peninsula and Sicily (Adams et al. 2008). M183 was first described by Karafet et al. (2008) and appears to be an extremely dominant subclade within E-M81, to the point that E-M81 * (xM183) individuals are very rare; for all practical purposes, E-M81 and E-M183 can be considered equivalent (Fig. 11.1). This evidence suggests that E-M81 is an important candidate for exploring which must be considered to explore the historical and demographical processes that gave rise to current North African populations. However, little is known about the phylogeographic structure of this haplogroup, and its origin and emergence are still very controversial. While some studies pointed to a Palaeolithic origin (Bosch et al. 2001), other authors claimed that E-M81 may have a Neolithic origin (Arredi et al. 2004). The most likely scenario, as suggested by Fadhlaoui-Zid et al. (2013), is that the origin of E-M81 is more recent than the Neolithic, as previously reported.

We analysed whole Y chromosome sequences from 32 North African individuals selected by carrying the derived allele at M183 (Solé-Morata et al. 2017b). The analysis of whole Y chromosome data enabled the characterization of five new subclades

within this Y chromosomal branch (Fig. 11.1). Although four of these five SNPs had previously been described, to the best of our knowledge this is the first academic publication describing their phylogenetic relationships and population frequencies. The first variant, named SM001, clusters together all E-M183 samples but one, a sample from Lebanon. Another subclade is defined by Z5009, which groups two Moroccan samples, and one sample each from the western Sahara, Libya, Palestine and France. CTS12227 is shared by three Berber samples (two Tunisian Berbers and one Zenata Berber), two Arab Algerians and a sample from Saudi Arabia. Finally, three Libyan samples, two Moroccan and one Tunisian are derived for PF6794, and all but one of these samples are also derived for PF6789.

In order to study the geographical structure and phylogenetic robustness of E-M183, we genotyped the five subclades described using whole Y chromosome data in 250 North African samples (Fig. 11.3). We also analysed the genetic diversity within and between North African populations by genotyping the set of Y-STRs contained in the AmpFlSTR®YFiler® kit.

We have estimated with ρ the divergence of the E-M183 branch from its sister, E-M78, around 9700 ya when using a fast mutation rate and ~12,700 ya when a slow mutation rate is considered. The TMRCA of a certain haplogroup can provide some constraints on the time of its spread. Here, we used both Y chromosome SNP and STR data to obtain those estimates. Regardless of using a Bayesian or a Rho-based approach, our findings when using SNP data suggest that E-M183 originated around 2000 ya. It is worth noting that when the tree is calibrated with a slow mutation rate (Fu et al. 2014; Poznik et al. 2016), the TMRCA of E-M183 reaches ~3000 ya. However, age estimates computed using STR data strongly support that the coalescence time for this haplogroup is around 2000 ya. We have also computed the coalescence times for each subclade by using Y-STRs. The TMRCAs of E-SM001,

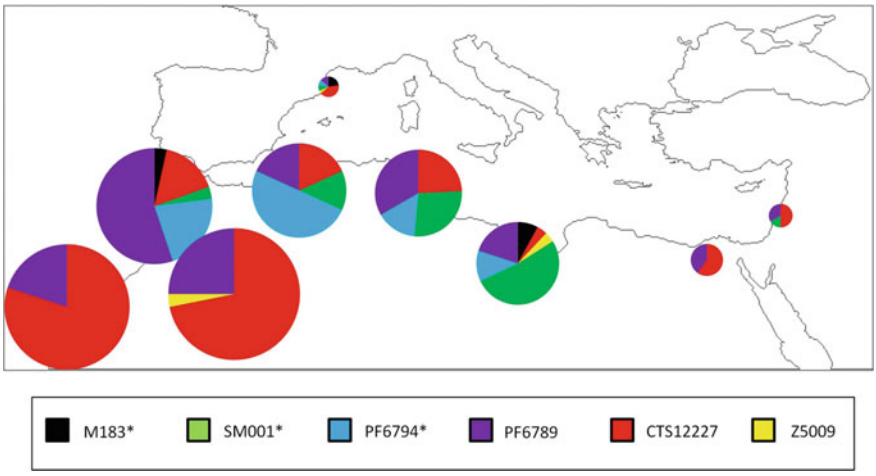


Fig. 11.3 Frequency of subhaplogroups within E-M183. As above, asterisks indicate paragroups. Pie areas are proportional to the overall frequency of E-M183 in that population

E-CTS1227 and E-Z5009 are all ~2000 ya, and their confidence intervals broadly overlap with each other and with that of the whole of E-M183, pointing to a rapid radiation. On the contrary, E-PF6794 and its subclade E-PF6789 appear to be more recent, at ~1500 ya. Interestingly, E-PF6789 is present in most of North Africa and the Near East; if, as discussed below, E-M183 may have expanded from east to west, then ~1500 ya sets an upper limit for this expansion. We found that Y-STRs are extremely homogeneous across E-M183 subhaplogroups, with the same haplotype shared by samples belonging to different subclades. This extreme homogeneity could be attributed to a recent and rapid radiation of this Y chromosomal branch (Larmuseau et al. 2014; Solé-Morata et al. 2014b), which is also seen in the fact that most of its subclades seem to have appeared almost simultaneously (Solé-Morata et al. 2017a).

Although the dating of this lineage has been controversial, with dates proposed ranging from Palaeolithic to Neolithic and to more recent times (Bosch et al. 1997; Arredi et al. 2004; Fadhlaoui-Zid et al. 2013), our results suggested that the origin of E-M183 is much more recent than was previously thought. While other studies have relied only on STR data to provide time estimates, we used Y chromosomal sequence data to calculate the TMRCA for E-M183. In addition to the recent radiation suggested by the high haplotype resemblance, the pattern showed by E-M183 implied that subhaplogroups originated within a relatively short time period, in a burst similar to those happening in many Y chromosome haplogroups (Poznik et al. 2016).

Regarding the geographical origin of E-M183, a previous study (Arredi et al. 2004) suggested that an expansion from the Near East could explain the observed east–west cline of genetic variation that extends into the Near East. Indeed, our results also showed a reduction in STR heterozygosity towards the west, which may be taken to support the hypothesis of an expansion from the Near East. In addition, previous studies based on genome-wide SNPs (Arauna et al. 2017) reported that a North African autochthonous component increases towards the west, whereas the Near Eastern decreases towards the same direction, which again support an expansion from the Near East. However, our correlations should be taken carefully because our analysis includes only six locations on the longitudinal axis, none from the Near East. As a result, we do not have sufficient statistical power to confirm a Near Eastern origin. In addition, rather than showing a west-to-east cline of genetic diversity, the overall picture shown by this correlation analysis suggests overall low genetic diversity in western Sahara, which indeed could be also caused by the small sample size ($n = 26$) in this region. Alternatively, given the high frequency of E-M183 in the Maghreb, a local origin of E-M183 in NW Africa could be envisaged, which would fit the clear pattern of longitudinal isolation by distance reported in genome-wide studies (Henn et al. 2012; Arauna et al. 2017). Moreover, the presence of autochthonous North African E-M81 lineages in the indigenous population of Canary Island strongly points to North Africa as the most probable origin of the Guanche ancestors (Fregel et al. 2009). This, together with the fact that the oldest indigenous individuals have been dated 2210 ± 60 ya, supports a local origin of E-M183 in NW Africa. Within this scenario, it is also worth noting that the paternal lineage of an early Neolithic Moroccan individual appeared to be distantly related to the typically North African

E-M81 haplogroup (Fregel et al. 2017), suggesting again a NW African origin of E-M183. A local origin of E-M183 in NW Africa >2200 ya is supported by our TMRCA estimates, which can be taken as 2000–3000, depending on the data, methods and mutation rates used.

In summary, either in Iberia or the Maghreb, most men share a common ancestor that lived just a few hundred generations ago. While human genealogies do tend to coalesce quite quickly (Rohde et al. 2004), they can do so through many different paths, alternating fathers and mothers. In this case, probably due to male-specific demographic properties, it is precisely through exactly one of these possible paths, namely the patrilineage, that these men are related to each other. And given the lack of recombination in the MSY, unlike in autosomes, we can trace this path for common descent.

References

Adams SMM, Bosch E, Balaresque PLL et al (2008) The genetic legacy of religious diversity and intolerance: paternal lineages of Christians, Jews, and Muslims in the Iberian Peninsula. Am J Hum Genet 83:725–736. https://doi.org/10.1016/j.ajhg.2008.11.007

Airey J (2012) The politics of rape: sexual atrocity, propaganda wars, and the restoration stage. University of Delaware Press, Newark, NJ

Allentoft ME, Sikora M, Sjögren K-G et al (2015) Population genomics of bronze age Eurasia. Nature 522:167–172. https://doi.org/10.1038/nature14507

Alvarez-Cubero MJ, Santiago O, Martínez-Labarga C et al (2018) Methodology for Y chromosome capture: a complete genome sequence of Y chromosome using flow cytometry, laser microdissection and magnetic streptavidin-beads. Sci Rep 8:9436. https://doi.org/10.1038/s41598-018-27819-x

Arauna LR, Mendoza-Revilla J, Mas-Sandoval A et al (2017) Recent historical migrations have shaped the gene pool of Arabs and Berbers in North Africa. Mol Biol Evol 34:318–329. https://doi.org/10.1093/molbev/msw218

Arredi B, Poloni ES, Paracchini S et al (2004) A predominantly neolithic origin for Y-chromosomal DNA variation in North Africa. Am J Hum Genet 75:338–345

Auton A, Abecasis GR, Altshuler DM et al (2015) A global reference for human genetic variation. Nature 526:68–74. https://doi.org/10.1038/nature15393

Barbujani G, Magagni A, Minch E, Cavalli-Sforza LL (1997) An apportionment of human DNA diversity. Proc Natl Acad Sci U S A 94:4516–4519

Batini C, Hallast P, Zadik D et al (2015) Large-scale recent expansion of European patrilineages shown by population resequencing. Nat Commun 6:7152. https://doi.org/10.1038/ncomms8152

Bosch E, Calafell F, Comas D et al (2001) High-resolution analysis of human Y-chromosome variation shows a sharp discontinuity and limited gene flow between northwestern Africa and the Iberian Peninsula. Am J Hum Genet 68:1019–1029. https://doi.org/10.1086/319521

Bosch E, Calafell F, Pérez-Lezaun A et al (2000) Y chromosome STR haplotypes in four populations from northwest Africa. Int J Legal Med 114:36–40. https://doi.org/10.1007/s004140000136

Bosch E, Calafell F, Pérez-Lezaun A et al (1997) Population history of North Africa: evidence from classical genetic markers. Hum Biol 69:295–311

Bouckaert R, Heled J, Kühnert D et al (2014) BEAST 2: a software platform for bayesian evolutionary analysis. PLoS Comput Biol 10:e1003537. https://doi.org/10.1371/journal.pcbi.1003537

Burgarella C, Navascués M (2011) Mutation rate estimates for 110 Y-chromosome STRs combining population and father–son pair data. Eur J Hum Genet 19:70–75. https://doi.org/10.1038/ejhg.2010.154

Busby GBJ, Brisighelli F, Sanchez-Diz P et al (2012) The peopling of Europe and the cautionary tale of Y chromosome lineage R-M269. Proc R Soc B Biol Sci 279:884–892. https://doi.org/10.1098/rspb.2011.1044

Calafell F, Larmuseau MHD (2017) The Y chromosome as the most popular marker in genetic genealogy benefits interdisciplinary research. Hum Genet 136:559–573. https://doi.org/10.1007/s00439-016-1740-0

Cruciani F, Trombetta B, Antonelli C et al (2011) Strong intra- and inter-continental differentiation revealed by Y chromosome SNPs M269, U106 and U152. Forensic Sci Int Genet 5:e49–e52. https://doi.org/10.1016/j.fsigen.2010.07.006

Cruz-Dávalos DI, Nieves-Colón MA, Sockell A et al (2018) In-solution Y-chromosome capture-enrichment on ancient DNA libraries. BMC Genom 19:608. https://doi.org/10.1186/s12864-018-4945-x

de Filippo C, Bostoen K, Stoneking M, Pakendorf B (2012) Bringing together linguistic and genetic evidence to test the Bantu expansion. Proc R Soc B Biol Sci 279:3256–3263. https://doi.org/10.1098/rspb.2012.0318

Dupanloup I, Pereira L, Bertorelle G et al (2003) A recent shift from polygyny to monogamy in humans is suggested by the analysis of worldwide Y-chromosome diversity. J Mol Evol 57:85–97. https://doi.org/10.1007/s00239-003-2458-x

Fadhlaoui-Zid K, Haber M, Martínez-Cruz B et al (2013) Genome-wide and paternal diversity reveal a recent origin of human populations in North Africa. PLoS ONE 8:e80293. https://doi.org/10.1371/journal.pone.0080293

Fregel R, Gomes V, Gusmão L et al (2009) Demographic history of Canary Islands male gene-pool: replacement of native lineages by European. 14:1–14. https://doi.org/10.1186/1471-2148-9-181

Fregel R, Méndez FL, Bokbot Y et al (2017) Neolithization of North Africa involved the migration of people from both the Levant and Europe. 1–14

Fu Q, Li H, Moorjani P et al (2014) Genome sequence of a 45,000-year-old modern human from western Siberia. Nature 514:445–449. https://doi.org/10.1038/nature13810

Fu Q, Posth C, Hajdinjak M et al (2016) The genetic history of Ice Age Europe. Nature 534:200–205. https://doi.org/10.1038/nature17993

Gusmao L, Sanchez-Diz P, Calafell F et al (2005) Mutation rates at Y chromosome specific microsatellites. Hum Mutat 26:520–528. https://doi.org/10.1002/humu.20254

Haak W, Lazaridis I, Patterson N et al (2015) Massive migration from the steppe was a source for Indo-European languages in Europe. Nature 522:207–211. https://doi.org/10.1038/nature14317

Henn BM, Botigue LR, Gravel S et al (2012) Genomic ancestry of North Africans supports back-to-Africa migrations. PLoS Genet 8:e1002397

Jobling MA, Hollox EJ, Hurles ME et al (2014) Human evolutionary genetics, 2nd edn. Garland Science, New York and London

Karafet TM, Mendez FL, Meilerman MB et al (2008) New binary polymorphisms reshape and increase resolution of the human Y chromosomal haplogroup tree. Genome Res 18:830–838

Kayser M, Caglià A, Corach D et al (1997) Evaluation of Y-chromosomal STRs: a multicenter study. 110:125–133

Kohler TA, Smith ME, Bogaard A et al (2017) Greater post-Neolithic wealth disparities in Eurasia than in North America and Mesoamerica. Nature 551:619–622. https://doi.org/10.1038/nature24646

Kuderna LFK, Lizano E, Julià E et al (2019) Selective single molecule sequencing and assembly of a human Y chromosome of African origin. Nat Commun 10:4. https://doi.org/10.1038/s41467-018-07885-5

Larmuseau MHD, Calafell F, Princen SA et al (2018) The black legend on the Spanish presence in the low countries: verifying shared beliefs on genetic ancestry. Am J Phys Anthropol. https://doi.org/10.1002/ajpa.23409

Larmuseau MHD, Vanderheyden N, Van Geystelen A et al (2014) Recent radiation within Y-chromosomal haplogroup R-M269 resulted in high Y-STR haplotype resemblance. Ann Hum Genet 78:92–103. https://doi.org/10.1111/ahg.12050

McEvedy C, Jones R (1978) Atlas of world population history. Penguin, Harmondsworth (UK)

Mendizabal I, Sandoval K, Berniell-Lee G et al (2008) Genetic origin, admixture, and asymmetry in maternal and paternal human lineages in Cuba. BMC Evol Biol 8:213. https://doi.org/10.1186/1471-2148-8-213

Myres NM, Rootsi S, Lin AA et al (2011) A major Y-chromosome haplogroup R1b Holocene era founder effect in Central and Western Europe. Eur J Hum Genet 19:95–101. https://doi.org/10.1038/ejhg.2010.146

Parker G (1990a) The Dutch revolt: the revised version. Penguin Books

Parker G (1990b) Spain and the Netherlands, 1559–1659. Fontana Press, Waukegan, Il

Parker G (1973) Mutiny and discontent in the Spanish army of flanders. Past Present 58:38–52

Parker G, González de León F (1999) The grand strategy of Philip II and the revolt of the Netherlands. In: Van Nierop H, Venard M, Benedict P (eds) Reformation, revolt, and civil war in France and the Netherlands 1555–1585. Koninklijke Nederlandse Akademie van Wetenschappen, Amsterdam, pp 215–232

Poznik GD (2016) Identifying Y-chromosome haplogroups in arbitrarily large samples of sequenced or genotyped men. bioRxiv 088716. https://doi.org/10.1101/088716

Poznik GD, Xue Y, Mendez FL et al (2016) Punctuated bursts in human male demography inferred from 1,244 worldwide Y-chromosome sequences. Nat Genet 48:593–599. https://doi.org/10.1038/ng.3559

Purps J, Siegert S, Willuweit S et al (2014) A global analysis of Y-chromosomal haplotype diversity for 23 STR loci. Forensic Sci Int Genet 12:12–23. https://doi.org/10.1016/j.fsigen.2014.04.008

Raghavan M, DeGiorgio M, Albrechtsen A et al (2014) The genetic prehistory of the New World Arctic. Science 345:1255832. https://doi.org/10.1126/science.1255832

Ralf A, Montiel González D, Zhong K, Kayser M (2018) Yleaf: software for human Y-chromosomal haplogroup inference from next generation sequencing data. Mol Biol Evol. https://doi.org/10.1093/molbev/msy032

Redd AJ, Agellon AB, Kearney VA et al (2002) Forensic value of 14 novel STRs on the human Y chromosome. 130:97–111

Rocca RA, Magoon G, Reynolds DF et al (2012) Discovery of Western European R1b1a2 Y chromosome variants in 1000 genomes project data: an online community approach. PLoS ONE 7:e41634

Rohde DLT, Olson S, Chang JT (2004) Modelling the recent common ancestry of all living humans. Nature 431:562–566. https://doi.org/10.1038/nature02842

Romualdi C, Balding D, Nasidze IS et al (2002) Patterns of human diversity, within and among continents, inferred from biallelic DNA polymorphisms. Genome Res 12:602–612

Seielstad MT, Minch E, Cavalli-Sforza LL (1998) Genetic evidence for a higher female migration rate in humans. 20:278–280

Soen V (2008) ¿Más allá de la leyenda negra? Léon van der Essen y la historiografía reciente en torno al castigo de las ciudades rebeldes en los Países Bajos (siglos XIV a XVI). In: Janssens G, Van der Essen L (eds) El Ejército Español en Flandes 1567–1584. Academia de Yuste, Yuste, pp 45–72

Solé-Morata N, Bertranpetit J, Comas D, Calafell F (2014a) Recent radiation of R-M269 and high Y-STR haplotype resemblance confirmed. Ann Hum Genet 78. https://doi.org/10.1111/ahg.12066

Solé-Morata N, Bertranpetit J, Comas D, Calafell F (2015) Y-chromosome diversity in Catalan surname samples: insights into surname origin and frequency. Eur J Hum Genet 23:1549–1557. https://doi.org/10.1038/ejhg.2015.14

Solé-Morata N, Bertranpetit J, Comas D, Calafell F (2014b) Recent radiation of R-M269 and high Y-STR haplotype resemblance confirmed. Ann Hum Genet 78:253–254. https://doi.org/10.1111/ahg.12066

Solé-Morata N, García-Fernández C, Urasin V et al (2017a) Whole Y-chromosome sequences reveal an extremely recent origin of the most common North African paternal lineage E-M183 (M81). Sci Rep 7:15941. https://doi.org/10.1038/s41598-017-16271-y

Solé-Morata N, Villaescusa P, García-Fernández C et al (2017b) Analysis of the R1b-DF27 haplogroup shows that a large fraction of Iberian Y-chromosome lineages originated recently in situ. Sci Rep 7. https://doi.org/10.1038/s41598-017-07710-x

Stamatakis A (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30:1312–1313. https://doi.org/10.1093/bioinformatics/btu033

Swart K (1975) The black legend during the eighty years war. In: Bromley J, Kossmann E (eds) Britain and the Netherlands. Springer, The Hague, pp 36–57

Valverde L, Illescas MJ, Villaescusa P et al (2016) New clues to the evolutionary history of the main European paternal lineage M269: dissection of the Y-SNP S116 in Atlantic Europe and Iberia. Eur J Hum Genet 24:437–441. https://doi.org/10.1038/ejhg.2015.114

van Oven M, Van Geystelen A, Kayser M et al (2014) Seeing the wood for the trees: a minimal reference phylogeny for the human Y chromosome. Hum Mutat 35:187–191. https://doi.org/10.1002/humu.22468

Verdu P, Austerlitz F, Estoup A et al (2009) Origins and genetic diversity of pygmy hunter-gatherers from Western Central Africa. Curr Biol 19:312–318

Wei W, Ayub Q, Chen Y et al (2013) A calibrated human Y-chromosomal phylogeny based on resequencing. Genome Res 23:388–395

Wooding S, Ostler C, Prasad BV et al (2004) Directional migration in the Hindu castes: inferences from mitochondrial, autosomal and Y-chromosomal data. Hum Genet 115:221–229

Chapter 12
Vertebrate Genome Size and the Impact of Transposable Elements in Genome Evolution

Maria A. Biscotti, Federica Carducci, Ettore Olmo and Adriana Canapa

Abstract In eukaryotes, the haploid DNA content (C-value) varies widely across lineages without an apparent correlation with the complexity of organisms. This incongruity has been called the C-value paradox and has been solved by demonstrating that not all DNA is constituted by genes but, on the contrary, most of it is made up of repetitive DNA. In vertebrates, the increasing number of sequenced genomes has shown that differences in genome size between lineages are ascribable to a variation in transposon content. These mobile elements, previously perceived as "junk DNA" or "selfish DNA," are now recognized as the major players in shaping genomes. During vertebrate evolution, transposable elements have been repeatedly co-opted and exapted to generate regulatory sequences, coding exons, or entirely new genes that lead to evolutionary advantages for the host. Moreover, transposable elements are also responsible for substantial rearrangements such as insertions, deletions, inversions, and duplications potentially associated with, or following, speciation events.

List of Abbreviations and Acronyms

7SL RNA	eukaryotic small cytoplasmic RNA
Alu	*Arthrobacter luteus* restriction endonuclease
CR1	Chicken Repeat 1
en	endonuclease
env	envelope
ISL-1	Insulin gene enhancer protein ISL-1
L1	LINE1
L2	LINE2
L3	LINE3
LINE	Long Interspersed Nuclear Elements
LTR	Long Terminal Repeat

M. A. Biscotti (✉) · F. Carducci · E. Olmo · A. Canapa
Dipartimento di Scienze della Vita e dell'Ambiente, Università Politecnica delle Marche, Via Brecce Bianche, 60131 Ancona, Italy
e-mail: m.a.biscotti@univpm.it

P. Pontarotti (ed.), *Evolution, Origin of Life, Concepts and Methods*,
https://doi.org/10.1007/978-3-030-30363-1_12

MITEs	Miniature Inverted-repeated Transposable Elements
pg/N	picograms/Nucleus
POMC	ProopioMelanoCortin gene
RAG1	Recombination-ActivatinG Protein 1
RAG2	Recombination-ActivatinG Protein 2
RT	Reverse Transcriptase
SatDNA	Satellite DNA
SINE	Short Interspersed Nuclear Elements
SINE-R	Short Interspersed Nuclear Elements-R, where R indicates a sequence of Retroviral origin
SNPs	Single Nucleotide Polymorphisms
SETMAR	SET domain and Mariner transposase fusion gene
SVA	SINE-VNTR-*Alu*
TEs	Transposable Elements
TIR	Terminal Inverted Repeat
UTRs	UnTranslated Regions
VDJ	Variable Diversity Joining
VNTR	Variable Number of Tandem Repeats

12.1 Genome Size Variation: A Fascinating Enigma

In 1948, Vendrely and Vendrely (1948) reported a "remarkable constancy in the nuclear DNA content of all the cells in all the individuals within a given animal species." This constancy, referred to the haploid nuclear DNA content, was defined as the *C-value* (Swift 1950).

Comparing the haploid DNA content or *C-value* in eukaryotes, there is a notable lack of correlation between DNA content and organism complexity (Thomas 1971). Indeed, DNA is the stuff of genes and the more complex the organism is, the more genes it should have, and thus more DNA. However, it has been demonstrated that simple organisms such as some amoebas have hundreds of times more DNA than humans (Thomas 1971). The lack of correlation between the size and complexity of eukaryotic genomes is known as the *C-value paradox* and remained a mystery for almost half a century. Research in this field revealed that this incongruity is only apparent since not all DNA is made up of genes but, on the contrary, most of it is constituted by non-coding DNA and often repetitive DNA. The comparison between the genomes of prokaryotes and multicellular eukaryotes has shown an increase in genome size that is associated not only with a proliferation of repeated elements but also with an increased number of genes. Moreover, an expansion in the size and number of introns has also been related to the gigantism of genomes (Lynch and Conery 2003).

The presence of repetitive DNA in the genome of eukaryotes opens up a number of questions regarding, first of all, why some species possess a great amount of

repetitive DNA and others present compact genomes. It would be interesting to reveal the mechanisms by which repetitive DNA spreads or is deleted from genomes during evolution, together with the effects and functions that it might have on chromosomes, nuclei, cells, and organisms. It is also intriguing to investigate whether this DNA has an adaptive role, and if this is not the case, it would be interesting to understand why natural selection has tolerated so much extra DNA.

According to the adaptive theory, this DNA, given its amount, influences phenotype. Indeed, the amount of DNA can directly or indirectly increase nuclear and cell size, the duration of mitosis and meiosis, the rate of basal metabolism, as well as embryonic development time and body size (Bennett 1971; Cavalier-Smith 1978; Vinogradov 1995, 1997). It has been pointed out that for birds, it is advantageous to have a smaller genome size because this implies smaller cells and thus a higher surface-to-volume ratio which provides a consequently more efficient transfer of oxygen through the cell membrane (Olmo 1983; Hughes and Hughes 1995). This hypothesis is also supported by the reduced genome size of bats compared with other mammals (Burton et al. 1989; Van den Bussche et al. 1995).

In 2002, Petrov (2002) suggested the hypothesis of the "mutation equilibrium model" according to which genome size is determined as the equilibrium between the rate of sequence loss by the deletion bias of small indels and the rate of sequence gain by long insertions.

The extra DNA could also have a protective function since mutations can statistically occur more frequently in the non-coding repetitive fraction (Vinogradov 1998).

Repetitive DNA also includes mobile elements that, given their ability to replicate themselves, have been defined as "selfish DNA" (Doolittle and Sapienza 1980). The proliferation of such elements in the genome depends on the strength of natural selection so that the final genome size is the highest tolerable value. However, smaller genomes seem to be favored by natural selection also in relation to the population size: species that have experienced a population reduction present larger genomes and are more prone to extinction (Vinogradov 2003, 2004; Kraaijeveld 2010).

Alternatively, DNA accumulation might be non-adaptive and thus useless. For this reason, repetitive DNA was initially labeled as "junk DNA," fixed by random drift and carried passively in the genome (Ohno 1972). On the contrary, an increasing number of studies are now supporting an unexpected dynamicity of repetitive DNA which was originally thought to be silent and inert (Biscotti et al. 2015a, b; Biemont 2010).

Several approaches have been adopted to estimate the nuclear DNA content. In the 1970s and 1980s, it became relatively common to use reassociation kinetics to assess the composition and size of genomes. This method consists in the extraction of DNA from cells and denaturation by heating. The solution of denaturated genomic DNA is placed in an environment conducive to renaturation. The rates of reassociation of the DNA strands are proportional to the number of times that specific sequences are found in the genome, providing information on the repetitive and low-copy components. Moreover, they can be calibrated against a standard to give an estimate of absolute DNA content. DNA reassociation of a eukaryotic genome is

described by a Cot curve where Co is the starting concentration of nucleotides and t is the reassociation time; low values correspond to highly repetitive DNA while high values indicate single and low-copy DNA sequences (Britten et al. 1974). Thus, Cot analysis provides considerable information on the size and structure of eukaryotic genomes. However, reassociation kinetics for large size genomes are very slow and not particularly accurate and consequently, alternative methods have been adopted. Flow cytometry is a technique used in several applications including genome sizing. It consists in isolating the nuclei and staining them with a fluorescent dye that binds stoichiometrically to DNA. The amount of fluorescent light emitted by each nucleus is converted into a digital signal and compared with a known amount of DNA in order to determine the absolute DNA content in the species of interest. Using this method, the sources of error are the choice of fluorochrome and the presence of inhibitors or cytoplasmatic constituents released during nuclei isolation. Feulgen microdensitometry is a method based on staining nuclei with the Feulgen technique and then measuring the amount of light absorbed by the stain. However, this method can be influenced by chromatin condensation since histones, in particular, restrict the accessibility of DNA to fluorochromes and thus, the stoichiometry of DNA staining is affected. This depends on cell types but also on the fluorochromes.

The use of next-generation genome sequencing techniques has provided information on genome size, organization, and composition in an increased number of species. However, GC-rich regions or arrays of repetitive DNA are under-represented in assemblies (Peona et al. 2018). Until sequencing technologies will not allow obtaining scaffolds which span the entire length of individual chromosomes, genome assemblies will continue to be far from complete. Moreover, organisms such as lamprey are characterized by the physical restructuring of the genome during development that consists in the elimination of about 0.5 Gb of DNA from the 2.3 Gb genome. As a consequence, somatic cell types possess a smaller gene complement compared to germ cells possessing a full complement (Smith et al. 2018).

An understanding of the C-value paradox will only be achieved through studies on the non-coding portion of the genome; the so-called dark matter, which currently, given the technical difficulties in identifying and understanding its function, is a subject of interest for many research groups (Blaxter 2010; Kapranov and Laurent 2012).

12.1.1 Types of Repetitive DNA Sequences

Repetitive DNA includes sequences present in multiple copies in the genome and can account for up to 90% of the genome size in some species (Biscotti et al. 2015a; Lopez-Flores and Garrido-Ramos 2012).

This DNA, together with sequences removed from mRNA before translation (5′ and 3′ untranslated regions and introns) and pseudogenes, is part of non-coding DNA (Fig. 12.1).

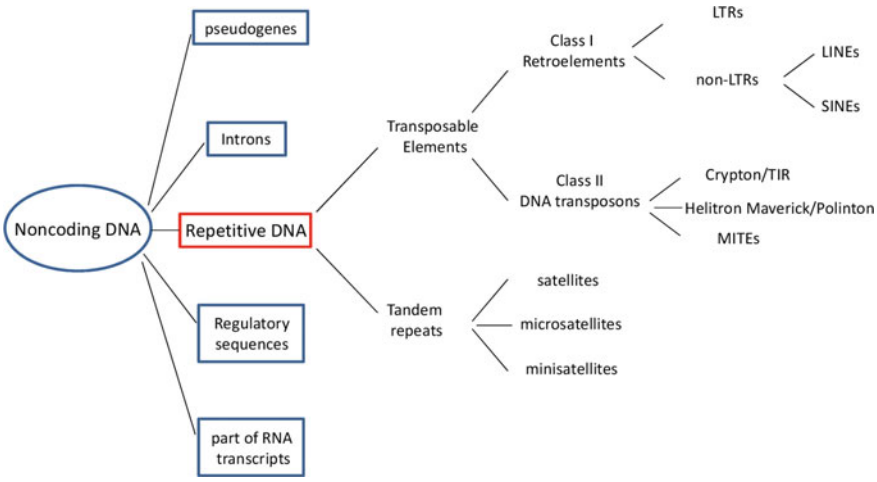


Fig. 12.1 Scheme reporting the main sequence types included in non-coding DNA with in-deep focus onto repetitive DNA in eukaryotes

Repetitive DNA is constituted by sequences repeated thousands of times that can be grouped into two main types: transposable elements (TEs) and tandem repeats (Fig. 12.1).

Among repetitive sequences, TEs are mostly responsible for the pronounced differences in genome size (Garrido-Ramos 2017). They are genetic elements characterized by their ability to insert themselves in novel genome locations of the host and to increase in number by replication. On the basis of their transposition mechanism, TEs can be further distinguished into (i) Class I retroelements or retrotransposons and (ii) Class II DNA transposons (Goerner-Potvin and Bourque 2018; Bourque et al. 2018).

Retroelements (Class I) are provided by an RNA intermediate that is then reverse transcribed into complementary DNA using a copy and paste mechanism. In Class I, long terminal repeat (LTR) retrotransposons and non-LTR (non-LTR) retrotransposons can both be found.

LTR retrotransposons are characterized by direct LTR-flanking sequences of about 250–600 bp, necessary for the transcription and consequent insertion into the host genome. LTR retrotransposons are structurally very similar to retroviruses with the exception of the envelope gene that is only present in retroviruses (Naville et al. 2016). Moreover, these elements, unlike retroviruses, are not able to move between cells and to infect them (Malik et al. 2000; Ribet et al. 2008). Besides direct LTR-flanking sequences, LTR retrotransposons are constituted by some genes, essential for the complete synthesis of all the components of reverse transcriptase machinery: *gag* protein, reverse transcriptase (RT), protease (prt), RNAse H, and integrase (int).

After the RT-mediated cDNA synthesis, integrase inserts the cDNA into a new position of the genome. The subclassification of LTR retrotransposons includes three main TE superfamilies in vertebrates: *Ty1/Copia* (Pseudoviridae), *Ty3-gypsy-like* (Metaviridae), and *BEL/Pao* (Chalopin et al. 2015).

Non-LTR retroelements are defined as autonomous retrotransposons and are mainly represented by long interspersed nuclear elements (LINEs) and short interspersed nuclear elements (SINEs). LINEs are non-LTR retrotransposons that may contain one or two open reading frames (ORFs) (Wicker et al. 2007). One of the two ORFs encodes for a reverse transcriptase (RT) and an endonuclease (en) domain encoding for a protein responsible for integration of the TE copy in a different region of the genome. On the contrary, SINEs are non-autonomous retrotransposons and do not code for a RT, thus they need LINE reverse transcriptase in order to be transposed (Kramerov and Vassetzsky 2011). The origin of SINEs can be ascribed to the reverse transcription of Pol III transcripts (Lopez-Flores and Garrido-Ramos 2012).

Class II DNA transposons are transposed by moving their genomic DNA copies from one chromosomal location to another without any RNA intermediates (Goerner-Potvin and Bourque 2018; Bourque et al. 2018) and can be divided into subclasses I and II. In subclass I, Crypton elements and terminal inverted repeat (TIR) transposons can be found. For these elements, both DNA strands are cleaved and transposed following the canonical *cut and paste* mechanism of transposition. In this case, the number of these elements remains unchanged. In subclass II, the major representatives are Helitrons and Maverick/Polinton elements in which the transposition follows the *copy and paste* mechanism (Wicker et al. 2007; Kapitonov and Jurka 2008). Class II also comprises MITEs, non-autonomous transposons originated from DNA transposons, not encoding for a transposase and therefore unable to copy themselves autonomously. Thus, they exploit transposase encoded by autonomous elements in order to transpose (Feschotte et al. 2003).

Beside TEs, repetitive DNA includes tandem repeats that are constituted by satellite DNAs (satDNAs), minisatellites, and microsatellites (Fig. 12.1). The former are highly repetitive DNA sequences contributing considerably to the C-value (Biscotti et al. 2015a) while the latter show a lower number of repeats and are classified as moderately repeated sequences. However, there are some exceptions, as found in snakes which are characterized by a great number of minisatellites and microsatellites (Adams et al. 2016; Pasquesi et al. 2018). SatDNAs are organized as long arrays of head-to-tail linked repeats and are mainly localized at telomeric, centromeric, or pericentromeric level. Their preferential localization may suggest a possible involvement in biological or functional processes (Biscotti et al. 2015b) such as centromeric DNA packaging (Heslop-Harrison and Schwarzacher 2011; Levitsky et al. 2013), chromosome segregation, and kinetochore formation (Schueler et al. 2001). These sequences may be characterized by a high dynamism so as to be species or genus-specific (Garrido-Ramos 2015) but may also be conserved in some lineages for long evolutionary periods (Chaves et al. 2017; Mravinac et al. 2005; Petraccioli et al. 2015; Biscotti et al. 2018).

12.2 Genome Size in Vertebrates: An Overview

The genome size database (http://www.genomesize.com/) contains information about the DNA content of 6222 species of which 3793 are vertebrates (Gregory 2019). Currently, the genome size of the smallest animal is estimated as 0.02 pg/N and belongs to the plant-parasitic nematode *Pratylenchus coffeae* while the largest animal genome is found in the marbled lungfish *Protopterus aethiopicus* and is estimated as 132.83 pg/N. Changes in genome size are mainly ascribable to the gain or loss of repetitive DNA. The advent of high throughput sequencing technologies has led to an increase in the number of sequenced genomes and this has been extremely important for extending our knowledge on the amount and the landscape of TEs in several species.

Although data available on primitive metazoans are limited, they seem to suggest a low genome size and probably a low percentage of transposons at the origin of metazoans. In protostomes, the variability in genome size is mainly related to the expansion of various classes of transposons (Canapa et al. 2015). Indeed, a positive correlation exists between genome size and the percentage of transposons. Most of the data on protostomes derives from species belonging to the Arthropoda phylum. Within insects, genome size is mainly comprised between 0.09 and 4 pg/N, with the exception of orthopterans that reach 16.93 pg/N. These differences are attributable to the percentage of transposons showing variable rates of amplification in the different groups of insects (Canapa et al. 2015). Although in invertebrates, genome size rarely exceeds 20 pg/N, in crustaceans, some species have genomes of over 50 pg/N. It has been noted that such values are restricted to species adapted to extreme environments, such as polar regions or hydrothermal vents (Bonnivard et al. 2009; Dufresne and Jeffery 2011).

Among deuterostomes, primitive chordates (urochordates and cephalochordates) present smaller genomes than vertebrates. This seems to be related to the whole-genome duplication events (WGDs) that have affected vertebrates during evolution. These events determined the appearance of new features leading to an increase in complexity. However, the variation in genome size does not follow a common trend among the different vertebrate lineages. Indeed, comparing the estimated genome size among vertebrates, some classes experienced expansions while others experienced strong contractions (Fig. 12.2). For example, the genome of the teleost fish *Takifugu rubripes* is one-eighth of that of *Homo sapiens* even if both contain a comparable number of protein-coding genes (Aparicio et al. 2002); amphibians and lungfish show the widest range of genome size variation (Fig. 12.2) and both taxa have organisms with the largest genomes among vertebrates.

Regarding agnathes, the genome size data are available for seven species of Myxiniformes and for 10 species of Petromyzontiformes (Gregory 2019). These organisms have moderate genomes ranging from 1.29 to 4.59 pg/N, thus ranking after birds among vertebrates (Fig. 12.2). Data on transposon contribution are available only for the sea lamprey *Petromyzon marinus* for which the analysis of the sequenced genome revealed that 34.7% is made up of mobile elements (Smith et al. 2013). More

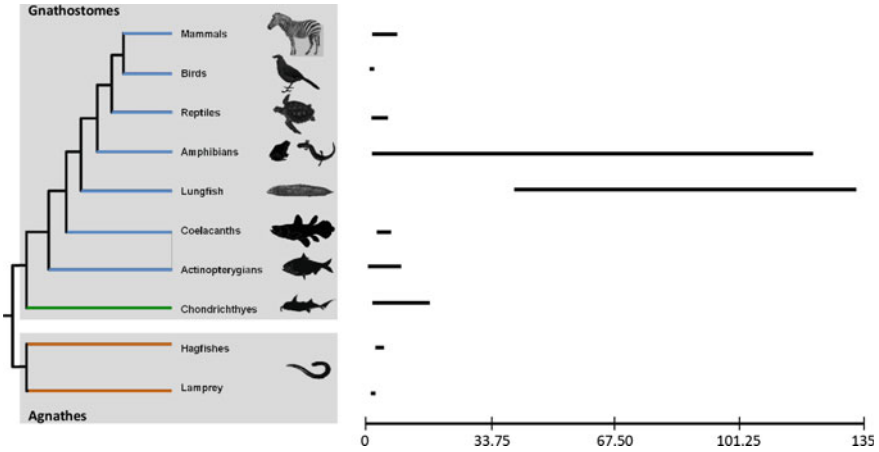


Fig. 12.2 DNA content in vertebrate lineages. On the left: the evolutionary relationships between vertebrate lineages. On the right: bars indicating the range of DNA amount for each lineage. Orange branches represent jawless fishes; green branch represents Chondrichthyes; blue branches represent Osteichthyes. The scale indicates the amount of DNA in pg/nucleus. Note that for coelacanths, only one species has been analyzed but with different methods. This justifies the presence of a range of C-values in the figure

than 20% of these mobile elements are unknown while the remaining portion (about 15%) is constituted by LINEs, LTRs, and DNA transposons (Chalopin et al. 2015).

Considerable genomic dimensions are reported for some species of cartilaginous fish, reaching 17 pg/N. The analysis of the sequenced genome of the elephant shark *Callorhinchus milii* demonstrated that more than 40% is composed of transposons with a major contribution of LINE retroelements (Chalopin et al. 2015; Venkatesh et al. 2014).

Actinopterygians comprise of Polypteriformes, Acipenseriformes, holostei, and teleosts. The latter is one of the most successful groups, representing more than 99.8% of ray-finned fishes. The genome size of bony fish is comprised between 0.34 and 9.32 pg/N, including organisms with the most compact genome among vertebrates, such as some species belonging to the Tetraodontidae family (Gregory 2019). In this lineage, the link between genome size variation and transposon contribution is particularly evident. Indeed, within actinopterygians, the content of mobile elements is extremely variable, ranging between a very low amount in puffer fish (about 6%) to more than 50% in zebrafish. Moreover, ray-finned fish genomes present a higher mobile element diversity than other vertebrate lineages which is also maintained in the most compact genomes. Overall, in teleosts, the major contribution is provided by DNA transposons (Chalopin and Volff 2017) while compact genomes do not show a prevalence of any transposon type. In non-teleost species, such as the spotted gar *Lepisosteus oculatus*, a predominance of non-LTR retrotransposons is evident (Chalopin et al. 2015; Chalopin and Volff 2017).

Lobe-finned fish includes two species of coelacanths (*Latimeria chalumnae* and *Latimeria menadoensis*) and six species of lungfish (*Protopterus annectens*, *Protopterus dolloi*, *Protopterus aethiopicus*, *Protopterus amphibius*, *Lepidosiren paradoxa*, and *Neoceratodus forsteri*). While the former presents a moderate amount of DNA of about 3.5 pg/N, the latter have a genome size ranging from 40 to 132.83 pg/N. The contribution of transposons has been well evaluated in *L. chalumnae* for which genome sequencing is available (Amemiya et al. 2013). Analyses performed on the genome of this taxon revealed that 20% is made up of transposons with about one-third of SINEs (Chalopin et al. 2015). In lungfish, the huge genome size represents a drawback for current sequencing techniques and assembly procedures. However, a study performed on a small portion of the *N. forsteri* genome estimated that 40% of the genome is made up of transposons and suggested that CR1 and L2 (non-LTR) are predominant (Metcalfe et al. 2012). These data obtained in basal sarcopterygians at genome level reflect the results obtained by analyzing the activity of mobile elements in the transcriptomes of the Indonesian coelacanth *L. menadoensis* (Forconi et al. 2014) and the West African lungfish *P. annectens* (Biscotti et al. 2016).

Among vertebrates, the Amphibia class shows the widest range of genome size from 0.95 to 120.60 pg/N (Fig. 12.2). Most of the analyzed species belong to the Anura and Urodela orders, while only three species have been investigated for the Gymnophiona order (Gregory 2019). In Anura and Gymnophiona, the genome size does not exceed 14 pg/N unlike Urodela in which the values range from 10.12 to 120.60 pg/N in species belonging to the Proteidae family (Gregory 2019). Mobile elements constitute from 20% to over 40% of the genome with a predominance of DNA transposons in *Xenopus tropicalis* (Chalopin et al. 2015; Sun et al. 2012, 2015) and of LTR in urodeles (Canapa et al. 2015; Nowoshilow et al. 2018) and in the Tibetan frog *Nanorana parkeri* (Sun et al. 2015). However, in amphibians, the genomic gigantism observed is not only due to the higher amount of repetitive DNA but also to longer introns, as found mainly in salamanders (Sun et al. 2012; Voss et al. 2013; Nowoshilow et al. 2018).

The genome size of the 420 non-bird reptile species analyzed to date ranges from 1.05 to 5.44 pg/N. In Squamata and Crocodylia, values are comprised between 1.05 and 3.95 pg/N while in Testudines and in the unique analyzed species of Sphenodontia, the genome size exceeds 4.00 pg/N reaching the value of 5.44 pg/N in *Testudo graeca*. In Squamata and Crocodylia, about 30% of the genome is TE-derived and the major contribution is ascribable to non-LTR and DNA transposons (Alföldi et al. 2011; Green et al. 2014; Castoe et al. 2011, 2013). The genomes of turtles sequenced to date indicate that around 10% is represented by TEs and that non-LTR retrotransposons constitute the predominant part of the mobilome, as is the case for the other two orders (Shaffer et al. 2013; Wang et al. 2013).

Birds present the smallest genomes among vertebrates (Fig. 12.2). Indeed, the 898 species included in the genome size database show values comprised between 0.91 and 2.16 pg/N (Gregory 2019). The compressed genomes of birds are characterized by a very low number of mobile elements with the loss of certain TE families; however, their mobilome is mainly constituted by retroelements (Chalopin et al. 2015). Studying fossil cell size in dinosaurs, the contraction of the genome size can be dated

at 230–250 Mya in saurischians, the lineage from which birds originated. Comparative genomic analyses on the current descendants of this evolutionary lineage showed that the reduction observed in saurischians might be due to a strong reduction in non-LTR elements (Organ et al. 2007). It has been proposed that the genome size between 2.5 and 5.0 pg/N represents the ancestral condition of the sarcopterygian lineage and consequently, the large genomes of lungfish and salamanders, together with the contracted genomes of birds, are secondarily derived (Organ et al. 2011). Moreover, it is worth noting that the increase in the amount of DNA experienced in dipnoans and amphibians accompanied their transition from water to land life which, being one of the most important steps during vertebrate evolution, probably required significant changes at genome level.

Finally, the genome size of mammals ranges from 1.63 to 8.40 pg/N (Gregory 2019) and the TE content accounts for more than 30% of the mammalian genomes sequenced to date. Non-LTR elements are the most prominent type even if in therians (Metatheria and Eutheria) there is a prevalence of L1 LINE retroelements while in monotremes there is a predominance of L2 LINE retroelements (Chalopin et al. 2015). Moreover, in mammals, most species having small genomes (less than 2.5 pg/N of DNA) belong to the Chiroptera order. These organisms are the only group of mammals to have evolved powered flight and, interestingly, they are characterized by reduced genomes, similar to birds. However, while in the latter, there is a prevalence of retroelements, and in bats, there is an accumulation of DNA transposons (Pagán et al. 2012; Ray et al. 2007).

In conclusion, the general trend that can be extrapolated is that retroelements have shaped the genomes of jawless fish, cartilaginous fish, coelacanths, lungfish, birds, and mammals while DNA transposons have played a key role in ray-finned fish and *Xenopus* genome size.

Moreover, if some lineages are characterized by high transposon diversity others have experienced a reduction in retroelement diversity with the complete extinction of some families in certain lineages. Indeed, in mammals, only three families of retrotransposons are present: the non-LTR retrotransposons L1 (LINE1), L2, and L3/CR1 while in birds, L1 and L2 have been completely lost (Wicker et al. 2005). The extinction of ancient families of TEs has also been identified in teleosts, in which the non-LTR retrotransposon Rex3 is widespread, but not in salmonids (Volff et al. 2001; Carducci et al. 2018).

It is also interesting to note that the impact of the same TE family can be very different: the L1 family is highly dispersed in mammalian genomes while a much lower copy number is present in fish genomes (Volff et al. 2003, Furano et al. 2004); the L3/CR1 family is the major group of TEs in birds with 96,000 copies compared to the larger genome of placental mammals with only 8000 copies (Wicker et al. 2005).

Finally, some TEs were not present in the common ancestor of vertebrates but were introduced/originated *ex novo* in some lineages. This is the case of *Alu* elements derived from 7SL RNA, or SVA elements originated from *Alu* and SINE-R. Both these elements are non-autonomous retroelements specific of primates.

Overall, the TE content in a given species could be the result of an equilibrium between TE transposition, defense mechanisms of the genome, and natural selection constraints allowing genome functionality to be maintained.

The determination of the TE landscape in a genome depends on the methods used to identify and annotate TEs. This issue requires considerable efforts due to the great variability of TEs and to the accumulation of mutations in old and inactive TE sequences. Currently, there is no reliable strategy to overcome this problem. However, three approaches are commonly used: library-based methods, signature-based methods, and de novo consensus methods (Goerner-Potvin and Bourque 2018; Lerat 2010; Girgis 2015; Tempel 2012). The library-based methods use Repeat-Masker program, usually in association with Repbase. The signature-based methods identify specific traits such as long terminal repeats. The methods based on de novo consensus such as the REPET package combine both the previous strategies. Each of these approaches presents advantages and disadvantages and thus, different strategies or pipelines have been developed to improve TE annotation (Guizard et al. 2016; Su et al. 2019).

12.3 The Impact of Transposable Elements on Genome Evolution: Adaptations and Speciation

Vertebrates represent a highly successful lineage that has adapted to a wide range of different environments from arid deserts to cold polar regions or from high altitudes to deep oceans. They originated during the Cambrian explosion and the appearance of relevant innovations allowed the colonization of a broad range of biotopes.

The comprehension of the evolution of organisms is based on knowledge of the functional and structural complexity of their genomes. Given the significant impact of TEs in genome plasticity, it is mandatory to get information on content, diversity, distribution, and the family abundance of mobile elements.

In this context the sequenced genomes available allowed comparative analyses to be performed and the evolution of vertebrates (Fig. 12.3), the origin, and the successful diversification of these organisms to be traced.

TEs are widely present in the genomes of mammals, non-bird reptiles, sharks, lamprey, and some fish but are poorly represented in the genomes of birds and some fish, such as puffer fish. In fact, mammals contain 10 times more TEs than birds; the zebrafish genome is composed of 55% TEs while Tetraodon has less than 6%. Although all types of TEs have been identified in vertebrates, the composition, the copy number and the age of the mobile elements are extremely variable (Chalopin et al. 2015; Warren et al. 2015). TE diversity decreases from basal sarcopterygians to mammals. Indeed, few TE superfamilies are present in the genomes of mammals and birds, unlike reptiles and amphibians.

Within superfamilies, the human genome contains 20% of L1 while the zebrafish genome harbors more than 30 different L1 families even if with a lower copy number

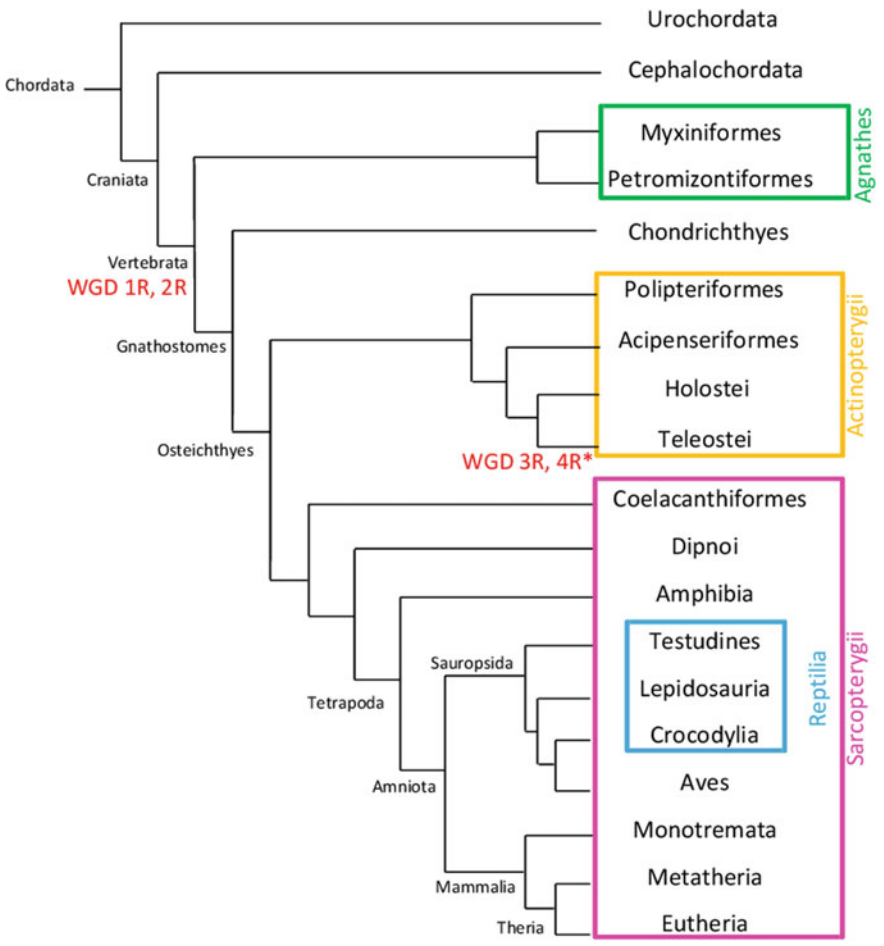


Fig. 12.3 Cladogram showing evolutionary relationships between the main lineages of chordates. whole-genome duplication (WGD) events in vertebrate evolution are indicated in red: 1R and 2R occurred before the divergence of Vertebrata, 3R in Teleost and 4R* in salmonids

(Furano et al. 2004). The prevalence of a specific TE family could be due to competition, rate of transposition, rate of DNA elimination, population size, mode of reproduction, and host defense mechanisms. Moreover, horizontal transfer can also affect TE diversity, leading to the insertion of mobile elements from distant species into a new genome.

Overall, these observations indicate that TEs might have had a different impact on genome evolution in various lineages. Indeed, it is well-known that TEs are responsible for the origin of key adaptations leading to evolutionary advantages and the success of host species (Chalopin et al. 2015; Warren et al. 2015) and thus could be among the main drivers of speciation and major evolutionary transitions. Notably,

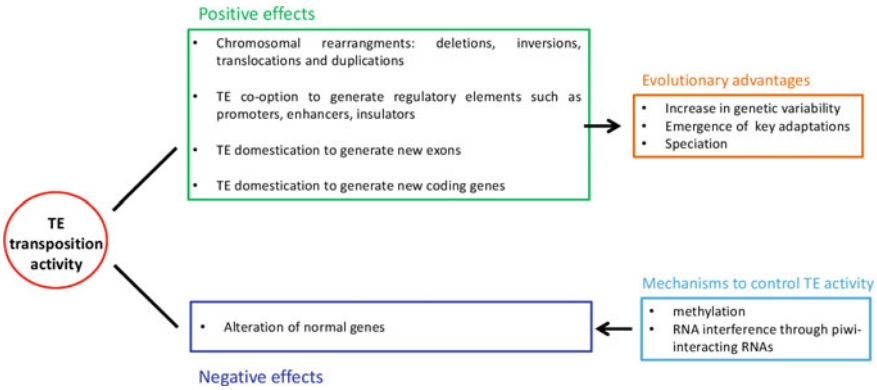


Fig. 12.4 Main effects of TE transposition activity. Positive effects are included in the green box while negative effect in blue box. On the upper right side, evolutionary advantages due to positive effects are indicated; on the lower right side, the defense mechanisms adopted by the host against the negative effect of TE transposition are listed

one of the most important events in the evolution of vertebrates was the transition from water to land life that was accompanied by drastic changes in genome size and in the percentage of TEs, as observed in lungfish and salamanders.

Given their activity, TEs play a key role in genome organization through chromosomal rearrangements such as deletions, inversions, translocations, and duplication events (Fig. 12.4) that have provoked a rapid evolution of a specific lineage followed by reproductive isolation, thereby, determining species diversification (Rebollo et al. 2010).

Moreover, mobile elements have significantly contributed to the complexity of vertebrate transcriptome and proteome (Horie et al. 2007). In fact, several reports have discussed the ability of TEs to generate regulatory elements, genetic novelties, and functional innovations (Fig. 12.4). In humans, 4% of genes contains coding sequences derived from TEs as well as 25% of promoters (Nekrutenko and Li 2001; van de Lagemaat et al. 2003).

TEs can insert near promoter regions and can be coopted to alter the gene expression of the nearby genes (Thornburg et al. 2006). Among the TE-derived regulatory sequences, the involvement of ERV elements is well-documented in the emergence of the placenta in mammals, which was one of the most important innovations in vertebrate evolution (Chuong et al. 2013). Indeed, promoters derived from these mobile elements trigger the expression of placenta-specific genes.

The neuronal enhancer responsible for the expression of the proopiomelanocortin gene (POMC) is responsible for encoding the prohormone of the adenocorticotropic hormone, the melanocyte-stimulating hormone, and endorphin derived from a SINE retroelement in mammals. The absence of this element in other non-mammalian vertebrates suggests that this event occurred in the common ancestor of placentals, marsupials, and monotremes (Santangelo et al. 2007). Another example of lineage-specific recruitment of regulatory sequences from TEs is the enhancer derived from a

LF-SINE that controls the expression of the neurodevelopmental gene ISL1 encoding a LIM homeobox transcription factor required for motor neuron differentiation. This regulatory element has been found in mammals, chicken, and frogs suggesting that the co-option event occurred in the common ancestor of tetrapods (Bejerano et al. 2006).

In addition, TEs contribute to the occurrence of new exons in an existing host-coding protein through a process called exonization (Sorek 2007) and this seems to be very frequent in humans in which this process is mainly due to *Alu* elements (Sela et al. 2010). The insertion of these elements occurs preferentially at the beginning of the coding sequence in both human and mouse genes. Moreover, the analysis of SNPs has highlighted a population-specific pattern indicating that exonization may enhance divergence and thus speciation (Sela et al. 2010). Similarly, in primates, the histone methyltransferase SETMAR presents an exon derived from a mariner-like DNA transposon (Cordaux et al. 2006).

TEs can generate not only new exons but also entire new coding genes through a process called molecular domestication or exaptation. The emergence of new genes enriches the gene repertoire in genomes and thus represents an important contribution to the evolution of organisms. The immune system adaptability of jawed vertebrates took advantage of the emergence of the genes *RAG1* and *RAG2* derived from a transposon about 500 Mya. Indeed, these genes encode the recombinase which catalyzes the V(D)J recombination responsible for the generation of a wide repertoire of antibodies (Kapitonov and Jurka 2005; Schatz and Swanson 2011). A key event in the evolution of mammals is certainly represented by the emergence of the placenta, a specialized organ whose purpose is to guarantee the exchange of water, nutrients, and gasses between the mother and the developing fetus. Several genes expressed in this structure are derived from retrotransposons (Henke et al. 2015). One example is represented by the genes *syncytin-1* and *syncytin-2* that derived from the exaptation of retroviral *env* genes and are involved in cell–cell fusion and in the differentiation of the trophoblast layer in the placenta (Vernochet et al. 2014). Moreover, the mammalian protein CENP-B that binds the centromeric 17 base-pair CENP-B box derived from a pogo-like transposase before the divergence of placental mammals, marsupials, and monotremes (Casola et al. 2008).

Polyploidization represents a drastic event that is accompanied by substantial rearrangements useful for bringing the genome back to a state of diploidy. Although the mechanisms involved in this reinstatement are not completely understood, large scale movements, in particular, due to TEs, have been hypothesized as playing a major role in shaping genomes. Therefore, these observations suggest that polyploidization is associated with bursts of TE activity (Matzke and Matzke 1998). Moreover, these events have led to an increase in gene redundancy and less selective constraints against insertional mutagenesis resulting in an increase in TE content. In vertebrates, two rounds of whole-genome duplication are known to have occurred in the agnathe-gnathostome ancestor after the divergence from urochordates and cephalochordates,

a third event occurred in teleost ancestor and a fourth in salmonids (Allendorf and Thorgaard 1984) (Fig. 12.3). Analyses have shown that bursts of transposon activity took place after genome duplication in salmonids and coincide with speciation events in this lineage (de Boer et al. 2007).

Mobile elements have also been proposed as responsible for the rapid adaptation of invasive species to new environments despite the reduction in genetic variation characterizing these species as a result of a genetic bottleneck. In fact, adaptation to novel habitats represents a stress condition that induces changes in the epigenetic control of TEs; consequently, TE transposition is altered and mobile elements contribute to increase genetic diversity (Stapley et al. 2015).

Despite the positive effects that TEs may have for the host genome, their movement could have strongly deleterious consequences and therefore, organisms have developed various mechanisms to control TE activity (Fig. 12.4). Mobile elements can be inactivated by methylation and/or interference of small RNAs such as piwi-interacting RNAs (Malone and Hannon 2009; Biscotti et al. 2017). However, under stress conditions, these mechanisms can be neutralized leading to an increase in TE activity (Piacentini et al. 2014).

The resolution of the C-value paradox is an ambitious challenge that many research groups are addressing in order to unravel why a parsimonious energy system such as the cellular one can tolerate a great amount of repetitive DNA. The advent of next-generation sequencing technologies has certainly provided a greater availability of genomic data that, as discussed in the present chapter, have strongly contributed to gaining insight into the functional, structural, and evolutionary meaning of repetitive DNA.

The ever-increasing number of available transcriptomes, together with more accurate annotations of TEs, will also allow information to be obtained on the transcriptional activity of the mobilome.

References

Adams RH, Blackmon H, Reyes-Velasco J et al (2016) Microsatellite landscape evolutionary dynamics across 450 million years of vertebrate genome evolution. Genome 59:295–310

Alföldi J, Di Palma F, Grabherr M et al (2011) The genome of the green anole lizard and a comparative analysis with birds and mammals. Nature 477:587–591

Allendorf FW, Thorgaard GH (1984) Tetraploidy and the evolution of salmonid fishes. In: Turner BJ (eds) Evolutionary genetics of fishes. Monographs in evolutionary biology. Springer, Boston, MA

Amemiya CT, Alfoldi J, Lee AP et al (2013) The African coelacanth genome provides insight into tetrapod evolution. Nature 496:311–316

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

Bejerano G, Lowe CB, Ahituv N et al (2006) A distal enhancer and an ultraconserved exon are derived from a novel retroposon. Nature 441:87–90

Bennett MD (1971) The duration of meiosis. Proc R Soc Lond B Biol Sci 178:259–275

Biemont (2010) A brief history of the status of transposable elements: from junk DNA to major players in evolution. Genetics 186:1085–1093
Biscotti MA, Olmo E, Heslop-Harrison JS (2015a) Repetitive DNA in eukaryotic genomes. Chromosome Res 23:415–420
Biscotti MA, Canapa A, Forconi M, Olmo E, Barucca M (2015b) Transcription of tandemly repetitive DNA: functional roles. Chromosome Res 23:463–477
Biscotti MA, Gerdol M, Canapa A et al (2016) The lungfish transcriptome: a glimpse into molecular evolution events at the transition from water to land. Sci Rep 6:21571
Biscotti MA, Canapa A, Forconi M et al (2017) The small noncoding RNA processing machinery of two living fossil species, lungfish and coelacanth, gives new insights into the evolution of the argonaute protein family. Genome Biol Evol 9:438–453
Biscotti MA, Barucca M, Canapa A (2018) New insights into the genome repetitive fraction of the Antarctic bivalve *Adamussium colbecki*. PLoS ONE 13(3):e0194502
Blaxter M (2010) Genetics. Revealing the dark matter of the genome. Science 330:1758–1759
Bonnivard E, Catrice O, Ravaux J et al (2009) Survey of genome size in 28 hydrothermal vent species covering 10 families. Genome 52:524–536
Bourque G, Burns KH, Gehring M et al (2018) Ten things you should know about transposable elements. Genome Biol 19:199
Britten RJ, Graham DE, Neufeld BR (1974) Analysis of repeating DNA sequences by reassociation. Method Enzymol 29:363–405
Burton DW, Bickham JW, Genoways HH (1989) Flow-cytometric analyses of nuclear DNA content in four families of neotropical bats. Evolution 43:756–765
Canapa A, Barucca M, Biscotti MA et al (2015) Transposons, genome size, and evolutionary insights in animals. Cytogenet Genome Res 147:217–239
Carducci F, Barucca M, Canapa A et al (2018) Rex retroelements and teleost genomes: an overview. Int J Mol Sci 19:11
Casola C, Hucks D, Freschotte C et al (2008) Convergent domestication of pogo-like transposases into centromere-binding proteins in fission yeast and mammals. Mol Biol Evol 25:29–41
Castoe TA, de Konig AP, Hall KT et al (2013) The Burmese python genome reveals the molecular basis for extreme adaptation in snakes. Proc Natl Acad Sci USA 110:20645–20650
Castoe TA, Hall KT, Guibotsy Mboulas ML et al (2011) Discovery of highly divergent repeat landscapes in snake genomes using high-throughput sequencing. Genome Biol Evol 3:641–653
Cavalier-Smith T (1978) Nuclear volume control by nucleoskeletal DNA, selection for cell volume and cell growth rate, and the solution of the DNA C-value paradox. J Cell Sci 43:247–278
Chalopin D, Naville M, Plard F et al (2015) Comparative analysis of transposable elements highlights mobilome diversity and evolution in vertebrates. Genome Biol Evol 7:567–580
Chalopin D, Volff JN (2017) Analysis of the spotted gar genome suggests absence of causative link between ancestral genome duplication and transposable element diversification in teleost fish. J Exp Zool B Mol Dev Evol 328:629–637
Chaves R, Ferreira D, Mendes-da-Silva et al (2017) FA-SAT is an old satellite DNA frozen in several bilateria genomes. Genome Biol Evol 9:3073–3087
Chuong EB, Rumi MAK, Soares MJ et al (2013) Endogenous retroviruses function as species-specific enhancer elements in the placenta. Nat Genet 45:325–329
Cordaux R, Udit S, Batzer MA et al (2006) Birth of a chimeric primate gene by capture of the transposase gene from a mobile element. Proc Natl Acad Sci USA 103:8101–8106
de Boer JG, Yazawa R, Davidson WS et al (2007) Bursts and horizontal evolution of DNA transposons in the speciation of pseudotetraploid salmonids. BMC Genom 8:422
Doolittle WF, Sapienza C (1980) Selfish genes, the phenotype paradigm and genome evolution. Nature 284:601–603
Dufresne F, Jeffery N (2011) A guided tour of large genome size in animals: what we know and where we are heading. Chromosome Res 19:925–938

Feschotte C, Swamy L, Wessler SR (2003) Genome-wide analysis of *mariner*-like transposable elements in rice reveals complex relationships with stowaway miniature inverted repeat transposable elements (MITEs). Genetics 163:747–758
Forconi M, Chalopin D, Barucca M et al (2014) Transcriptional activity of transposable elements in coelacanth. J Exp Zool B Mol Dev Evol 322:379–389
Furano AV, Duvernell DD, Boissinot S (2004) L1 (LINE-1) retrotransposon diversity differs dramatically between mammals and fish. Trends Genet 20:9–14
Garrido-Ramos MA (2015) Satellite DNA in plants: more than just rubbish. Cytogenet Genome Res 146:153–170
Garrido-Ramos MA (2017) Satellite DNA: an evolving topic. Genes (Basel) 8(9)
Girgis HZ (2015) Red: an intelligent, rapid, accurate tool for detecting repeats de-novo on the genomic scale. BMC Bioinf 16:227
Goerner-Potvin P, Bourque G (2018) Computational tools to unmask transposable elements. Nat Rev Genet 19:688–704
Green RE, Braun EL, Armstrong J et al (2014) Three crocodilian genomes reveal ancestral patterns of evolution among archosaurs. Science 346:1254449
Gregory TR (2019) Animal genome size database. http://www.genomesize.com
Guizard S, Piégu B, Arensburger P et al (2016) Deep landscape update of dispersed and tandem repeats in the genome model of the red jungle fowl, *Gallus gallus*, using a series of de novo investigating tools. BMC Genom 17(1):659
Henke C, Strissel PL, Schubert MT et al (2015) Selective expression of sense and antisense transcripts of the sushi-ichi-related retrotransposon-derived family during mouse placentogenesis. Retrovirology 12:9
Heslop-Harrison JS, Schwarzacher T (2011) Organisation of the plant genome in chromosomes. Plant J 66:18–33
Horie K, Saito ES, Keng VW et al (2007) Retrotransposons influence the mouse transcriptome: implication for the divergence of genetic traits. Genetics 176:815–827
Hughes AL, Hughes MK (1995) Small genomes for better fliers. Nature 377:391
Kapitonov VV, Jurka J (2005) RAG1 core and V(D)J recombination signal sequences were derived from Transib transposons. PLoS Biol 3:e181
Kapitonov VV, Jurka J (2008) A universal classification of eukaryotic transposable elements implemented in Repbase. Nat Rev Genet 9:411–412
Kapranov P, Laurent GS (2012) Genomic 'dark matter': implications for understanding human disease mechanisms, diagnostics, and cures. Front Genet 3:95
Kraaijeveld K (2010) Genome size and species diversification. Evol Biol 37:227–233
Kramerov DA, Vassetzky NS (2011) Origin and evolution of SINEs in eukaryotic genomes. Heredity 107:487–495
Lerat E (2010) Identifying repeats and transposable elements in sequenced genomes: how to find your way through the dense forest of programs. Heredity 104:520–533
Levitsky VG, Babenko VN, Vershinin AV (2013) The roles of the monomer length and nucleotide context of plant tandem repeats in nucleosome positioning. J Biomol Struct Dyn 32:115–126
López-Flores I, Garrido-Ramos MA (2012) The repetitive DNA content of eukaryotic genomes. In: Garrido-Ramos MA (ed) Repetitive DNA. Genome Dyn 7:1–28
Lynch M, Conery JS (2003) The origins of genome complexity. Science 302:1401–1404
Malik HS, Henikoff S, Eickbush TH (2000) Poised for contagion: evolutionary origins of the infectious abilities of invertebrate retroviruses. Genome Res 10:1307–1318
Malone CD, Hannon GJ (2009) Small RNAs as guardians of the genome. Cell 136:656–668
Matzke MA, Matzke AJ (1998) Polyploidy and transposons. Trends Ecol Evol 13:241
Metcalfe CJ, Filée J, Germon I et al (2012) Evolution of the Australian lungfish (Neoceratodus forsteri) genome: a major role for CR60 and L2 LINE elements. Mol Biol Evol 29:3529–3539
Mravinac B, Plohl M, Ugarković D (2005) Preservation and high sequence conservation of satellite DNAs suggest functional constraints. J Mol Evol 61:542–550

Naville M, Warren IA, Haftek-Terreau Z et al (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

Nekrutenko A, Li WH (2001) Transposable elements are found in a large number of human protein-coding genes. Trends Genet 17:619–621

Nowoshilow S, Schloissnig S, Fei JF et al (2018) The axolotl genome and the evolution of key tissue formation regulators. Nature 554:50–55

Ohno S (1972) So much 'junk' DNA in our genome. In: Smith HH (ed) Evolution of genetic systems. Gordon and Breach, New York

Olmo E (1983) Nucleotype and cell size in vertebrates: a review. Basic Appl Histochem 27:227–256

Organ CL, Shedlock AM, Meade A et al (2007) Origin of avian genome size and structure in non-avian dinosaurs. Nature 446:180–184

Organ CL, Canoville A, Reisz RR et al (2011) Paleogenomic data suggest mammal-like genome size in the ancestral amniote and derived large genome size in amphibians. J Evol Biol 24:372–380

Pagán HJ, Macas J, Novák P et al (2012) Survey sequencing reveals elevated DNA transposon activity, novel elements, and variation in repetitive landscapes among vesper bats. Genome Biol Evol 4:575–585

Pasquesi GIM, Adams RH, Card DC et al (2018) Squamate reptiles challenge paradigms of genomic repeat element evolution set by birds and mammals. Nat Commun 9:2774

Peona V, Weissensteiner MH, Suh A (2018) How complete are "complete" genome assemblies?-An avian perspective. Mol Ecol Resour 18(6):1188–1195

Petraccioli A, Odierna G, Capriglione T et al (2015) A novel satellite DNA isolated in Pecten jacobaeus shows high sequence similarity among molluscs. Mol Genet Genomics 290:1717–1725

Petrov DA (2002) Mutational equilibrium model of genome size evolution. Theor Popul Biol 61:531–544

Piacentini L, Fanti L, Specchia V et al (2014) Transposons, environmental changes, and heritable induced phenotypic variability. Chromosoma 123:345–354

Ray DA, Pagan HJ, Thompson ML et al (2007) Bats with hATs: evidence for recent DNA transposon activity in genus *Myotis*. Mol Biol Evol 24:632–639

Ribet D, Harper F, Esnault C et al (2008) The GLN family of murine endogenous retroviruses contains an element competent for infectious viral particle formation. J Virol 82:4413–4419

Rebollo R, Horard B, Hubert B et al (2010) Jumping genes and epigenetics: towards new species. Gene 454:1–7

Santangelo AM, de Souza FSJ, Franchini LF et al (2007) Ancient exaptation of a CORE–SINE retroposon into a highly conserved mammalian neuronal enhancer of the proopiomelanocortin gene. PLoS Genet 3:1813–1826

Schatz DG, Swanson PC (2011) V(D)J recombination: mechanisms of initiation. Annu Rev Genet 45:167–202

Schueler MG, Higgins AW, Rudd MK, Gustashaw K, Willard HF (2001) Genomic and genetic definition of a functional human centromere. Science 294:109–115

Sela N, Mersch B, Hotz-Wagenblatt A et al (2010) Characteristics of transposable element exonization within human and mouse. PLoS ONE 5:e10907

Shaffer HB, Minx P, Warren DE et al (2013) The western painted turtle genome, a model for the evolution of extreme physiological adaptations in a slowly evolving lineage. Genome Biol 14:R28

Smith JJ, Kuraku S, Holt C et al (2013) Sequencing of the sea lamprey (*Petromyzon marinus*) genome provides insights into vertebrate evolution. Nat Genet 45:415–421

Smith JJ, Timoshevskaya N, Ye C et al (2018) The sea lamprey germline genome provides insights into programmed genome rearrangement and vertebrate evolution. Nat Genet 50(2):270–277

Sorek R (2007) The birth of new exons: mechanisms and evolutionary consequences. RNA 13:1603–1608

Stapley J, Santure AW, Dennis SR (2015) Transposable elements as agents of rapid adaptation may explain the genetic paradox of invasive species. Mol Ecol 24:2241–2252

Su W, Gu X, Peterson T (2019) TIR-learner, a new ensemble method for TIR transposable element annotation, provides evidence for abundant new transposable elements in the maize genome. Mol Plant 12:447–460

Sun C, Lopez Arriaza JR, Mueller RL (2012) Slow DNA loss in the gigantic genomes of salamanders. Genome Biol Evol 4:1340–1348

Sun YB, Xiong ZJ, Xiang XY et al (2015) Whole–genome sequence of the Tibetan frog *Nanorana parkeri* and the comparative evolution of tetrapod genome. Proc Natl Acad Sci USA 112:1257–1262

Swift H (1950) The constancy of desoxyribose nucleic acid in plant nuclei. Proc Natl Acad Sci USA 36:643–654

Tempel S (2012) Using and understanding RepeatMasker. Methods Mol Biol 859:29–51

Thomas CA Jr (1971) The genetic organization of chromosomes. Annu Rev Genet 5:237–256

Thornburg BG, Gotea V, Makalowski W (2006) Transposable elements as a significant source of transcription regulating signals. Gene 365:104–110

Van den Bussche RA, Longmire JL, Baker RJ (1995) How bats achieve a small C-value: frequency of repetitive DNA in Macrotus. Mamm Genome 6:521–525

van de Lagemaat LN, Landry JR, Mager DL et al (2003) Transposable elements in mammals promote regulatory variation and diversification of genes with specialized functions. Trends Genet 19:530–536

Vendrely R, Vendrely C (1948) La teneur du noyau cellulaire en acide désoxyribonucléique à travers les organes, les individus et les espèces animales: techniques et premiers résultats. Experientia 4:434–436

Venkatesh B, Lee AP, Ravi V et al (2014) Elephant shark genome provides unique insights into gnathostome evolution. Nature 505:174–179

Vernochet C, Redelsperger F, Harper F et al (2014) The captured retroviral envelope syncytin-A and syncytin-B genes are conserved in the Spalacidae together with hemotrichorial placentation. Biol Reprod 91:148

Vinogradov AE (1995) Nucleotypic effect in homeotherms: body-mass-corrected basal metabolic rate of mammals is related to genome size. Evolution 49:1249–1259

Vinogradov AE (1997) Nucleotypic effect in homeotherms: body-mass independent resting metabolic rate of passerine birds is related to genome size. Evolution 51:220–225

Vinogradov AE (1998) Buffering: a possible passive-homeostasis role for redundant DNA. J Theor Biol 193:197–199

Vinogradov AE (2003) Selfish DNA is maladaptive: evidence from the plant red list. Trends Genet 19:609–614

Vinogradov AE (2004) Testing genome complexity. Science 304:389–390

Volff JN, Körting C, Froschauer A et al (2001) Non-LTR retrotransposons encoding a restriction enzyme-like endonuclease in vertebrates. J Mol Evol 52:351–360

Volff JN, Bouneau L, Ozouf-Costaz C et al (2003) Diversity of retrotransposable elements in compact pufferfish genomes. Trends Genet 19:674–678

Voss SR, Putta S, Walker JA et al (2013) Salamander Hox clusters contain repetitive DNA and expanded non-coding regions: a typical Hox structure for non-mammalian tetrapod vertebrates? Hum Genomics 7:9

Wang Z, Pascual-Anaya J, Zadissa A et al (2013) The draft genomes of soft-shell turtle and green sea turtle yield insights into the development and evolution of the turtle specific body plan. Nat Genet 45:701–706

Warren IA, Naville M, Chalopin D et al (2015) Evolutionary impact of transposable elements on genomic diversity and lineage-specific innovation in vertebrates. Chromosome Res 23:505–531

Wicker T, Sabot F, Hua-Van A et al (2007) A unified classification system for eukaryotic transposable elements. Nat Rev Genet 8:973–982

Wicker T, Robertson JS, Schulze SR et al (2005) The repetitive landscape of the chicken genome. Genome Res 15:126–136

Chapter 13
Genetic Diversity of the Apoptotic Pathway in Insects

Mélanie Ribeiro Lopes, Nicolas Parisot, Patrick Callaerts and Federica Calevro

Abstract Programmed cell death is an intrinsic part of normal development, physiology and organismal homeostasis. Apoptosis is a widespread form of regulated cell death, controlled by a genetically encoded machinery conserved throughout evolution. In the highly diverse group of insects, apoptotic pathways have been characterized in only a few dipteran and lepidopteran species, where they have been shown to be essential for development, metamorphosis and immunity-related processes. The lack of studies in other insect orders clearly limits our understanding of the role of apoptosis in the life history of insects.

13.1 Introduction

Animal cells can die in a variety of ways but apoptosis is the most widespread and conserved form of regulated cell death, with evidence for a genetically encoded apoptotic machinery conserved throughout metazoan evolution, from nematodes to arthropods, from fishes to mammals (Lamkanfi et al. 2002; Kumar 2007; Berthelet and Dubrez 2013). By selectively eliminating unwanted or damaged cells, apoptosis plays a fundamental role in many biological processes, such as development (Miura 2012), tissue homeostasis (Pellettieri and Sanchez Alvarado 2007) and immune response (Opferman and Korsmeyer 2003).

Apoptosis can be triggered by either external or internal signals that both ultimately lead to the activation of evolutionarily conserved proteases named caspases (**c**ysteine-dependent **asp**artate-specific prote**ases**), which constitute the effector arm of the apoptotic machinery (Lamkanfi et al. 2002; Kumar 2007). Once activated,

M. Ribeiro Lopes (✉) · N. Parisot · F. Calevro (✉)
Univ Lyon, INSA-Lyon, INRA, UMR0203, BF2I, F-69621, Villeurbanne, France
e-mail: melanie.ribeiro-lopes@insa-lyon.fr

F. Calevro
e-mail: federica.calevro@insa-lyon.fr

P. Callaerts
Department of Human Genetics, Laboratory of Behavioral and Developmental Genetics, KU Leuven, B-3000 Leuven, Belgium

P. Pontarotti (ed.), *Evolution, Origin of Life, Concepts and Methods*,
https://doi.org/10.1007/978-3-030-30363-1_13

caspases selectively cleave key cellular substrates within the cell, thus promoting its ordered disassembly. Apoptosis is associated with typical morphological changes, i.e., rounding up of the cell, reduction of cellular volume (pyknosis), chromatin condensation, nuclear fragmentation (karyorrhexis) and plasma membrane blebbing (Saraste and Pulkki 2000).

Since apoptosis is a very potent and fast process, caspase activation and activity are subject to complex regulation by a wide variety of cellular pro- and anti-apoptotic regulators (Berthelet and Dubrez 2013; Vasudevan and Ryoo 2015; Kale et al. 2018). The most documented regulators of apoptosis include proteins of the Bcl-2 (**B**-**c**ell CLL/**l**ymphoma **2**) and IAP (**i**nhibitor of **a**poptosis **p**rotein) families, as well as several IAP antagonists. These regulators are conserved throughout the Animal Kingdom but their relative contribution to the apoptotic process varies widely from one species to another. For instance, proteins of the Bcl-2 family are the main apoptotic regulators in mammals but only have a minor role in the fruit fly *Drosophila melanogaster* (Berthelet and Dubrez 2013).

Moreover, even though apoptotic proteins have been identified in a wide range of organisms, our current knowledge of the apoptotic processes comes almost entirely from studies conducted on three species: humans, *Caenorhabditis elegans* and *Drosophila melanogaster* (Parrish et al. 2013; Denton and Kumar 2015). These studies have led to tremendous advances in our knowledge on the functioning of apoptosis pathways, but they are not necessarily representative of all the species belonging to the same phylogenetic group. Indeed, some non-model species display an expansion of apoptosis-related genes (Bryant et al. 2010), which may indicate the emergence of new functions and/or activation patterns.

In the highly diverse group of insects, which comprises several species of high agronomical and medical interest, apoptosis has been investigated in only a few species from the dipteran (Kumar and Doumanis 2000; Blitvich et al. 2002; Clem 2005; Li et al. 2007; Bryant et al. 2008; Cooper et al. 2009b; Orme and Meier 2009; Liu and Clem 2011; Wang and Clem 2011; Denton et al. 2013) and lepidopteran (Ahmad et al. 1997; Zhang et al. 2010; Courtiade et al. 2011; Suganuma et al. 2011; Huang et al. 2013) orders, and functional studies are very often restricted to *D. melanogaster*. These studies indicate that apoptosis is essential for development, metamorphosis and immunity-related processes. Given the high evolutionary distance among insects and the differences they show in terms of habitats, shapes, sizes, lifestyles and developmental and immune processes, the expectation is that apoptotic processes will also show variation. As an example, fruit flies and mosquitoes, which are both members of the order Diptera, are separated by ~250 million years of evolution and have been exposed to very different evolutionary pressures, with mosquitoes having evolved an aquatic larval stage and relying on vertebrate blood feeding for reproduction (Grimaldi and Engel 2005). Thus, even though apoptosis actors are the same, the way in which they interact with each other and are regulated might vary widely to allow for these differences.

In the present paper, after describing apoptotic pathways in metazoans, we summarize what is currently known about apoptosis in insects. We will emphasize the

differences existing between insects and mammals, but also among insects, thus highlighting how the focus on a few species limits our understanding of the role of apoptosis in insect life history.

13.2 The Main Actors of Apoptosis in Metazoa

13.2.1 Caspases

Caspases are the main executors of apoptosis, and their activation is necessary and sufficient for its completion. The importance of caspases in apoptosis was first demonstrated in *C. elegans* with the identification of the *ced-3* gene, work that was honored with the 2002 Nobel Prize in Physiology or Medicine (Yuan et al. 1993). The *ced-3* gene encodes a cysteine protease with similar properties to the mammalian **i**nterleukin-1-beta **c**onverting **e**nzyme (ICE, today known as caspase-1). We now know that caspases have been conserved in metazoans throughout evolution (Lamkanfi et al. 2002; Kumar 2007), and proteins related to caspases have been identified in plants, fungi and even bacteria (Tsiatsiani et al. 2011). In all these organisms, caspases and caspase-like proteins seem to play a role in cell death, highlighting the fundamental importance of apoptosis.

Although caspase sequences are poorly conserved between organisms, they all share a similar structure consisting of a prodomain and a catalytic domain (Fuentes-Prior and Salvesen 2004; Kumar 2007; Parrish et al. 2013). The catalytic domain, named "peptidase C14" or "CASc," can be further divided into large (p20) and small (p10) catalytic subunits (Fig. 13.1a). The active site of the catalytic domain contains a conserved Q-A-C-[RQG]-[GN] sequence, which forms a substrate-binding groove with a catalytic cysteine at its center (Fig. 13.1c). Most caspases cleave their substrate after an aspartate to activate or inhibit the target protein. This cleavage is dependent on caspase activation. Indeed, caspases are initially produced as inactive zymogen precursor proteins and are activated through proteolytic cleavages that separate the small from the large subunit, and often remove the prodomain. Two large and two small subunits form the active protease (Fig. 13.1b). This results in a significant conformational change of the substrate-binding groove, which rotates to accommodate substrates and brings the catalytic cysteine into position for catalysis. Activation of caspases ensures that the cellular components are degraded in a controlled manner, with minimal effects of cell death on surrounding tissues.

While the catalytic subunits are structurally conserved, prodomains can vary significantly, in both length and composition. Depending on their size, caspases can be grouped into two classes: effector caspases (containing a short prodomain of about 30 amino acids) and initiator caspases (containing a long prodomain of 80 or more amino acids) (Kumar 2007).

Long prodomains often harbor protein–protein interaction domains, such as CARD (**c**aspase **a**ctivation and **r**ecruitment **d**omain) or DED (**d**eath **e**ffector

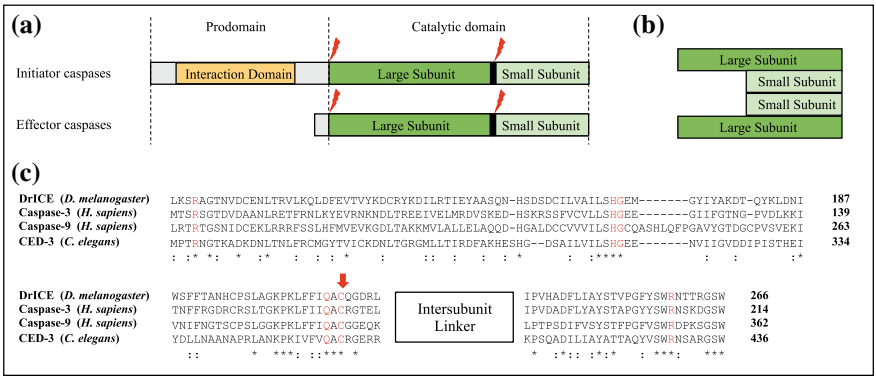


Fig. 13.1 Conservation of the domain structure and sequence of caspases. **a** Typical domain composition of initiator and effector caspases. All caspases have a C-terminal catalytic domain composed of two (large and small) subunits, separated by a linker. Initiator caspases contain a long prodomain at the N-terminus that harbors one or more protein–protein interaction domains (i.e., CARD (**C**aspase **A**ctivation and **R**ecruitment **D**omain), DED (**D**eath **E**ffector **D**omain)). Activation of caspases requires cleavage of the zymogen protein between the prodomain and the catalytic domain and between the large and small subunits (red lightning symbols). **b** The activated caspase forms a tetramer composed of two large and two small subunits. **c** Alignment of the partial catalytic domains of selected caspases. Residues predicted to be involved in substrate binding and recognition are highlighted in red, and the catalytic cysteine is indicated with a red arrow. Strictly conserved residues are indicated by «*» and conserved substitutions by «:»

domain), which bind similar motifs in adaptor proteins following the reception of death stimuli (Kumar 2007). Because different initiator caspases have different interaction domains, they can respond to different types of stimuli. For instance, in mammals, there are at least two distinct pathways of caspase activation (Parrish et al. 2013). In the extrinsic pathway, recognition of extracellular death signals by membrane receptors, such as Fas, leads to the recruitment of the adaptor protein FADD (**F**as-**a**ssociated protein with **d**eath **d**omain), which then recruits and activates the initiator caspases-8 and 10. In the intrinsic pathway, intracellular stress signals lead to the activation of pro-apoptotic regulators that trigger the release of cytochrome c from mitochondria. This small protein interacts with the adaptor protein Apaf-1 (**a**poptotic **p**eptidase **a**ctivating **f**actor **1**), which can then recruit the initiator caspase-9 to form the apoptosome complex in which caspase-9 is activated. The two pathways join after the activation of the initiator caspases, which, once activated, can in turn cleave and activate the effector caspases. In mammals, the main effector caspases are caspases-3 and 7 (Fig. 13.2).

In *D. melanogaster* and *C. elegans,* only one pathway has been identified but the activation cascade remains similar: An apoptotic stimulus leads to the recruitment and activation of an initiator caspase by an adaptor protein and the activated initiator caspase can then activate the effector caspases (Parrish et al. 2013) (Fig. 13.2).

Once activated, effector caspases are implicated in the degradation of over 600 cellular components, some of which are necessary to induce the morphological changes

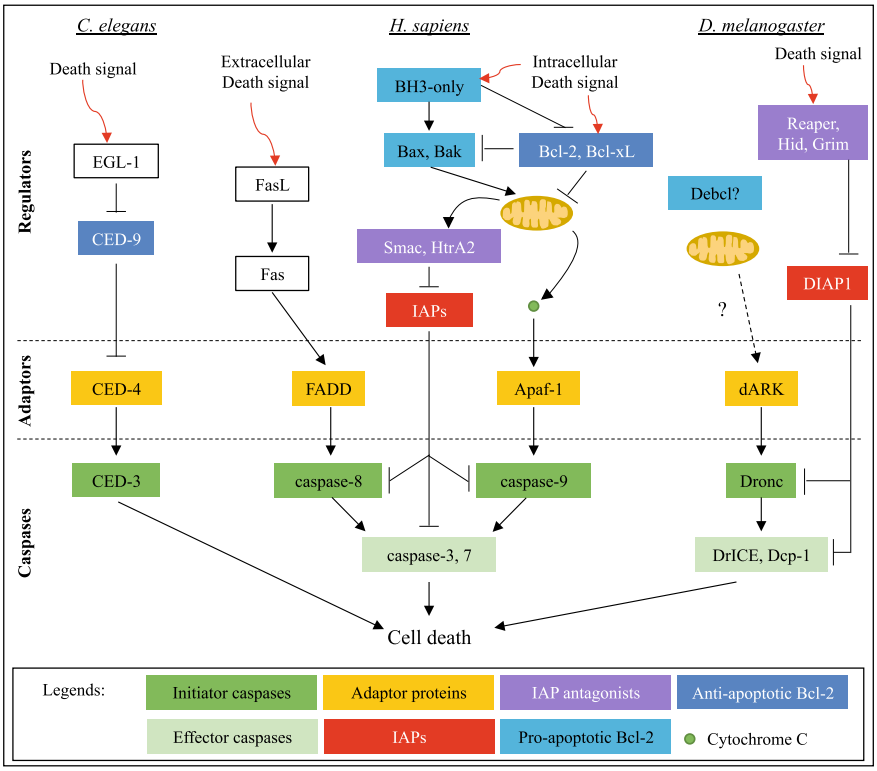


Fig. 13.2 Components of the apoptosis pathway are conserved across evolution. Many of the core proteins of the apoptotic pathway are conserved in *Caenorhabditis elegans*, *Drosophila melanogaster* and humans, and these protein families share some functional similarities (noted by matching colors in the figure)

underlying apoptosis (Fuentes-Prior and Salvesen 2004). Since caspases are so effective, it is important that their activation is finely regulated, which is reflected in many cellular pro- and anti-apoptotic regulators that have been identified.

13.2.2 Apoptosis Regulators

The most documented regulators of apoptosis include Bcl-2 family members as well as IAPs and IAP antagonists.

IAP proteins are among the main negative regulators of caspases. Even though they were first discovered in baculoviruses (Crook et al. 1993), they have been found in almost all organisms ranging from yeasts to mammals (Uren et al. 1998), where they play important roles in regulating apoptosis but also cell division, cell mobility and immunity. They can inhibit caspase function by sterically blocking the active site

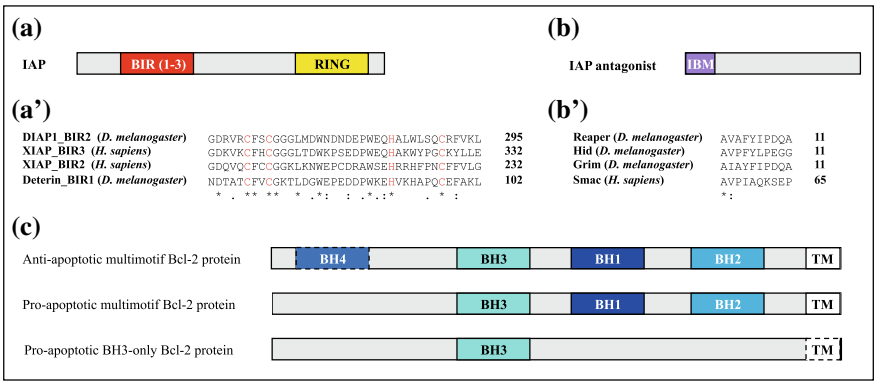


Fig. 13.3 Conservation of the domain/motif structure and sequence of apoptosis regulators. **a** Typical domain composition of an IAP (**i**nhibitor of **a**poptosis **p**rotein). All IAP family members contain from one to three BIR (**b**aculoviral **I**AP **r**epeat) domains and many of them have a RING (**R**eally **I**nteresting **N**ew **G**ene) domain at the C-terminus. Other domains found in IAPs include UBC (**ub**iquitin **c**onjugating enzyme domain), UBA (**ub**iquitin **a**ssociated domain), CARD (**c**aspase **a**ctivation and **r**ecruitment **d**omain) and NACHT (nucleoside triphosphatase domain). **(a')** Alignment of the partial BIR domains of selected IAPs. Zinc-chelating residues are shown in red. **b** Typical domain composition of an IAP antagonist. IAP antagonists share little similarity apart from their small IBM (**I**AP-**b**inding **m**otif) at the N-terminus. **b'** Alignment of the IBM motif of selected IAP antagonists. **c** Typical motif composition of Bcl-2 family members from the three Bcl-2 subfamilies. All Bcl-2 family members share 1–4 BH (**B**cl-2 **H**omology) motifs. Most Bcl-2 family members also possess a **t**rans**m**embrane domain (TM). Dashed lines indicate that the region has only been found in some proteins of the concerned Bcl-2 subfamily. For all alignments, strictly conserved residues are indicated by « * » and conserved substitutions by «:»

of caspases, by sequestering them away from their substrates or by promoting their degradation (Tenev et al. 2005; Domingues and Ryoo 2012; Berthelet and Dubrez 2013). They are characterized by their ability to directly bind and inhibit activated caspases thanks to an evolutionary conserved **b**aculovirus **I**AP **r**epeat (BIR) domain (Fig. 13.3a). The core component of BIRs is a consensus Cys/His motif (G-x(2)-Y-x(4)-D-x(3)-C-x(2)-C-x(6)-W-x(9)-H-x(6,10)-C) that coordinates a single zinc ion (Berthelet and Dubrez 2013) (Fig. 13.3a'). This forms a surface hydrophobic groove that enables binding to caspases and IAP antagonists. IAPs contain between one and three BIR domains that often have different substrate specificity. Most IAPs also possess one or more additional domains. The most common is a C-terminal RING (**R**eally **I**nteresting **N**ew **G**ene) zinc finger with ubiquitin ligase (E3) activity Budhidarmo and Day 2015). The RING domain targets caspases, IAP antagonists and IAPs themselves for ubiquitination, often leading to their degradation by the proteasome. Alternatively, IAPs can contain UBC (**ub**iquitin **c**onjugating), UBA (**ub**iquitin **a**ssociated), NACHT (**NA**IP (neuronal apoptosis inhibitor protein), **C**2TA (MHC class 2 transcription activator), **H**ET-E (incompatibility locus protein from *Podospora anserina*) and **T**P1 (telomerase-associated protein)) or CARD domains (Budhidarmo and Day 2015).

Several proteins, such as the *C. elegans* protein Bir-1 or the mammalian protein Survivin, contain a structurally distinct BIR domain (Silke and Vaux 2001). These proteins are unable to interact with caspases and do not (or are unlikely to) inhibit apoptosis. Because of that, they are commonly referred to by the term BIRC (**BIR C**ontaining) protein, while the term IAPs is used only for proteins working as apoptosis inhibitors.

The RHG (**R**eaper, **H**id and **G**rim) proteins of *D. melanogaster* were the first IAP antagonists identified, and most of our current knowledge on these proteins comes from studies in this insect (White et al. 1994). IAP antagonist genes encode proteins without any enzymatic activity that all contain an IBM (**I**AP-**b**inding **m**otif) motif, sometimes called RHG motif (Shi 2002) (Fig. 13.3b). This motif is defined by a sequence of four to ten partially conserved residues at the N-terminus of the protein, with an unmodified alanine residue in first position (which can be generated by posttranslational cleavage or by removal of the initiator methionine). It can be identified by the pattern A-[KTVI]-[PAE]-[FEISY] (Shi 2002; Berthelet and Dubrez 2013) (Fig. 13.3b'). IAP antagonists share very little sequence similarity except for their IBM motifs, which complicates their identification. In mammals, the two proteins Smac and HtrA2 contain such a motif and function similarly to the RHG proteins (Hegde et al. 2002; Verhagen et al. 2002). The IBM motif directly binds to a groove within the BIR domain, allowing caspases to be liberated from IAPs. This binding also prevents caspase inhibition by competing for caspase binding and by promoting IAP auto-ubiquitination and subsequent degradation (Yoo et al. 2002; Chai et al. 2003; Orme and Meier 2009). Some IAP antagonists also have additional functions, with an active role in cell death. For instance, Reaper and Grim inhibit general protein translation thus inhibiting de novo IAP protein synthesis (Yoo et al. 2002).

Proteins of the Bcl-2 family are dominant regulators of programmed cell death that can either promote or inhibit apoptosis (Kvansakul and Hinds 2015; Kale et al. 2018). All members of this family share one or more of the four characteristic **B**cl-2 **h**omology motifs (BH1, BH2, BH3 or BH4) that are necessary for their activity. Depending on their functions, proteins of the Bcl-2 family can be further divided into three subfamilies: (1) anti-apoptotic multimotif Bcl-2 proteins, (2) pro-apoptotic multimotif Bcl-2 proteins and (3) BH3-only pro-apoptotic Bcl-2 proteins (Fig. 13.3c). Most of these proteins also contain a transmembrane domain that anchors them to the membrane of various subcellular compartments (i.e., mitochondria, nuclear membrane, endoplasmic reticulum, Golgi apparatus). Their main localization remains in the mitochondria where they execute their cell death-related functions. In mammals, depending on their function, activated Bcl-2 proteins either promote or inhibit **m**itochondrial **o**uter **m**embrane **p**ermeabilization (MOMP). MOMP allows for the release of pro-apoptotic factors, such as cytochrome c or IAP antagonists, otherwise sequestered in the mitochondria intermembrane space (Kalkavan and Green 2018; Kale et al. 2018). For instance, BH3-only pro-apoptotic Bcl-2 proteins can activate pro-apoptotic multimotif Bcl-2 proteins, such as BAX and BAK, which, following various conformation changes, form pores in the mitochondrial outer membrane. Inversely, anti-apoptotic Bcl-2 proteins are able to bind to the BH3 motif of proteins

from the two other subfamilies to inhibit their functions. In contrast, the *C. elegans* Bcl-2-like protein CED-9 directly binds to the initiator caspase CED-4 and blocks its interaction with the effector caspase CED-3 (Malin and Shaham 2015).

13.3 Apoptosis in *Drosophila melanogaster*

D. melanogaster is a versatile model system, thanks to its easy maintenance, rapid generation time and extensive genetic toolkit. This insect has a holometabolous lifecycle, characterized by a complete metamorphosis accompanied by drastic phases of rapid remodeling of the entire organism during development (Denton et al. 2013). Given the important role of apoptosis therein, this offers multiple opportunities to study apoptosis and its regulation.

The *D. melanogaster* genome encodes seven caspases, four IAPs, six IAP antagonists and two Bcl-2 family members (Kornbluth and White 2005; Hay and Guo 2006; Xu et al. 2009) (Fig. 13.4). Among these proteins, some have a more prominent role: the initiator caspase Dronc and its activating protein Ark, the effector caspase DrICE, the caspase negative regulator DIAP1 and the RHG IAP antagonists.

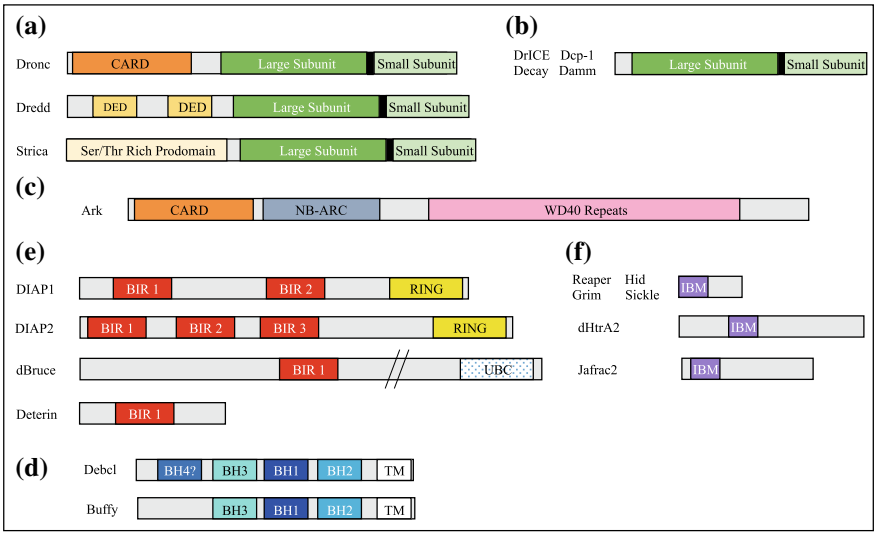


Fig. 13.4 ***Drosophila melanogaster*** **caspases and apoptosis regulators.** Schematic representation of *D. melanogaster* initiator caspases **a**, effector caspases **b**, adaptor proteins **c**, IAP proteins **d**, IAP antagonists **e** and Bcl-2 family members **f**. Each protein is represented with its characteristic domains/motifs. Abbreviations: CARD, **c**aspase **a**ctivation and **r**ecruitment **d**omain; DED, **d**eath **e**ffector **d**omain; NB-ARC, **n**ucleotide-**b**inding **a**daptor shared by **A**PAF-1, **R** proteins and **C**ED-4; BIR, **b**aculoviral **I**AP **r**epeat; RING, **R**eally **I**nteresting **N**ew **G**ene; UBC, **ub**iquitin **c**onjugating **e**nzyme; IBM, **I**AP-**b**inding **m**otif; BH 1-4, **B**cl-2 **H**omology 1-4

This small ensemble of proteins is required for almost all apoptotic cell death and is regulated by complex mechanisms, involving multiple cross-interactions and feedback loops. Unlike in mammals, where there are at least two distinct pathways of caspase activation, a single pathway appears to be involved in apoptotic cell death in *D. melanogaster*.

The other pro- and anti-apoptotic proteins of *D. melanogaster* seem to play a more limited role. They have tissue-specific functions yet are able to functionally substitute the core proteins when the latter are mutated (Hay and Guo 2006).

13.3.1 Caspases

The *D. melanogaster* genome encodes seven caspases: three initiators (Dronc, Dredd and Strica; Fig. 13.4b) and four effectors (DrICE, Dcp-1, Decay and Damm) (Kumar and Doumanis 2000) (Fig. 13.4c).

Dronc is by far the best characterized of these proteins and is considered the main initiator caspase of *D. melanogaster* as it is required for almost all developmental and stress-induced apoptotic events (Dorstyn et al. 1999a; Quinn et al. 2000; Xu et al. 2005; Cooper et al. 2009b).

Dronc is a functional homolog of mammalian caspase-9 and nematode caspase CED-3, and, like these two proteins, it contains a CARD domain that enables binding to the adaptor protein Ark through CARD/CARD domain interaction (Dorstyn et al. 1999a). The association of Dronc and Ark proteins allows the formation of the apoptosome, a protein complex which is the spearhead of apoptosis (Pang et al. 2015). Like its mammalian homolog Apaf-1, *D. melanogaster* Ark contains, in addition to the CARD domain, a NB-ARC domain and WD40 repeats that provide the structural backbone of the apoptosome (Fig. 13.4a). In mammals, the WD40 repeats are autoinhibitory and block the interaction of the CARD domains of Apaf-1 and caspase-9 (Hu et al. 1999; Benedict et al. 2000). For the apoptosome to form, cytochrome c, released from the mitochondria following an apoptotic stimulus, needs to bind to the WD40 repeats. This interaction releases the CARD of Apaf-1 and allows interaction with the CARD of caspase-9. In the fruit fly, these repeats, while necessary for apoptosome stabilization, seem to have a completely different role. Indeed, while Ark can be immunoprecipitated in association with cytochrome c (Rodriguez et al. 1999), this protein is not required for the formation of the apoptosome complex in most tissues, with the exception of the retina and during spermatid individualization (Arama et al. 2003; Dorstyn et al. 2004; Mendes et al. 2006). In fact, in most *D. melanogaster* cell types, apoptosome formation and Dronc activation occur constitutively, even in the absence of apoptotic stimuli.

The interaction with Ark strongly promotes and stabilizes Dronc dimerization and catalytic activity by allowing Dronc auto-processing (Dorstyn and Kumar 2008; Pang et al. 2015). Dronc has a unique substrate specificity among caspases as it not only cleaves after an aspartate residue, like other caspases, but also after glutamate residues. This unusual substrate profile may be due to the unique amino acid sequence

P-F-C-R-G in its catalytic site (Hawkins et al. 2000). It allows Dronc auto-processing in an Ark-dependent manner, by cleaving at a glutamate residue between the large and small subunits as well as at an aspartate residue following the CARD domain. Activated Dronc cleaves and activates the effector caspase DrICE that is responsible for most apoptotic phenotypes in *D. melanogaster*.

The two remaining initiator caspases, Dredd and Strica, also contain a long N-terminal prodomain but, unlike Dronc, have minor roles in apoptosis. Exactly like caspase-8 of mammals, Dredd contains 2 DED in its N-terminal prodomain and is able to interact with the caspase adaptor, dFADD (*Drosophila* **F**as-**a**ssociated **d**eath **d**omain-containing protein) (Chen et al. 1998; Hu and Yang 2000). However, there is no clear extrinsic apoptotic pathway in *D. melanogaster* and, while *dredd* mutations are able to dominantly suppress IAP antagonist-induced cell death (Chen et al. 1998), the main function of Dredd is in the innate immune response following infection by Gram-negative bacteria (Meinander et al. 2012). Dredd is required for the cleavage and activation of Relish, one of the NF-kB family members, which is essential for the expression of antimicrobial peptides and for the activation of IKK and JNK (components of the *Drosophila* antibacterial immunity pathway) (Stoven et al. 2003). Strica possesses a unique serine and threonine-rich prodomain of as yet unknown function (Doumanis et al. 2001). It is able to induce apoptosis when expressed ectopically but in natural condition it has mainly redundant roles with Dronc (defects during mid- and late oogenesis appear only in *dronc/strica* double mutants but not in either single mutant) and tissue-specific functions (Leulier et al. 2006b; Baum et al. 2007). In particular, Strica is involved in larval salivary gland development as *strica* RNAi results in delayed elimination of cells specified to die in this tissue (Leulier et al. 2006b).

DrICE, the closest homolog to mammalian caspase-3, is the most important effector caspase of *D. melanogaster* (Muro et al. 2006). Its known targets include lamins, DIAP1, Dronc, the baculovirus caspase inhibitor p35 and ICAD (**i**nhibitor of **c**aspase-activated **D**Nase) (Kumar and Doumanis 2000). DrICE is involved in most developmental and stress-related apoptotic events (Muro et al. 2006). It is for instance involved in salivary gland remodeling and in programmed cell death during oogenesis (Leulier et al. 2006b). However, cell death is decreased, but not completely abolished, in *drICE* mutant embryos, larval and pupal tissues, possibly because of redundant functions with the caspase Dcp-1.

Dcp-1 and Decay are the other caspase-3-like proteins of *D. melanogaster* (Song et al. 1997; Dorstyn et al. 1999b). Phylogenetic analysis actually revealed that DrICE and Dcp-1 arose by duplication of an ancestral effector caspase after the divergence of mosquitoes and flies (Bryant et al. 2010). *dcp-1* and *decay* mutants are viable and fertile, without apparent adult defects or obvious morphological abnormalities (Kumar and Doumanis 2000). However, *dcp-1* and *drICE* double mutant embryos show further reduction of cell death compared to *drICE* single mutant embryos and, while the single mutants are homozygous viable, *dcp-1/drICE* double mutants are pupal lethal, similar to *dronc* mutants (Leulier et al. 2006b; Xu et al. 2006). In the case of Decay, possible redundancy with Strica has been hypothesized, as RNAi assays

show that Decay and Strica might act redundantly downstream of Hid-induced cell death (Leulier et al. 2006b).

Damm is the last effector caspase of *D. melanogaster* (Harvey et al. 2001). It can induce cell death in vitro and upon ectopic expression in the eye. However, no mutant is available, leaving its actual function undetermined. This is further complicated by its relationship to Strica. According to the available phylogenetic trees, Damm and Strica are more closely related to each other than any of the other caspases in *D. melanogaster*. Whether Damm derives from a gene duplication event, and whether it is an effector or initiator caspase, is thus unclear (Bryant et al. 2008).

Because Dronc is constitutively activated, and cleavage of DrICE by activated Dronc is sufficient to induce apoptosis, all *D. melanogaster* cells would be expected to undergo apoptosis. The only reason why they do not is the simultaneous constitutive expression of IAPs that inhibit caspase activity.

13.3.2 IAPs

The *D. melanogaster* genome encodes four IAPs: DIAP1, DIAP2, dBruce and Deterin (Orme and Meier 2009) (Fig. 13.4e).

DIAP1 is the main apoptosis inhibitor of *D. melanogaster*. Inhibition of DIAP1, either by IAP antagonists or by knocking down or out the gene, is sufficient to initiate apoptosis in virtually all cells (Yoo et al. 2002; Chai et al. 2003; Orme and Meier 2009). The presence of two BIR domains in DIAP1 enables interaction with several caspases among Dronc, DrICE and Dcp-1. For instance, the BIR2 domain of DIAP1 can interact with both monomeric and complexed Dronc by binding to a twelve residue motif located between the prodomain and the large catalytic subunit (Chai et al. 2003). Binding alone of DIAP1 to target caspases is not sufficient for total inhibition. This requires their ubiquitination, mediated by the DIAP1 RING domain (Tenev et al. 2005; Orme and Meier 2009). In the case of Dronc, this ubiquitination inhibits caspase processing and ultimately leads to its degradation. Interestingly, for complexed Dronc, Dronc-mediated cleavage of Ark is required for this degradation to take place (Shapiro et al. 2008). This indicates that Dronc and Ark are able to suppress their mutual levels in vivo and that DIAP1 promotes this mutual suppression. This may be a mechanism to quickly remove any inappropriately processed Dronc caspase or to keep Dronc activity low in non-apoptotic cells when high caspase activity is not desired.

DrICE and Dcp-1 binding involves the BIR1 domain of DIAP1, which binds to a motif that is revealed following cleavage of the caspases (Tenev et al. 2005). This means that, contrary to what happens with Dronc, DIAP1 can only bind to processed effector caspases. Once the interaction is established, DIAP1 blocks enzymatic activity through steric interference with binding of substrate. Nevertheless, for this inhibition to be completely effective, DIAP1 still needs to ubiquitinate the caspases, and this ubiquitination is dependent on DrICE. Indeed, the RING domain of DIAP1, while necessary, is not sufficient to inhibit effector caspases. DrICE also needs to

cleave DIAP1, removing 20 N-terminal amino acids. This results in a conformational change that relieves auto-inhibition of DIAP1 and exposes a new docking site for UBR-containing E3 ligases of the N-end rule pathway (Ditzel et al. 2008). Recruitment of these proteins allows for ubiquitination of DrICE.

The principal function of DIAP2 is connected to NF-kB-related innate immune responses, but it also plays a minor role in apoptosis (Leulier et al. 2006a). *diap2* mutants fail to show any detectable phenotype in developmental apoptosis, but are very sensitive to mild stress (Ribeiro et al. 2007). DIAP2 is able to rescue apoptosis triggered by RNAi-mediated depletion of DIAP1 (Leulier et al. 2006a) indicating that there is a level of functional redundancy between the two proteins. The anti-apoptotic function of DIAP2 relies on the specific binding of its BIR3 domain to DrICE. Following this, DrICE and DIAP2 are trapped in a covalent association, which allows proximal ubiquitination of the associated DrICE by the DIAP2 RING domain (Ribeiro et al. 2007). This in turn leads to degradation of DrICE.

Despite its exceptionally long size (five splicing variants, leading to proteins with lengths ranging from 4852 to 4976 amino acids), the dBruce protein only contains two domains: one BIR domain and one UBC domain that confer E2 ubiquitin conjugating enzyme activity (Domingues and Ryoo 2012). dBruce has not been reported to bind caspases and does not block the activity of Dronc. It does, however, bind to IAP antagonists and specifically regulates Reaper and Grim-induced cell death (see Sect. 2.3) (Domingues and Ryoo, 2012). It does not promote their degradation directly, but its anti-apoptotic action does require their IBM motifs to be intact. Moreover, the BIR domain is not sufficient for this activity, and the E2 domain is likely required.

Finally, the small IAP Deterin, the homolog of mammalian Survivin, contains a single N-terminal BIR repeat (Jones et al. 2000). Survivin-like proteins are best known for their ability to control cytokinesis and regulate chromosome segregation during mitosis (Silke and Vaux 2001). Nevertheless, Deterin has been reported to suppress apoptosis when overexpressed and to block Reaper-dependent cell death (Jones et al. 2000).

As Dronc is constitutively active at a basal level, survival of the cells not programmed to die is enabled by the constant anti-apoptotic effect of DIAP1. Full activation of Dronc thus depends on the removal of the inhibitory effects of DIAP1. This can be achieved either by increasing the intracellular levels of DIAP1 antagonists, or by increasing the concentration of Dronc and Ark. Both scenarios are seen during developmental cell death in *D. melanogaster*. Ecdysone signaling leads to increased caspase expression that makes cells sensitive to Reaper and Hid-induced apoptosis (Dorstyn et al. 1999a; Kilpatrick et al. 2005). In addition, in cells programmed to die, binding of the IAP antagonist proteins to DIAP1 changes the substrate specificity of the RING domain, leading to auto-ubiquitination and degradation of DIAP1 by the proteasome (Yoo et al. 2002; Chai et al. 2003; Orme and Meier 2009) (Fig. 13.5).

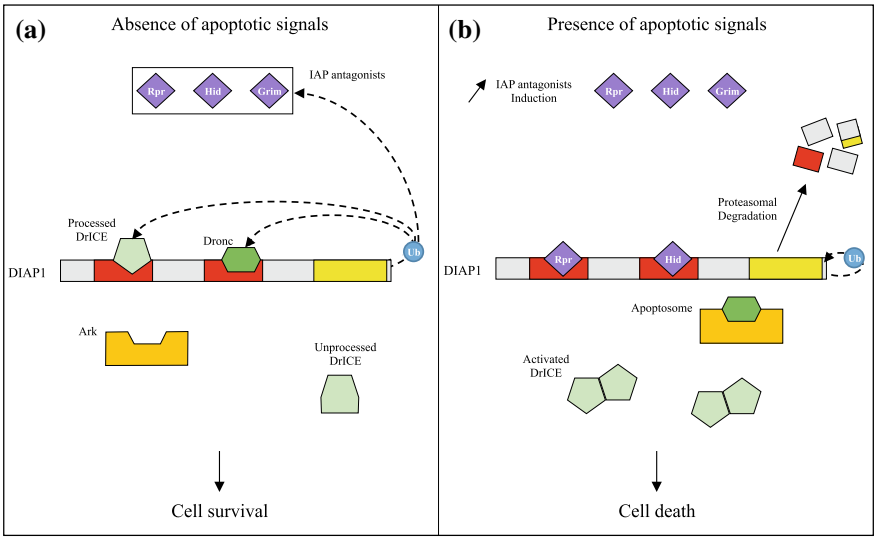


Fig. 13.5 Regulation of the caspase cascade by IAPs in *Drosophila melanogaster*. a In the absence of apoptotic signals, the caspase activating cascade is kept in check by a direct interaction of the BIR domains of DIAP1 with the caspases DrICE (must be processed prior to binding) and Dronc (DIAP1 can bind both unprocessed Dronc and processed Dronc-Ark complex). This promotes the ubiquitination of the caspases by the RING domain of DIAP1 and their subsequent inactivation. The RING domain of DIAP1 is also capable of ubiquitinating IAP antagonists. **b** In the presence of apoptotic signals, the IAP antagonists displace the caspases from DIAP1 and stimulate auto-ubiquitination and proteasomal degradation of DIAP1. Dronc is free to complex with Ark for apoptosome formation leading to activation of DrICE for cell death induction

13.3.3 IAP Antagonists

Six IAP antagonists have been identified in *D. melanogaster*: Reaper, Hid, Grim, Sickle, Jafrac2 and HtrA2 (also known as dOmi) (Chen et al. 1996; Xu et al. 2009; Vasudevan and Ryoo 2015) (Fig. 13.4f).

Apart from their IBM motif, these proteins are poorly conserved, and they are not interchangeable. The slight differences between their respective IBM motifs allow selective binding to specific BIR domains and dictate their specific cell killing abilities (Wing et al. 1998; Zachariou et al. 2003). For example, while Reaper and Grim can bind to both BIR1 and BIR2 of DIAP1 with equal affinity, Hid, Sickle, HtrA2 and Jafrac2 preferentially bind the BIR2 domain. Thus, since distinct caspases bind to different BIR domains, different IAP antagonists can differentially displace distinct sets of caspases. Even when they bind to the same BIR domains, IAP antagonists are not functionally identical, and they are involved in different apoptotic processes. For instance, Grim is more effective at inducing cell death in the embryonic central

nervous system (Chen et al. 1996), Hid is essential for cell death during eye development (Yu et al. 2002), and apoptosis in post-embryonic neuroblasts specifically requires the function of Reaper (Wing et al. 1998).

DIAP1, as the main regulator of apoptosis in *D. melanogaster*, is the preferential target of IAP antagonists, but DIAP2 and dBruce (but not Deterin) can also be recognized (Vasudevan and Ryoo 2015).

The essential nature of these proteins for apoptosis is supported by the fact that virtually all apoptosis, even developmental, is abolished in the absence of these genes; whereas, their overexpression is sufficient to kill cells (White et al. 1994; Vasudevan and Ryoo 2015). Their activation is finely regulated through a complex combination of transcriptional and/or posttranscriptional mechanisms (Xu et al. 2009; Vasudevan and Ryoo 2015), which are, for the most part, conserved over very long evolutionary distances. For example, Reaper expression is restricted to cells fated to die during embryonic development and absent from cells fated to survive (White et al. 1994).

The best documented transcriptional regulatory mechanism of IAP antagonists is their induction by ecdysone during development and metamorphosis. Expression of *grim*, *reaper* and *hid* is induced by the ecdysone receptor (EcR), which binds to a *reaper*-upstream enhancer sequence (Jiang et al. 2000). The tumor suppressor protein p53 can also directly bind to a regulatory sequence that lies between *reaper* and *sickle* to induce their expression (Brodsky et al. 2000) and the EGF receptor/MAP kinase survival signaling pathway regulates *hid* transcription (Kurada and White 1998). Moreover, IAP antagonist expression triggers a transcriptional feedback loop that induces other IAP antagonists to increase their pro-apoptotic effects, and this is mediated by JNK signaling and p53 (Shlevkov and Morata 2012). At the post-transcriptional level, several miRNAs interacting with IAP antagonists have been identified: *bantam* targets Hid for translational suppression, *miR-2* targets sites in Reaper, Grim and Sickle transcripts, *miR-14* targets Reaper and *miR-6* and *miR-11* are regulators of all four IAP antagonists (Vasudevan and Ryoo 2015).

There are two main differences between mammalian and *D. melanogaster* IAP antagonists that might explain why these proteins have a greater role in the latter.

First of all, contrary to the two mammalian proteins Smac and HtrA2, *D. melanogaster* IAP antagonists have no N-terminal mitochondrial localization sequence, and they do not enter the mitochondrial intermembrane space. The fact that Smac and HtrA2 are sequestered into the mitochondrial intermembrane space makes them incapable of inhibiting IAPs that reside in the cytoplasm in non-apoptotic conditions (Du et al. 2000; Verhagen and Vaux 2002). On the other hand, *D. melanogaster* IAP antagonists are anchored to the mitochondrial outer membrane, a localization essential for their pro-apoptotic function, with their IAP-binding motif facing the cytoplasm, thus enabling direct interaction with their targets (Olson et al. 2003; Abdelwahid et al. 2007).

Secondly, IAP antagonists in *Drosophila* have potent autonomous apoptotic activity, independent of DIAP1, which has not been reported for mammalian IAP antagonists (Kornbluth and White 2005).

13.3.4 BCL-2 Family Members

BCL-2 signaling seems to play a limited role in apoptosis in *D. melanogaster* and, despite numerous studies, only two members of this family have been identified so far: Debcl (also named Drob-1/dBorg-1/dBok) and Buffy (dBorg-2) (Colussi et al. 2000; Zhang et al. 2000; Igaki and Miura 2004) (Fig. 13.4d). These two proteins show the highest similarity with the mammalian pro-apoptotic protein Bok/Mtd, even though Buffy lacks a BH4 motif (Zhang et al. 2000). Debcl, at least, seems to have conserved this pro-apoptotic function. Overexpression of Debcl induces cell death both in *D. melanogaster* and mammalian cells. Silencing of this gene in early stage embryos results in reduced cell death and excessive numbers of glial cells. The expression pattern of Debcl in the embryos, larval tissues and adult ovaries correlates with the distribution of TUNEL-positive dying cells (Colussi et al. 2000; Igaki and Miura 2004).

Debcl contains a C-terminal hydrophobic tail, which presumably confers mitochondrial localization and appears to be important for its pro-apoptotic activity (Igaki and Miura 2004). Interestingly, it was shown that Debcl induces cytochrome c release from mitochondria when expressed in mammalian cells (Zhang et al. 2000) but the fact that cytochrome c is not necessary for apoptosis in *D. melanogaster* suggests that Debcl pro-apoptotic function might depend on another mechanism. While this mechanism remains to be elucidated, a possibility, supported by the fact that Ark loss-of-function does not affect Debcl-induced cell death, is that Debcl might kill cells in a caspase-independent fashion by compromising mitochondrial function. Consistent with this, it was recently shown that mitochondrial oxidative stress, and the resulting impaired mitochondrial dynamics, takes an important place in Debcl-induced cell death (Colin et al. 2015; Clavier et al. 2016).

Results for Buffy are even more controversial. It was shown that its overexpression in S2 cells can induce cell death (Igaki and Miura 2004), even if less efficiently than Debcl, but another study demonstrated that Buffy acts as an anti-apoptotic factor in vivo (Quinn et al. 2003). The discovery that Debcl and Buffy can heterodimerize led to the hypothesis that Buffy might hold the Dronc/Ark complex in check, and full activation is achieved by Debcl-induced removal of Buffy (Quinn et al. 2003).

13.4 Apoptosis in Other Diptera

Diptera are among the most diverse order of insects with more than 150,000 known species of flies and mosquitoes. However, laboratory studies have been conducted almost exclusively on the model *D. melanogaster* or on disease vector mosquitoes, especially the yellow fever mosquito *Aedes aegypti* and the malaria vector *Anopheles gambiae*. This is also true for studies on apoptotic pathways.

If the core apoptotic pathway seems to be conserved in Diptera, several differences in the number of caspases and regulatory proteins have been reported, and studies

on the functions of these proteins are lacking. This is particularly surprising for mosquitoes as numerous reports suggest that cell death might play a role in the transmission of certain infections (Bartholomay et al. 2004; Girard et al. 2007; Li et al. 2007).

13.4.1 Species-Specific Expansion of Apoptotic Proteins in Diptera

Only the caspases Dronc, Dredd and Decay have clear orthologs across all the 11 Drosophilid genomes analyzed by Bryant and colleagues (Bryant et al. 2010). The other caspases have undergone species- and lineage-specific duplications. For instance, three and four Strica paralogs have been found in *Drosophila mojavensis* and *Drosophila willistoni*, respectively, and two Dcp-1 paralogs have been found in the obscura group (*Drosophila pseudooscura* and *Drosophila persimilis)*.

In the case of mosquitoes, sequencing of the *Ae. aegypti* and *An. gambiae* genomes allowed for the identification of 11 (AeDronc, AeDredd, AeCASPS7-8, AeCASPS15-21) and 13 caspases (AgCASPL1-2, AgCASPS1-9, AgCASPS13-14), respectively (Christophides et al. 2002; Waterhouse et al. 2007; Bryant et al. 2008) (Table 13.1). While many of the predicted caspases in *An. gambiae* have not been

Table 13.1 *Drosophila melanogaster* caspases and their homologs in the mosquitoes *Aedes aegypti* and *Anopheles gambiae*

Drosophila melanogaster caspases	*Aedes aegypti* homologs	*Anopheles gambiae* homologs
Dronc	AeDronc	AgCASPL2
Dredd	AeDredd	AgCASPL1
Strica/Damm	AeCASPS15	AgCASPS9
	AeCASPS16	AgCASPS13
	AeCASPS17	
	AeCASPS21	
DrICE/Dcp-1	AeCASPS7	AgCASPS7
	AeCASPS8	AgCASPS8
Decay	AeCASPS18	AgCASPS1
	AeCASPS19	AgCASPS2
		AgCASPS3
		AgCASPS4
		AgCASPS5
		AgCASPS6
		AgCASPS14
/	AeCASPS20	/

confirmed yet, this suggests overall that several duplications of caspase-like genes occurred in mosquitoes.

Mosquito effector caspases can be grouped into two distinct clades (Bryant et al. 2010). Clade I contains all the caspases that share sequence similarity with *D. melanogaster* Decay and include two *Ae. aegypti* (AeCASPS18 and AeCASPS19) and eight *An. gambiae* caspases (AgCASPS1-6, AgCASPS11 and AgCASPS14). Clade II contains caspases that share sequence similarity with *D. melanogaster* DrICE and Dcp-1 and includes two caspases from *Ae. aegypti* (AeCASPS7 and AeCASPS8) and two caspases from *An. gambiae* (AgCASPS7 and AgCASPS8). One additional caspase, *Ae. aegypti* CASPS20, shares sequence similarity with *D. melanogaster* DrICE but does not fall within either clade. The third clade includes the caspases that are closest phylogenetically to Damm and Strica. This clade includes caspase from *Ae. aegypti* with both short (AeCASPS15, 17, and 21) and long prodomains (AeCASPS16).

Out of all *Ae. aegypti* caspases, AeCASPS7 is actually the closest relative to DrICE and Dcp-1, while AeCASPS8 seems to be a closely related mosquito-specific clade that appears to have arisen by duplication of AeCASPS7 after the divergence between mosquitoes and fruit flies (Bryant et al. 2010; Liu and Clem 2011). Transcripts for AeCASPS7, 8 and 20 can be found at all developmental stages and in the adult midgut (Bryant et al. 2008).

The role and importance of these duplications have not been elucidated yet. Some of them may function as decoy caspases. Indeed, comparative sequence analysis revealed that several of the duplicated caspases display substitutions in critical amino acids and are therefore unlikely to be functional (Bryant et al. 2008). This concerns three caspases from *An. gambiae* (AgCASPS1, AgCASPS2 and AgCASPS14), which contain a serine or threonine instead of an alanine in the putative active site, and one from *Ae. aegypti* (AeCASPS18), which contains a serine in place of cysteine. AeCASPS18 functions as a decoy caspase and enhances the activity of AeCASPS19, both in cells and in vitro (Bryant et al. 2010).

These duplications do not seem to be limited to pro-apoptotic proteins as seven putative IAPs were identified in the *An. gambiae* genome (Waterhouse et al. 2007).

No expansion is seen for IAP antagonists of which only two (Michelob_x and IMP) have been identified in the *Ae. aegypti* genome so far (Zhou et al. 2005).

13.4.2 Conservation of the Function of Apoptotic Proteins in Diptera

The *Ae. aegypti* genome contains clear orthologs of *D. melanogaster* Dronc and Ark, and silencing one or both completely inhibits apoptosis triggered by several different apoptotic stimuli (Liu and Clem 2011). This confirms their prominent role in the apoptosis of this species.

Their regulation also seems to be similar, as *Aedronc* transcription is directly up-regulated in response to ecdysone (Cooper et al. 2007b). *Aedronc* transcripts are also expressed ubiquitously and, similar to *D. melanogaster dronc*, are found at highest levels in developmental stages experiencing pulses of ecdysone. The involvement of cytochrome c in mosquito apoptosis has not been fully investigated yet, but in vitro studies have shown that it is not required for caspase activation in Aag2 cells. Addition of purified cytochrome c does not induce caspase activation in the cytoplasmic extract, contrary to the addition of recombinant mosquito IAP antagonists, which activates caspases (Wang and Clem 2011).

Despite being very similar in size, AeDronc (449 amino acids) and Dronc (450 amino acids) sequences have limited sequence similarity and identity of 48% and 31%, respectively. Nevertheless, the most critical structures are conserved (Cooper et al. 2007b). AeDronc contains a CARD domain and a six-α-helix bundle in the prodomain, nearly identical to that found in *D. melanogaster* and similar to those found in human caspases-2 and -9. There is also limited conservation of the predicted DIAP1 binding site suggesting that regulation of AeDronc activation may occur through similar mechanisms.

The amino acid sequence surrounding the catalytic cysteine of AeDronc, S-I-C-R-G, is unique (Cooper et al. 2007b). The unusual sequence surrounding the catalytic cysteine in the *D. melanogaster* Dronc confers an unusual substrate profile, allowing the protein to promote auto-processing and strongly promoting its catalytic activity. By analogy, a similar mechanism might also be present in AeDronc even if this has not been confirmed yet. More recently, Accorsi et al. (2015) identified putative Dronc homologs in several other Dipterans, e.g., *Musca domestica* (Accession No. XP_005178429.1), *Ceratitis capitata* (XP_004519893.1) and *Culex quinquefasciatus* (XP_001844541.1). All these proteins present a different amino acid sequence in their active [PS]-[FI]-C-R-G sites, suggesting that this might be a specific characteristic of dipteran Dronc homologs.

Contrary to what happens with AeDronc, silencing of the effector caspases AeCASPS7 and AeCASPS8, which are the closest relatives to DrICE, only partially inhibits apoptosis. This indicates that more than one effector caspase is involved in apoptosis in *Ae. aegypti* (Liu and Clem 2011). It is possible that, like in *Drosophila*, redundancy between different effector caspases of *Ae. aegypti* rescues the phenotype induced by silencing of individual caspases.

AeIAP1 is highly similar to its *D. melanogaster* homolog, DIAP1. It can inhibit AeDronc, AeCASPS7 and AeCASPS8 and interact with the two IAP antagonists identified in the *Ae. aegypti* genome, Michelob_x and IMP (Wang and Clem 2011). AeIAP1 is also transcribed at all developmental stages of the mosquito, with higher levels at the pupal and adult stages as well as at specific time intervals when mosquitoes are subjected to various apoptotic stimuli (Puglise et al. 2016). Finally, silencing of the Aedes *iap1* gene causes spontaneous apoptosis in the *Ae. aegypti* cell line Aag2 and death in adult mosquitoes (Liu and Clem 2011, Wang and Clem 2011). This suggests that the function of DIAP1 as the main regulator of apoptosis is conserved in *Ae. aegypti*.

Overexpression of Michelob_x and IMP was sufficient to induce apoptosis in heterologous species (Wang and Clem 2011) suggesting that IAP antagonists in *Ae. aegypti* and *D. melanogaster* may play a similar role. In vivo data in mosquitoes are still necessary to confirm this and to determine if the mechanisms involved are similar to those used by IAP antagonists in *D. melanogaster*.

Apart from these proteins, the *Ae. aegypti* genome also encodes homologs of Dredd, Strica, DIAP2, Deterin and dBruce, but most of them have not been studied so far. AeDredd contains two death inducing domain (DID), interacts with a mosquito FADD adaptor and is required for IMD signaling and antibacterial immunity in *Ae. aegypti* (Cooper et al. 2007a, 2009a). The homolog of Strica, AeCASPS16, is predicted to contain a serine- and threonine-rich prodomain. Transcripts of AeCASPS16 can be detected in all developmental stages and in the adult midgut (Bryant et al. 2008). The function of their serine–threonine-rich prodomains remains unknown but they could represent novel functions, protein–protein interactions or cell death signaling pathways within insects.

A few pro- and anti-apoptotic proteins have also been identified in other mosquito species. For example, there have been reports of *iap1*-like genes in *Aedes triseriatus*, *Aedes albopictus*, *Culex pipiens* and *Culex tarsalis* (Blitvich et al. 2002; Beck et al. 2007; Li et al. 2007). They all encode proteins of similar sizes (401–410 amino acids) and structure with highly conserved domains (their BIR1 and BIR2 domains are 82 and 92% identical, respectively, the 36-amino acid RING finger of each mosquito IAP1 is identical). Orthologs of Michelob_x, with an invariable N-terminal IBM were found by BLAST in both *An. gambiae* and *Ae. albopictus* (Zhou et al. 2005; Bryant et al. 2008), and the AaCASPS16 of *Ae. albopictus* was predicted to contain a serine- and threonine-rich prodomain (Meng et al. 2016).

13.5 Apoptosis in Lepidoptera

Lepidoptera (butterflies and moths) is the only other insect order where apoptotic pathways have been studied in some detail. Studies of apoptosis in Lepidoptera began in the sixties, at the same time as in *Drosophila*, with the description of intersegmental muscle degeneration during metamorphosis in silkmoths (Lockshin and Williams 1965). However, since then, apoptosis research in lepidopterans has lagged far behind that of other organisms. The first lepidopteran apoptotic protein, (Sf-caspase-1), was identified in the fall armyworm *Spodoptera frudiperda* (Ahmad et al. 1997).

More recently, the genome sequencing of the domestic silkworm *Bombyx mori,* a representative of Lepidoptera with important economic value, has enabled the identification of several putative pro- and anti-apoptotic genes (Tambunan et al. 1998; Huang et al. 2001; Pei et al. 2002; Bryant et al. 2009; Zhang et al. 2010). Since then, an extensive survey of lepidopteran-derived EST datasets has partially

extended our knowledge, at least for caspases (Courtiade et al. 2011). However, these studies are for the most part descriptive, and there is a serious lack of information about the regulation of apoptosis in this insect order.

13.5.1 Identification of Apoptosis Proteins in Lepidopterans

Contrary to what has been observed in mosquitoes, no extensive duplication of caspases seems to have taken place in the order Lepidoptera.

Zhang and colleagues (Zhang et al. 2010) conducted an extensive study of pro- and anti-apoptotic genes in the silkworm genome, and they identified and cloned five caspases, four IAPs, two IAP antagonists and two Bcl-2 family members, showing a global conservation of the apoptotic machinery present in *D. melanogaster*. Consistent with this, Shu and colleagues recently conducted a similar study in *S. frugiperda*, using the transcriptome of Sf9 cells, and found results similar to what was described in *B. mori*: They identified four caspases, four IAPs, two IAP antagonists and one Bcl-2 family member (Shu et al. 2017).

The five *B. mori* caspases are named BmCaspase-1 (Pei et al. 2002), Bm-ICE (Yi et al. 2014), BmDredd, BmDronc (Suganuma et al. 2011) and BmCaspase-N. Based on the length of their prodomains, BmCaspase-1 and BmICE (short prodomains) are supposed to be effector caspases; whereas, the other three proteins identified are likely to be initiators as they have long prodomains. The four IAPs identified in the silkworm genome, named BmIAP1 (Huang et al. 2001), BmIAP2, BmBruce and BmSurvivin are homologs of DIAP1, DIAP2, dBruce and Deterin, respectively, indicating a one-to-one orthology with *D. melanogaster*. Exactly like their homologs in *D. melanogaster,* BmIAP has two BIR and one RING domain, while BmIAP2 has three BIR and one RING domain. BmBruce also has an ubiquitin-proteasome binding motif. The two IAP antagonists are homologs of Reaper and HtrA2 and were thus named BmReaper (Bryant et al. 2009) and BmHtra2. Finally, the two Bcl-2 homologs identified in the genome of *B. mori* are BmP109 (Tambunan et al. 1998) and BmBuffy. BmP109 contains the four conserved Bcl-2 regions, but BmBuffy lacks the BH4 motif. Courtiade and colleagues (Courtiade et al. 2011) made a more comprehensive survey of caspase proteins in Lepidoptera based on available EST datasets. Apart from some phylogenetic discrepancies, they globally confirmed the data from Zhang et al. (Zhang et al. 2010). They identified 66 putative caspase sequences in 27 species of lepidopterans.

Comparative phylogenetic analyses showed that these caspases can be grouped into six distinct subfamilies, Lep-caspase-1 to Lep-caspase-6, with similar structures and domain composition (Courtiade et al. 2011). Caspases from the Lep-caspase-5 and -6 groups have long prodomains and are homologs of initiator caspases Dronc and Dredd, respectively. These two groups include BmDronc and BmDredd.

Members of the Lep-caspase-1, -2 and -3 subfamilies have short prodomains. Lep-caspase-1 (including BmCaspase-1) and Lep-caspase-2 proteins are homologs of *D. melanogaster* DrICE and Dcp-1. Once again, these two proteins seem more related

than any of the lepidopteran caspases, confirming that they are the result of a dipteran-specific duplication event. Interestingly, Lep-caspase-2 appears to be noctuid-specific and to have arisen from a tandem duplication of the caspase-1 gene (Courtiade et al. 2011). It is possible that Lep-caspase-1 and -2 have redundant functions or that Lep-caspase-2 proteins have evolved new functions related to the noctuid lifestyle. However, this remains to be tested experimentally. Homology relationships of Lep-caspase-3 with effector caspases from *D. melanogaste*r were less clear, and no clear ortholog of Decay was identified in this study. This group includes BmICE.

The function of Lep-caspase-4 (including BmCaspase-N) remains unclear as it is closely related to both Damm and Strica, and its long prodomain does not seem to harbor any known death domain. Classification of Lep-caspase-4 is further complicated by the peculiar size of its catalytic domain small subunit, which is twice as long as any other small subunit described so far. No function has been proposed for this longer subunit but alignments show that, if the sequence corresponding to the last 140 residues is more variable compared to the rest of the subunit, some residues are conserved in all sequences, including three cysteine residues potentially involved in the formation of secondary structures (Courtiade et al. 2011).

This study identified at most one representative per caspase subfamily for each species, with the exception of the *Manduca sexta genome*, which contains two Lep-caspase-4. It is, however, important to point out that this could be due to methodological issues as EST datasets are not necessarily representative of the complete transcriptome of a given species. This could also explain why, for some species, caspases from only some subfamilies were found. For example, caspases from the Lep-caspase-5 subfamily were only found in three species (*B. mori*, *Helicoverpa armigera*, *Pieris rapae*), and for other species only one putative caspase was identified.

13.5.2 Regulation of Apoptosis in Lepidoptera

Only a few proteins related to apoptosis have been functionally characterized in Lepidoptera. These studies have shown that, like in mosquitoes, the core apoptosis pathway of *D. melanogaster* seems to be conserved in lepidopterans. For instance, BmReaper has been shown to bind to BmIAP1 via its IBM and induce apoptosis in both dipteran and lepidopteran cell lines (Bryant et al. 2009; Zhang et al. 2010). Its expression is also significantly up-regulated during metamorphosis suggesting a possibly conserved regulation compared to *D. melanogaster*. In *S. frugiperda*, RNAi of Sf-IAP-1 induces spontaneous apoptosis (Liu and Clem 2011), and silencing of SfDronc expression reduces the cleavage of Sf-caspase-1 and inhibits apoptosis in Sf9 cells (Huang et al. 2013). The caspase Hearm-1 of *H. armigera* is also inducible by ecdysone (Yang et al. 2008). However, these studies have also highlighted several differences in the regulation of apoptosis between Diptera and Lepidoptera.

Kumarswamya and colleagues recently used Sf9 cells to demonstrate that, contrary to *D. melanogaster*, cytosolic cytochrome c release is an essential event for

caspase activation in lepidopteran apoptosis, much like in mammals (Kumarswamy et al. 2009). If this is confirmed in vivo, this might suggest that lepidopteran Dronc homologs are not constitutively activated. Consistent with this, the silkmoth caspase BmDronc has the classical Q-A-C-R-G sequence surrounding the catalytic cysteine (Suganuma et al. 2011) suggesting that it does not possess the unusual substrate profile of *D. melanogaster* and other dipteran Dronc homologs. It should, therefore, be unable to promote its own auto-processing.

Another major difference is that, when expressed in mammalian cells, BmIAP1 is a specific inhibitor of mammalian caspase-9 (Dronc homolog), but not of the downstream effectors caspase-3 and caspase-7 (DrICE homologs). This suggests that BmIAP1 might not be a direct inhibitor of *B. mori* effector caspases, contrary to what happens in *D. melanogaster* with IAP1.

More recently, Yang and colleagues also used Sf9 cells to identify and characterize the caspase SfDredd (Yang et al. 2016). They showed that, despite being highly similar to other insect Dredd homologs and being able to function as an initiator caspase, purified SfDredd shows the strongest activity toward effector caspase substrates. To date, this caspase is the only caspase other than the human caspase-2, and the first insect caspase, that is homologous to initiator caspases while exhibiting effector caspase substrate specificity.

Interestingly, Zhang and colleagues (Zhang et al. 2010) have identified two splice variants of BmDronc with common translation initiation and termination sites. These two variants, named BmDroncL (415 amino acids) and BmDroncS (183 amino acids), may have different roles in apoptosis since BmDroncS only contains the small subunit of the catalytic domain of caspases; whereas, BmDroncL contains all the domains characteristic of these proteins. A similar observation has been made in humans, where a splice variant of caspase-9 lacks the catalytic site commonly found in other caspases. This isoform has been shown to be a dominant-negative inhibitor of apoptosis by blocking the caspase-9/Apaf-1 interaction (Seol and Billiar 1999). A similar regulation could be at play in *B. mori.* To our knowledge, no Dronc splice variants have been reported in *D. melanogaster*.

Zhang and colleagues (Zhang et al. 2010) also identified three splice variants for BmICE, but, despite their differences in exon numbers, BmICE (seven exons, 275 amino acids), BmICE-2 (eight exons, 284 amino acids) and BmICE-5 (nine exons, 312 amino acids) all have the characteristic Q-A-C-R-G sequence and the typical large/small subunit configuration of caspase family members. No hypothesis concerning their function(s) has been formulated yet.

Caspase regulation in Lepidoptera seems to happen also at the posttranscriptional level, as is the case for Sl-caspase-1, a putative effector caspase from *Spodoptera littoralis* (involved in apoptosis induced by UV irradiation and baculovirus infection) (Liu et al. 2005). This suggests that Lepidoptera may have an additional regulatory checkpoint in apoptotic pathway regulation compared to *D. melanogaster*.

13.6 Other Insect Orders

As illustrated in the previous paragraphs, the regulation of *D. melanogaster,* and more generally dipteran, apoptosis pathways, is not necessarily representative of what happens in other insect orders. Overall, insects are among the most diverse and speciose organisms on earth, and they display vast differences in terms of habitats, shapes, sizes and interactions with other organisms. These differences in terms of lifestyles appear to be accompanied by differences in their apoptotic pathways. To better understand how apoptosis pathways in insects have diversified and what the functional significance is thereof, it is of utmost importance to increase our efforts and expand apoptotic studies to other insects.

Several projects, such as the i5 k initiative (sequencing five thousand arthropod genomes, http://i5k.github.io/), aim to increase the number of sequenced insect genomes to facilitate specific and comparative research. However, these new genomes are made available much quicker than they are being exploited. This is supported by the fact that, while more than 350 insect genomes (corresponding to about 16 insect orders), are now available, the only species in which systematic research of apoptotic genes and their functions has been conducted are the ones cited above in this review. These few species belong to only two distinct insect orders, namely Diptera and Lepidoptera, and represent but a small part of the incredible diversity of insects (1 million of species currently described, distributed in 32 orders) (Misof et al. 2014, Stork 2018). Thus, while the information obtained on these insect apoptosis pathways is useful, it seems unlikely that it can be applied to all insects.

Diptera and Lepidoptera, for instance, share a holometabolous lifecycle. The complete metamorphosis that these insects undergo is accompanied by a rapid and drastic remodeling in which apoptosis plays a key role. Several authors have already tentatively linked this characteristic to the particular regulation of apoptosis in *D. melanogaster* (Kornbluth and White 2005). It has been proposed that the constitutive activation of caspases, without the need for cytochrome c, and their control by a single inhibitor could reflect the need to induce apoptosis in a rapid manner during the development of holometabolous insects. Our extensive analysis of available literature does not confirm this interpretation. In Lepidoptera, cytochrome c seems to be an essential event for caspase activation, which is an important difference with caspase activation in *D. melanogaster*. Differences with *D. melanogaster* are expected to be even more pronounced in hemimetabolous insects, such as those that constitute the order Hemiptera, as they do not undergo complete metamorphosis or major remodeling of the body during development (Sehnal et al. 1996). Thus, although apoptosis certainly occupies an important place in development and survival, it is possible that their apoptotic pathways and the way they are regulated differ from what was described in *Drosophila*.

Recent discoveries in the pea aphid, *Acyrthosiphon pisum,* are consistent with this hypothesis. *A. pisum* belongs to the order Hemiptera and, like other insects of this order, has evolved long-lasting interactions with symbiotic bacteria (The International Aphid Genomics Consortium 2010, Hansen and Moran 2014; Tagu et al. 2016).

The *A. pisum* genome encodes homologs for only three caspases of *D. melanogaster* (Dronc, DrICE and Decay). However, the pea aphid genome contains two paralogs for each of these caspases, even for Dronc (Simonet et al. 2018). This is worth noting as all Diptera and Lepidoptera contain only a single Dronc ortholog. This includes mosquitoes, for which an important expansion of caspases has been reported (Bryant et al. 2008). The two *A. pisum* Dronc homologs both contain a CARD and a catalytic CASc domain and are 57 and 75% identical at the amino acid and nucleotide level, respectively. They both have the conserved catalytic cysteine but, while one protein (ACYPI43249) possesses the classic Q-A-C-R-G motif found in most caspases, the second one contains a unique D-A-C-R-G motif (unpublished observations). This replacement of a neutral amino acid with a negatively charged amino acid could have important consequences for the specificity of this protein and its function. Similar observations can be made for IAPs. The *A. pisum* genome has no homolog of dBruce but encodes five orthologs of DIAP1/DIAP2 and several orthologs of Deterin. This expansion, and the caspase expansion in mosquitoes, suggests that duplication of apoptotic genes might be common in insects. We confirmed this by searching the newly sequenced genomes of several insects for those genes. We showed that caspase numbers vary wildly from one insect species to another, suggesting multiple gene duplication and pseudogenization events throughout evolution (Table 13.2).

In a study of the physiological regulation of cell death in *A. pisum*, we discovered that bacteriocytes, the cells housing symbiotic bacteria in insects, are eliminated through an unknown process, distinct from apoptosis- or autophagy-dependent cell death (Simonet et al. 2018). Bacteriocyte cell death is a dynamic process, resulting from a progressive ER-derived hypervacuolation of the cell, triggering cell stress responses, such as swelling of mitochondria, late induction of autophagy, reactive oxygen species (ROS) and the digestion of symbiotic bacteria by the lysosomal system. Even though a paralog for each aphid caspase group is induced in bacteriocyte cell death, the simultaneous induction of several IAPs among which four orthologs of DIAP1/DIAP2 may explain the absence of any apoptotic phenotype in these cells. This co-induction of IAPs and caspases has, to our knowledge, not been described in any other species. Interestingly, this cell death process is different from the process of bacteriocyte elimination described in the holometabolous weevil *Sitophilus orizae*, (Vigneron et al. 2014) where bacteriocytes are rapidly eliminated following caspase activation, in a process resembling apoptosis. Although different forms of non-apoptotic and non-autophagic cell death processes (e.g. paraptosis, necroptosis, oncosis, …)(Galluzzi et al. 2018) have recently been described, aphid bacteriocyte cell death is the first one discovered in insects.

This discovery reaffirms the necessity to extend the study of insect apoptosis to other insect models to better understand how they live, develop, feed and reproduce.

Table 13.2 Number of predicted caspase encoding genes in insects from different orders. For each of these organisms, the protein sequences of *Drosophila melanogaster* caspases were used as a query to perform BLAST searches against the reference assembly publicly available on NCBI. A domain search was subsequently conducted using the InterProscan software and proteins were considered putative caspases if they possessed a complete CASc domain, characteristic of this class of proteins

Order	Species (taxonomic name)	Species (common name)	Assembly	Number of predicted caspase encoding genes
Blattodea	*Blatella germanica*	German cockroach	GCA_003018175.1	12
	Zootermopsis nevadensis	Dampwood termite	GCA_000696155.1	7
Coleoptera	*Tribolium castaneum*	Red flour beetle	GCA_000002335.3	9
Diptera	*Aedes aegypti*[*]	Yellow fever mosquito	GCA_002204515.1	11[a]
	Anopheles gambiae[*]	African malaria mosquito	GCA_000005575.1	13[b]
	Zeugodacus cucurbitae	Melon fly	GCA_000806345.1	9
	Glossina morsitans	Common tsetse fly	GCA_001077435.1	6
	Drosophila melanogaster[*]	Fruit fly	GCA_000001215.4	7[c]
	Musca domestica	House fly	GCA_000371365.1	8
Hemiptera	*Acyrthosiphon pisum*[*]	Pea aphid	GCA_000142985.2	6[d]
	Bemisia tabaci[*]	Sweet potato whitefly	GCA_001854935.1	3[e]
	Cimex lectularius	Common bed bug	GCA_000648675.3	4
	Diaphorina citri	Asian citrus psyllid	GCA_000475195.1	6
	Nilaparvata lugens[*]	Brown planthopper	GCA_000757685.1	5[f]
	Rhodnius prolixus	Kissing bug	GCA_000181055.3	4
Hymenoptera	*Acromyrmex echinatior*	Panamanian leafcutter ant	GCA_000204515.1	5
	Apis mellifera	European honeybee	GCA_003254395.2	5
	Bombus terrestris	Buff-tailed bumblebee	GCA_000214255.1	5
	Camponotus floridanus	Florida carpenter ant	GCA_003227725.1	5

(continued)

Table 13.2 (continued)

Order	Species (taxonomic name)	Species (common name)	Assembly	Number of predicted caspase encoding genes
	Nasonia vitripennis	Jewel wasp	GCA_000002325.2	9
Lepidoptera	*Bombyx mori*[*]	Domestic silkworm	GCA_000151625.1	5 [g]
	Spodoptera litura	Cotton leafworm	GCA_002706865.1	7
	Manduca sexta	Carolina sphinx moth	GCA_000262585.1	7
Orthoptera	*Locusta migratoria*	Migratory locust	GCA_000516895.1	10
Phthiraptera	*Pediculus humanus*	Human body louse	GCA_000006295.1	4
Siphonaptera	*Ctenocephalides felis*	Cat flea	GCA_003426905.1	9
Thysanoptera	*Frankliniella occidentalis*	Western flower thrips	GCA_000697945.3	11

[*] indicates organisms for which the total number of caspases was previously known
[a] Bryant et al. 2008; [b] Waterhouse et al. 2007; [c] Kumar and Doumanis 2000; [d] Simonet et al. 2018; [e] Wang et al. 2018; [f] Huang et al. 2015; [g] Zhang et al. 2010

13.7 Conclusion

Insects are the most speciose class of the Animal Kingdom, are evolutionary ancient and cover a very broad evolutionary distance. Insects display an enormously rich variation of body plans with different shapes and sizes and distinct lifestyles occupying widely divergent habitats. Paralleling this diversity are changes in development and immunity. Given the important role of apoptosis, we surmise that the apoptotic pathway has undergone extensive modifications throughout evolution to support these changes in development and immunity. Finally, we propose that comparative studies of insects will not only uncover the extent of the modifications in the apoptotic pathway, but also reveal how this contributed to the marvelous wealth of insects we can observe today.

References

Abdelwahid E, Yokokura T, Krieser RJ, Balasundaram S, Fowle WH, White K (2007) Mitochondrial disruption in *Drosophila* apoptosis. Dev Cell 12(5):793–806

Accorsi A, Zibaee A, Malagoli D (2015) The multifaceted activity of insect caspases. J Insect Physiol 76:17–23

Ahmad M, Srinivasula SM, Wang L, Litwack G, Fernandes-Alnemri T, Alnemri ES (1997) *Spodoptera frugiperda* caspase-1, a novel insect death protease that cleaves the nuclear immunophilin FKBP46, is the target of the baculovirus antiapoptotic protein p35. J Biol Chem 272(3):1421–1424

Arama E, Agapite J, Steller H (2003) Caspase activity and a specific cytochrome C are required for sperm differentiation in *Drosophila*. Dev Cell 4(5):687–697

Bartholomay LC, Cho WL, Rocheleau TA, Boyle JP, Beck ET, Fuchs JF, Liss P, Rusch M, Butler KM, Wu RC, Lin SP, Kuo HY, Tsao IY, Huang CY, Liu TT, Hsiao KJ, Tsai SF, Yang UC, Nappi AJ, Perna NT, Chen CC, Christensen BM (2004) Description of the transcriptomes of immune response-activated hemocytes from the mosquito vectors *Aedes aegypti* and *Armigeres subalbatus*. Infect Immun 72(7):4114–4126

Baum JS, Arama E, Steller H, McCall K (2007) The *Drosophila* caspases Strica and Dronc function redundantly in programmed cell death during oogenesis. Cell Death Differ 14(8):1508–1517

Beck ET, Blair CD, Black WC, Beaty BJ, Blitvich BJ (2007) Alternative splicing generates multiple transcripts of the inhibitor of apoptosis protein 1 in *Aedes* and *Culex* spp. mosquitoes. Insect Biochem Mol Biol 37(11):1222–1233

Benedict MA, Hu Y, Inohara N, Nunez G (2000) Expression and functional analysis of Apaf-1 isoforms. Extra Wd-40 repeat is required for cytochrome c binding and regulated activation of procaspase-9. J Biol Chem 275(12):8461–8468

Berthelet J, Dubrez L (2013) Regulation of apoptosis by Inhibitors of Apoptosis (IAPs). Cells 2(1):163–187

Blitvich BJ, Blair CD, Kempf BJ, Hughes MT, Black WC, Mackie RS, Meredith CT, Beaty BJ, Rayms-Keller A (2002) Developmental- and tissue-specific expression of an inhibitor of apoptosis protein 1 homologue from *Aedes triseriatus* mosquitoes. Insect Mol Biol 11(5):431–442

Brodsky MH, Nordstrom W, Tsang G, Kwan E, Rubin GM, Abrams JM (2000) *Drosophila* p53 binds a damage response element at the reaper locus. Cell 101(1):103–113

Bryant B, Blair CD, Olson KE, Clem RJ (2008) Annotation and expression profiling of apoptosis-related genes in the yellow fever mosquito, *Aedes aegypti*. Insect Biochem Mol Biol 38(3):331–345

Bryant B, Zhang Y, Zhang C, Santos CP, Clem RJ, Zhou L (2009) A lepidopteran orthologue of reaper reveals functional conservation and evolution of IAP antagonists. Insect Mol Biol 18(3):341–351

Bryant B, Ungerer MC, Liu Q, Waterhouse RM, Clem RJ (2010) A caspase-like decoy molecule enhances the activity of a paralogous caspase in the yellow fever mosquito, *Aedes aegypti*. Insect Biochem Mol Biol 40(7):516–523

Budhidarmo R, Day CL (2015) IAPs: modular regulators of cell signalling. Semin Cell Dev Biol 39:80–90

Chai J, Yan N, Huh JR, Wu JW, Li W, Hay BA, Shi Y (2003) Molecular mechanism of Reaper-Grim-Hid-mediated suppression of DIAP1-dependent Dronc ubiquitination. Nat Struct Biol 10(11):892–898

Chen P, Nordstrom W, Gish B, Abrams JM (1996) grim, a novel cell death gene in *Drosophila*. Genes Dev 10(14):1773–1782

Chen P, Rodriguez A, Erskine R, Thach T, Abrams JM (1998) Dredd, a novel effector of the apoptosis activators reaper, grim, and hid in *Drosophila*. Dev Biol 201(2):202–216

Christophides GK, Zdobnov E, Barillas-Mury C, Birney E, Blandin S, Blass C, Brey PT, Collins FH, Danielli A, Dimopoulos G, Hetru C, Hoa NT, Hoffmann JA, Kanzok SM, Letunic I, Levashina EA, Loukeris TG, Lycett G, Meister S, Michel K, Moita LF, Muller HM, Osta MA, Paskewitz SM, Reichhart JM, Rzhetsky A, Troxler L, Vernick KD, Vlachou D, Volz J, von Mering C, Xu J, Zheng L, Bork P, Kafatos FC (2002) Immunity-related genes and gene families in *Anopheles gambiae*. Science 298(5591):159–165

Clavier A, Rincheval-Arnold A, Colin J, Mignotte B, Guenal I (2016) Apoptosis in *Drosophila*: which role for mitochondria? Apoptosis 21(3):239–251

Clem RJ (2005) The role of apoptosis in defense against baculovirus infection in insects. Curr Top Microbiol Immunol 289:113–129

Colin J, Garibal J, Clavier A, Szuplewski S, Risler Y, Milet C, Gaumer S, Guenal I, Mignotte B (2015) Screening of suppressors of bax-induced cell death identifies glycerophosphate oxidase-1 as a mediator of debcl-induced apoptosis in *Drosophila*. Genes Cancer 6(5–6):241–253

Colussi PA, Quinn LM, Huang DC, Coombe M, Read SH, Richardson H, Kumar S (2000) Debcl, a proapoptotic Bcl-2 homologue, is a component of the *Drosophila melanogaster* cell death machinery. J Cell Biol 148(4):703–714

Cooper DM, Pio F, Thi EP, Theilmann D, Lowenberger C (2007a) Characterization of *Aedes* Dredd: a novel initiator caspase from the yellow fever mosquito, *Aedes aegypti*. Insect Biochem Mol Biol 37(6):559–569

Cooper DM, Thi EP, Chamberlain CM, Pio F, Lowenberger C (2007b) *Aedes* Dronc: a novel ecdysone-inducible caspase in the yellow fever mosquito, *Aedes aegypti*. Insect Mol Biol 16(5):563–572

Cooper DM, Chamberlain CM, Lowenberger C (2009a) Aedes FADD: a novel death domain-containing protein required for antibacterial immunity in the yellow fever mosquito, *Aedes aegypti*. Insect Biochem Mol Biol 39(1):47–54

Cooper DM, Granville DJ, Lowenberger C (2009b) The insect caspases. Apoptosis 14(3):247–256

Courtiade J, Pauchet Y, Vogel H, Heckel DG (2011) A comprehensive characterization of the caspase gene family in insects from the order Lepidoptera. BMC Genom 12:357

Crook NE, Clem RJ, Miller LK (1993) An apoptosis-inhibiting baculovirus gene with a zinc finger-like motif. J Virol 67(4):2168–2174

Denton D, Kumar S (2015) Studying apoptosis in *Drosophila*. Cold Spring Harbor Protoc 2015(7):609–613

Denton D, Aung-Htut MT, Kumar S (2013) Developmentally programmed cell death in *Drosophila*. Biochim Biophys Acta 1833(12):3499–3506

Ditzel M, Broemer M, Tenev T, Bolduc C, Lee TV, Rigbolt KT, Elliott R, Zvelebil M, Blagoev B, Bergmann A, Meier P (2008) Inactivation of effector caspases through nondegradative polyubiquitylation. Mol Cell 32(4):540–553

Domingues C, Ryoo HD (2012) *Drosophila* BRUCE inhibits apoptosis through non-lysine ubiquitination of the IAP-antagonist REAPER. Cell Death Differ 19(3):470–477

Dorstyn L, Kumar S (2008) A biochemical analysis of the activation of the *Drosophila* caspase DRONC. Cell Death Differ 15(3):461–470

Dorstyn L, Colussi PA, Quinn LM, Richardson H, Kumar S (1999a) DRONC, an ecdysone-inducible *Drosophila* caspase. Proc Natl Acad Sci U S A 96(8):4307–4312

Dorstyn L, Read SH, Quinn LM, Richardson H, Kumar S (1999b) DECAY, a novel *Drosophila* caspase related to mammalian caspase-3 and caspase-7. J Biol Chem 274(43):30778–30783

Dorstyn L, Mills K, Lazebnik Y, Kumar S (2004) The two cytochrome c species, DC3 and DC4, are not required for caspase activation and apoptosis in *Drosophila* cells. J Cell Biol 167(3):405–410

Doumanis J, Quinn L, Richardson H, Kumar S (2001) STRICA, a novel *Drosophila melanogaster* caspase with an unusual serine/threonine-rich prodomain, interacts with DIAP1 and DIAP2. Cell Death Differ 8(4):387–394

Du C, Fang M, Li Y, Li L, Wang X (2000) Smac, a mitochondrial protein that promotes cytochrome c-dependent caspase activation by eliminating IAP inhibition. Cell 102(1):33–42

Fuentes-Prior P, Salvesen GS (2004) The protein structures that shape caspase activity, specificity, activation and inhibition. Biochem J 384(Pt 2):201–232

Galluzzi L, Vitale I, Aaronson SA, Abrams JM, Adam D, Agostinis P, Alnemri ES, Altucci L, Amelio I, Andrews DW, Annicchiarico-Petruzzelli M, Antonov AV, Arama E, Baehrecke EH, Barlev NA, Bazan NG, Bernassola F, Bertrand MJM, Bianchi K, Blagosklonny MV, Blomgren K, Borner C, Boya P, Brenner C, Campanella M, Candi E, Carmona-Gutierrez D, Cecconi F, Chan FK, Chandel NS, Cheng EH, Chipuk JE, Cidlowski JA, Ciechanover A, Cohen GM, Conrad M, Cubillos-Ruiz JR, Czabotar PE, D'Angiolella V, Dawson TM, Dawson VL, De Laurenzi V, De Maria R, Debatin KM, DeBerardinis RJ, Deshmukh M, Di Daniele N, Di Virgilio F, Dixit VM, Dixon SJ, Duckett CS, Dynlacht BD, El-Deiry WS, Elrod JW, Fimia GM, Fulda S, García-Sáez AJ, Garg AD, Garrido C, Gavathiotis E, Golstein P, Gottlieb E, Green DR, Greene LA,

Gronemeyer H, Gross A, Hajnoczky G, Hardwick JM, Harris IS, Hengartner MO, Hetz C, Ichijo H, Jäättelä M, Joseph B, Jost PJ, Juin PP, Kaiser WJ, Karin M, Kaufmann T, Kepp O, Kimchi A, Kitsis RN, Klionsky DJ, Knight RA, Kumar S, Lee SW, Lemasters JJ, Levine B, Linkermann A, Lipton SA, Lockshin RA, López-Otín C, Lowe SW, Luedde T, Lugli E, MacFarlane M, Madeo F, Malewicz M, Malorni W, Manic G, Marine JC, Martin SJ, Martinou JC, Medema JP, Mehlen P, Meier P, Melino S, Miao EA, Molkentin JD, Moll UM, Muñoz-Pinedo C, Nagata S, Nuñez G, Oberst A, Oren M, Overholtzer M, Pagano M, Panaretakis T, Pasparakis M, Penninger JM, Pereira DM, Pervaiz S, Peter ME, Piacentini M, Pinton P, Prehn JHM, Puthalakath H, Rabinovich GA, Rehm M, Rizzuto R, Rodrigues CMP, Rubinsztein DC, Rudel T, Ryan KM, Sayan E, Scorrano L, Shao F, Shi Y, Silke J, Simon HU, Sistigu A, Stockwell BR, Strasser A, Szabadkai G, Tait SWG, Tang D, Tavernarakis N, Thorburn A, Tsujimoto Y, Turk B, Vanden Berghe T, Vandenabeele P, Vander Heiden MG, Villunger A, Virgin HW, Vousden KH, Vucic D, Wagner EF, Walczak H, Wallach D, Wang Y, Wells JA, Wood W, Yuan J, Zakeri Z, Zhivotovsky B, Zitvogel L, Melino G, Kroemer G (2018) Molecular mechanisms of cell death: recommendations of the nomenclature committee on cell death. Cell Death Differ 25(3):486–541

Girard YA, Schneider BS, McGee CE, Wen J, Han VC, Popov V, Mason PW, Higgs S (2007) Salivary gland morphology and virus transmission during long-term cytopathologic West Nile virus infection in *Culex* mosquitoes. Am J Trop Med Hyg 76(1):118–128

Grimaldi D, Engel MS (2005) Evolution of the Insects, 755pp. Cambridge University Press, New York, Melbourne

Hansen AK, Moran NA (2014) The impact of microbial symbionts on host plant utilization by herbivorous insects. Mol Ecol 23(6):1473–1496

Harvey NL, Daish T, Mills K, Dorstyn L, Quinn LM, Read SH, Richardson H, Kumar S (2001) Characterization of the *Drosophila* caspase, DAMM. J Biol Chem 276(27):25342–25350

Hawkins CJ, Yoo SJ, Peterson EP, Wang SL, Vernooy SY, Hay BA (2000) The *Drosophila* caspase DRONC cleaves following glutamate or aspartate and is regulated by DIAP1, HID, and GRIM. J Biol Chem 275(35):27084–27093

Hay BA, Guo M (2006) Caspase-dependent cell death in *Drosophila*. Annu Rev Cell Dev Biol 22:623–650

Hegde R, Srinivasula SM, Zhang Z, Wassell R, Mukattash R, Cilenti L, DuBois G, Lazebnik Y, Zervos AS, Fernandes-Alnemri T, Alnemri ES (2002) Identification of Omi/HtrA2 as a mitochondrial apoptotic serine protease that disrupts inhibitor of apoptosis protein-caspase interaction. J Biol Chem 277(1):432–438

Hu S, Yang X (2000) dFADD, a novel death domain-containing adapter protein for the *Drosophila* caspase DREDD. J Biol Chem 275(40):30761–30764

Hu Y, Benedict MA, Ding L, Nunez G (1999) Role of cytochrome c and dATP/ATP hydrolysis in Apaf-1-mediated caspase-9 activation and apoptosis. EMBO J 18(13):3586–3595

Huang Q, Deveraux QL, Maeda S, Stennicke HR, Hammock BD, Reed JC (2001) Cloning and characterization of an inhibitor of apoptosis protein (IAP) from *Bombyx mori*. Biochim Biophys Acta 1499(3):191–198

Huang N, Civciristov S, Hawkins CJ, Clem RJ (2013) SfDronc, an initiator caspase involved in apoptosis in the fall armyworm *Spodoptera frugiperda*. Insect Biochem Mol Biol 43(5):444–454

Huang HJ, Bao YY, Lao SH, Huang XH, Ye YZ, Wu JX, Xu HJ, Zhou XP, Zhang CX (2015) Rice ragged stunt virus-induced apoptosis affects virus transmission from its insect vector, the brown planthopper to the rice plant. Sci Rep 5:11413

Igaki T, Miura M (2004) Role of Bcl-2 family members in invertebrates. Biochim Biophys Acta 1644(2–3):73–81

Jiang C, Lamblin AF, Steller H, Thummel CS (2000) A steroid-triggered transcriptional hierarchy controls salivary gland cell death during *Drosophila* metamorphosis. Mol Cell 5(3):445–455

Jones G, Jones D, Zhou L, Steller H, Chu Y (2000) Deterin, a new inhibitor of apoptosis from *Drosophila melanogaster*. J Biol Chem 275(29):22157–22165

Kale J, Osterlund EJ, Andrews DW (2018) BCL-2 family proteins: changing partners in the dance towards death. Cell Death Differ 25(1):65–80

Kalkavan H, Green DR (2018) MOMP, cell suicide as a BCL-2 family business. Cell Death Differ 25(1):46–55

Kilpatrick ZE, Cakouros D, Kumar S (2005) Ecdysone-mediated up-regulation of the effector caspase DRICE is required for hormone-dependent apoptosis in *Drosophila* cells. J Biol Chem 280(12):11981–11986

Kornbluth S, White K (2005) Apoptosis in *Drosophila*: neither fish nor fowl (nor man, nor worm). J Cell Sci 118(Pt 9):1779–1787

Kumar S (2007) Caspase function in programmed cell death. Cell Death Differ 14(1):32–43

Kumar S, Doumanis J (2000) The fly caspases. Cell Death Differ 7(11):1039–1044

Kumarswamy R, Seth RK, Dwarakanath BS, Chandna S (2009) Mitochondrial regulation of insect cell apoptosis: evidence for permeability transition pore-independent cytochrome-c release in the Lepidopteran Sf9 cells. Int J Biochem Cell Biol 41(6):1430–1440

Kurada P, White K (1998) Ras promotes cell survival in *Drosophila* by downregulating hid expression. Cell 95(3):319–329

Kvansakul M, Hinds MG (2015) The Bcl-2 family: structures, interactions and targets for drug discovery. Apoptosis 20(2):136–150

Lamkanfi M, Declercq W, Kalai M, Saelens X, Vandenabeele P (2002) Alice in caspase land. A phylogenetic analysis of caspases from worm to man. Cell Death Differ 9(4):358–361

Leulier F, Lhocine N, Lemaitre B, Meier P (2006a) The *Drosophila* inhibitor of apoptosis protein DIAP2 functions in innate immunity and is essential to resist gram-negative bacterial infection. Mol Cell Biol 26(21):7821–7831

Leulier F, Ribeiro PS, Palmer E, Tenev T, Takahashi K, Robertson D, Zachariou A, Pichaud F, Ueda R, Meier P (2006b) Systematic in vivo RNAi analysis of putative components of the *Drosophila* cell death machinery. Cell Death Differ 13(10):1663–1674

Li Q, Li H, Blitvich BJ, Zhang J (2007) The *Aedes albopictus* inhibitor of apoptosis 1 gene protects vertebrate cells from bluetongue virus-induced apoptosis. Insect Mol Biol 16(1):93–105

Liu Q, Clem RJ (2011) Defining the core apoptosis pathway in the mosquito disease vector *Aedes aegypti*: the roles of iap1, ark, dronc, and effector caspases. Apoptosis 16(2):105–113

Liu Q, Qi Y, Chejanovsky N (2005) *Spodoptera littoralis* caspase-1, a Lepidopteran effector caspase inducible by apoptotic signaling. Apoptosis 10(4):787–795

Lockshin RA, Williams CM (1965) Programmed cell death. V. Cytolytic enzymes in relation to the breakdown of the intersegmental muscles of silkmoths. J Insect Physiol 11(7):831–844

Malin JZ, Shaham S (2015) Cell death in *C. elegans* development. Curr Top Dev Biol 114:1–42

Meinander A, Runchel C, Tenev T, Chen L, Kim CH, Ribeiro PS, Broemer M, Leulier F, Zvelebil M, Silverman N, Meier P (2012) Ubiquitylation of the initiator caspase DREDD is required for innate immune signalling. EMBO J 31(12):2770–2783

Mendes CS, Arama E, Brown S, Scherr H, Srivastava M, Bergmann A, Steller H, Mollereau B (2006) Cytochrome c-d regulates developmental apoptosis in the *Drosophila* retina. EMBO Rep 7(9):933–939

Meng K, Li X, Wang S, Zhong C, Yang Z, Feng L, Liu Q (2016) The Strica homolog AaCASPS16 is involved in apoptosis in the yellow fever vector, *Aedes albopictus*. PLoS ONE 11(6):e0157846

Misof B, Liu S, Meusemann K, Peters RS, Donath A, Mayer C, Frandsen PB, Ware J, Flouri T, Beutel RG, Niehuis O, Petersen M, Izquierdo-Carrasco F, Wappler T, Rust J, Aberer AJ, Aspöck U, Aspöck H, Bartel D, Blanke A, Berger S, Böhm A, Buckley TR, Calcott B, Chen J, Friedrich F, Fukui M, Fujita M, Greve C, Grobe P, Gu PS, Huang Y, Jermiin LS, Kawahara AY, Krogmann L, Kubiak M, Lanfear R, Letsch H, Li Y, Li Z, Li J, Lu H, Machida R, Mashimo Y, Kapli P, McKenna DD, Meng G, Nakagaki Y, Navarrete-Heredia JL, Ott M, Ou Y, Pass G, Podsiadlowski L, Pohl H, von Reumont BM, Schütte K, Sekiya K, Shimizu S, Slipinski A, Stamatakis A, Song W, Su X, Szucsich NU, Tan M, Tan X, Tang M, Tang J, Timelthaler G, Tomizuka S, Trautwein M, Tong X, Uchifune T, Walzl MG, Wiegmann BM, Wilbrandt J, Wipfler B, Wong TK, Wu Q, Wu G, Xie Y, Yang S, Yang Q, Yeates DK, Yoshizawa K, Zhang Q, Zhang R, Zhang W, Zhang Y, Zhao J, Zhou C, Zhou L, Ziesmann T, Zou S, Li Y, Xu X, Zhang Y, Yang H, Wang J, Wang

J, Kjer KM, Zhou X (2014) Phylogenomics resolves the timing and pattern of insect evolution. Science 346(6210):763–767

Miura, M. (2012). Apoptotic and nonapoptotic caspase functions in animal development. Cold Spring Harb Perspect Biol 4(10)

Muro I, Berry DL, Huh JR, Chen CH, Huang H, Yoo SJ, Guo M, Baehrecke EH, Hay BA (2006) The *Drosophila* caspase Ice is important for many apoptotic cell deaths and for spermatid individualization, a nonapoptotic process. Development 133(17):3305–3315

Olson MR, Holley CL, Gan EC, Colon-Ramos DA, Kaplan B, Kornbluth S (2003) A GH3-like domain in reaper is required for mitochondrial localization and induction of IAP degradation. J Biol Chem 278(45):44758–44768

Opferman JT, Korsmeyer SJ (2003) Apoptosis in the development and maintenance of the immune system. Nat Immunol 4(5):410–415

Orme M, Meier P (2009) Inhibitor of apoptosis proteins in *Drosophila*: gatekeepers of death. Apoptosis 14(8):950–960

Pang Y, Bai XC, Yan C, Hao Q, Chen Z, Wang JW, Scheres SH, Shi Y (2015) Structure of the apoptosome: mechanistic insights into activation of an initiator caspase from *Drosophila*. Genes Dev 29(3):277–287

Parrish AB, Freel CD, Kornbluth S (2013) Cellular mechanisms controlling caspase activation and function. Cold Spring Harb Perspect Biol 5(6)

Pei Z, Reske G, Huang Q, Hammock BD, Qi Y, Chejanovsky N (2002) Characterization of the apoptosis suppressor protein P49 from the *Spodoptera littoralis* nucleopolyhedrovirus. J Biol Chem 277(50):48677–48684

Pellettieri J, Sanchez Alvarado A (2007) Cell turnover and adult tissue homeostasis: from humans to planarians. Annu Rev Genet 41:83–105

Puglise JM, Estep AS, Becnel JJ (2016) Expression profiles and RNAi silencing of inhibitor of apoptosis transcripts in *Aedes, Anopheles*, and *Culex* Mosquitoes (Diptera: Culicidae). J Med Entomol 53(2):304–314

Quinn LM, Dorstyn L, Mills K, Colussi PA, Chen P, Coombe M, Abrams J, Kumar S, Richardson H (2000) An essential role for the caspase dronc in developmentally programmed cell death in *Drosophila*. J Biol Chem 275(51):40416–40424

Quinn L, Coombe M, Mills K, Daish T, Colussi P, Kumar S, Richardson H (2003) Buffy, a *Drosophila* Bcl-2 protein, has anti-apoptotic and cell cycle inhibitory functions. EMBO J 22(14):3568–3579

Ribeiro PS, Kuranaga E, Tenev T, Leulier F, Miura M, Meier P (2007) DIAP2 functions as a mechanism-based regulator of drICE that contributes to the caspase activity threshold in living cells. J Cell Biol 179(7):1467–1480

Rodriguez A, Oliver H, Zou H, Chen P, Wang X, Abrams JM (1999) Dark is a *Drosophila* homologue of Apaf-1/CED-4 and functions in an evolutionarily conserved death pathway. Nat Cell Biol 1(5):272–279

Saraste A, Pulkki K (2000) Morphologic and biochemical hallmarks of apoptosis. Cardiovasc Res 45(3):528–537

Sehnal F, Svacha P, Zrzavy JAN (1996) Evolution of Insect Metamorphosis. In: Gilbert LI, Tata JR, Atkinson BG (eds) Metamorphosis Postembryonic reprograming of gene expression in amphibian and insect cells. Academic Press, San Diego, pp 3–58

Seol DW, Billiar TR (1999) A caspase-9 variant missing the catalytic site is an endogenous inhibitor of apoptosis. J Biol Chem 274(4):2072–2076

Shapiro PJ, Hsu HH, Jung H, Robbins ES, Ryoo HD (2008) Regulation of the *Drosophila* apoptosome through feedback inhibition. Nat Cell Biol 10(12):1440–1446

Shi Y (2002) A conserved tetrapeptide motif: potentiating apoptosis through IAP-binding. Cell Death Differ 9(2):93–95

Shlevkov E, Morata G (2012) A dp53/JNK-dependant feedback amplification loop is essential for the apoptotic response to stress in *Drosophila*. Cell Death Differ 19(3):451–460

Shu B, Zhang J, Sethuraman V, Cui G, Yi X, Zhong G (2017) Transcriptome analysis of *Spodoptera frugiperda* Sf9 cells reveals putative apoptosis-related genes and a preliminary apoptosis mechanism induced by azadirachtin. Sci Rep 7(1):13231

Silke J, Vaux DL (2001) Two kinds of BIR-containing protein - inhibitors of apoptosis, or required for mitosis. J Cell Sci 114(Pt 10):1821–1827

Simonet P, Gaget K, Balmand S, Ribeiro Lopes M, Parisot N, Buhler K, Duport G, Vulsteke V, Febvay G, Heddi A, Charles H, Callaerts P, Calevro F (2018) Bacteriocyte cell death in the pea aphid/*Buchnera* symbiotic system. Proc Natl Acad Sci USA

Song Z, McCall K, Steller H (1997) DCP-1, a *Drosophila* cell death protease essential for development. Science 275(5299):536–540

Stork NE (2018) How many species of insects and other terrestrial arthropods are there on earth? Annu Rev Entomol 63:31–45

Stoven S, Silverman N, Junell A, Hedengren-Olcott M, Erturk D, Engstrom Y, Maniatis T, Hultmark D (2003) Caspase-mediated processing of the *Drosophila* NF-kappaB factor Relish. Proc Natl Acad Sci U S A 100(10):5991–5996

Suganuma I, Ushiyama T, Yamada H, Iwamoto A, Kobayashi M, Ikeda M (2011) Cloning and characterization of a dronc homologue in the silkworm, *Bombyx mori*. Insect Biochem Mol Biol 41(11):909–921

Tagu D, Calevro F, Colella S, Gabaldón T, Sugio A (2016) Functional and evolutionary genomics in aphids. In: Vilcinskas A (ed) Biology and ecology of aphids CRC Press, Taylor & Francis Group, pp 52–88

Tambunan J, Kan Chang P, Li H, Natori M (1998) Molecular cloning of a cDNA encoding a silkworm protein that contains the conserved BH regions of Bcl-2 family proteins. Gene 212(2):287–293

Tenev T, Zachariou A, Wilson R, Ditzel M, Meier P (2005) IAPs are functionally non-equivalent and regulate effector caspases through distinct mechanisms. Nat Cell Biol 7(1):70–77

The International Aphid Genomics Consortium (2010) Genome sequence of the pea aphid *Acyrthosiphon pisum*. PLoS Biol 8(2):e1000313

Tsiatsiani L, Van Breusegem F, Gallois P, Zavialov A, Lam E, Bozhkov PV (2011) Metacaspases. Cell Death Differ 18(8):1279–1288

Uren AG, Coulson EJ, Vaux DL (1998) Conservation of baculovirus inhibitor of apoptosis repeat proteins (BIRPs) in viruses, nematodes, vertebrates and yeasts. Trends Biochem Sci 23(5):159–162

Vasudevan D, Ryoo HD (2015) Regulation of cell ceath by IAPs and their antagonists. Curr Top Dev Biol 114:185–208

Verhagen AM, Vaux DL (2002) Cell death regulation by the mammalian IAP antagonist Diablo/Smac. Apoptosis 7(2):163–166

Verhagen AM, Silke J, Ekert PG, Pakusch M, Kaufmann H, Connolly LM, Day CL, Tikoo A, Burke R, Wrobel C, Moritz RL, Simpson RJ, Vaux DL (2002) HtrA2 promotes cell death through its serine protease activity and its ability to antagonize inhibitor of apoptosis proteins. J Biol Chem 277(1):445–454

Vigneron A, Masson F, Vallier A, Balmand S, Rey M, Vincent-Monégat C, Aksoy E, Aubailly-Giraud E, Zaidman-Rémy A, Heddi A (2014) Insects recycle endosymbionts when the benefit is over. Curr Biol 24(19):2267–2273

Wang H, Clem RJ (2011) The role of IAP antagonist proteins in the core apoptosis pathway of the mosquito disease vector *Aedes aegypti*. Apoptosis 16(3):235–248

Wang XR, Wang C, Wang XW, Qian LX, Chi Y, Liu SS, Liu YQ, Wang XW (2018) The functions of caspase in whitefly *Bemisia tabaci* apoptosis in response to ultraviolet irradiation. Insect Mol Biol 27(6):739–751

Waterhouse RM, Kriventseva EV, Meister S, Xi Z, Alvarez KS, Bartholomay LC, Barillas-Mury C, Bian G, Blandin S, Christensen BM, Dong Y, Jiang H, Kanost MR, Koutsos AC, Levashina EA, Li J, Ligoxygakis P, Maccallum RM, Mayhew GF, Mendes A, Michel K, Osta MA, Paskewitz S, Shin SW, Vlachou D, Wang L, Wei W, Zheng L, Zou Z, Severson DW, Raikhel AS, Kafatos

FC, Dimopoulos G, Zdobnov EM, Christophides GK (2007) Evolutionary dynamics of immune-related genes and pathways in disease-vector mosquitoes. Science 316(5832):1738–1743

White K, Grether ME, Abrams JM, Young L, Farrell K, Steller H (1994) Genetic control of programmed cell death in *Drosophila*. Science 264(5159):677–683

Wing JP, Zhou L, Schwartz LM, Nambu JR (1998) Distinct cell killing properties of the *Drosophila* reaper, head involution defective, and grim genes. Cell Death Differ 5(11):930–939

Xu D, Li Y, Arcaro M, Lackey M, Bergmann A (2005) The CARD-carrying caspase Dronc is essential for most, but not all, developmental cell death in *Drosophila*. Development 132(9):2125–2134

Xu D, Wang Y, Willecke R, Chen Z, Ding T, Bergmann A (2006) The effector caspases drICE and dcp-1 have partially overlapping functions in the apoptotic pathway in *Drosophila*. Cell Death Differ 13(10):1697–1706

Xu D, Woodfield SE, Lee TV, Fan Y, Antonio C, Bergmann A (2009) Genetic control of programmed cell death (apoptosis) in *Drosophila*. Fly (Austin) 3(1):78–90

Yang D, Chai L, Wang J, Zhao X (2008) Molecular cloning and characterization of Hearm caspase-1 from *Helicoverpa armigera*. Mol Biol Rep 35(3):405–412

Yang Z, Zhou K, Liu H, Wu A, Mei L, Liu Q (2016) SfDredd, a novel initiator caspase possessing activity on effector caspase substrates in *Spodoptera frugiperda*. PLoS ONE 11(3):e0151016

Yi HS, Pan CX, Pan C, Song J, Hu YF, Wang L, Pan MH, Lu C (2014) BmICE-2 is a novel pro-apoptotic caspase involved in apoptosis in the silkworm, *Bombyx mori*. Biochem Biophys Res Commun 445(1):100–106

Yoo SJ, Huh JR, Muro I, Yu H, Wang L, Wang SL, Feldman RM, Clem RJ, Muller HA, Hay BA (2002) Hid, Rpr and Grim negatively regulate DIAP1 levels through distinct mechanisms. Nat Cell Biol 4(6):416–424

Yu SY, Yoo SJ, Yang L, Zapata C, Srinivasan A, Hay BA, Baker NE (2002) A pathway of signals regulating effector and initiator caspases in the developing *Drosophila* eye. Development 129(13):3269–3278

Yuan J, Shaham S, Ledoux S, Ellis HM, Horvitz HR (1993) The *C. elegans* cell death gene ced-3 encodes a protein similar to mammalian interleukin-1 beta-converting enzyme. Cell 75(4):641–652

Zachariou A, Tenev T, Goyal L, Agapite J, Steller H, Meier P (2003) IAP-antagonists exhibit non-redundant modes of action through differential DIAP1 binding. EMBO J 22(24):6642–6652

Zhang H, Huang Q, Ke N, Matsuyama S, Hammock B, Godzik A, Reed JC (2000) *Drosophila* pro-apoptotic Bcl-2/Bax homologue reveals evolutionary conservation of cell death mechanisms. J Biol Chem 275(35):27303–27306

Zhang JY, Pan MH, Sun ZY, Huang SJ, Yu ZS, Liu D, Zhao DH, Lu C (2010) The genomic underpinnings of apoptosis in the silkworm, *Bombyx mori*. BMC Genom 11:611

Zhou L, Jiang G, Chan G, Santos CP, Severson DW, Xiao L (2005) Michelob_x is the missing inhibitor of apoptosis protein antagonist in mosquito genomes. EMBO Rep 6(8):769–774

Chapter 14
Xenacoelomorpha, a Key Group to Understand Bilaterian Evolution: Morphological and Molecular Perspectives

Ulf Jondelius, Olga I. Raikova and Pedro Martinez

Abstract The Xenacoelomorpha is a clade of mostly marine animals placed as the sister group of the remaining Bilateria (Nephrozoa) in most phylogenomic and morphological analyses, although alternative hypotheses placing them within deuterostomes have been proposed. This key phylogenetic position has raised recently a great interest in the study of their constitutive clades, since they can provide us with character states that illuminate different aspects of the origin of bilateral animals. Moreover, the recent availability of genomic and transcriptomic data from different species has been used in inferring the internal relationships among xenacoelomorph clades and the deciphering of molecular mechanisms that contribute to the evolution of metazoan genomes. Having access to molecular data paves the way to the systematic analysis of the genetic control of xenacoelomorph development and, additionally, to a better-informed study of bilaterian innovations. Here we revisit what has been learned over the last decades on the morphology, genomics and phylogenetic relationships of the Xenacoelomorpha.

U. Jondelius (✉)
Swedish Museum of Natural History, Box 50007, 104 05 Stockholm, Sweden
e-mail: ulf.jondelius@nrm.se

Department of Zoology, Stockholm University, Svante Ahrrenius Väg 18b, 106 91 Stockholm, Sweden

O. I. Raikova (✉)
Zoological Institute RAS, Universitetskaya emb. 1, 199034 St. Petersburg, Russia
e-mail: oraikova@gmail.com

Faculty of Biology, Chair of Invertebrate Zoology, St. Petersburg State University, Universitetskaya emb. 7/9, 199034 St. Petersburg, Russia

P. Martinez (✉)
Department de Genètica, Microbiologia i Estadística, Universitat de Barcelona, Av. Diagonal, 645, 08028 Barcelona, Spain
e-mail: pedro.martinez@ub.edu

ICREA (Institut Català de Recerca i Estudis Avancats), Barcelona, Spain

P. Pontarotti (ed.), *Evolution, Origin of Life, Concepts and Methods*,
https://doi.org/10.1007/978-3-030-30363-1_14

14.1 Xenacoelomorpha: Taxonomic Status

Superphylum Xenacoelomorpha (Philippe et al. 2011) currently includes two phyla: Xenoturbellida (Bourlat et al. 2006) and Acoelomorpha (Ehlers 1985) with two subgroups that are traditionally considered to be orders: Nemertodermatida (Karling 1940) and Acoela (Uljanin 1870). Xenacoelomorpha lack any type of nephridia, which distinguishes them from the rest of Bilateria (Nephrozoa).

14.2 Xenacoelomorpha and Its Constitutive Clades: Phylogenetic Considerations

Xenacoelomorpha are soft-bodied marine hermaphroditic worms comprising three subgroups: Acoela, Nemertodermatida and *Xenoturbella* (see Fig. 14.1 for representatives). Acoela is by far the most diverse group with about 400 nominal species abundant in many marine habitats ranging from sandy beaches to deep-water mud and also with a few pelagic species. The 18 nominal species of Nemertodermatida are much rarer, many of them known from just a few specimens from marine sandy and muddy sediments below the intertidal zone (Sterrer 1998). Endosymbiotic species occur within both groups: *Meara stichopi* is an endosymbiotic nemertodermatid occurring in the intestine of the holothurian *Parastichopus tremulus* (Westblad 1949a), and within the acoel genus *Avagina,* there are three species that occur in echinoids (Dörjes 1972). Acoels and nemertodermatids are small animals with a maximum size of a few millimetres whereas the five species of *Xenoturbella* are much larger, attaining more than 20 cm in *X. monstrosa* (see below for detailed descriptions).

When they were first described, both Acoela and Nemertodermatida were classified as flatworms ("turbellarians"). The name Acoela (Uljanin 1870) refers to the lack of both intestinal cavity and body cavity in these animals. Nemertodermatida alludes to the thick gland-rich epidermis similar to that of nemertines. In contrast to acoels, nemertodermatids possess an epithelial gut, and they were classified in a separate order (Karling 1940). The first known *Xenoturbella*, X. *bocki,* was described in 1949 (Westblad 1949b) and, as the name indicates, it was considered to be a "strange turbellarian", i.e. also a flatworm, but strange due to its large size and very simple anatomy.

Acoela and Nemertodermatida are considered sister groups making up the group Acoelomorpha (Ehlers 1984). This hypothesis was first based on ultrastructural similarities in their epidermal ciliary rootlets and later received support from phylogenomic analyses (Hejnol et al. 2009).

Acoelomorphs, based on similarities of morphological characters (lack body cavities (acoelomate structure), a hindgut or an anus), were originally classified within the Platyhelminthes (i.e. Conway-Morris et al. 1985). However, this original classification of Acoelomorpha within Platyhelminthes was difficult to reconcile with some of their morphological features, e.g. the absence of protonephridia, the divergent

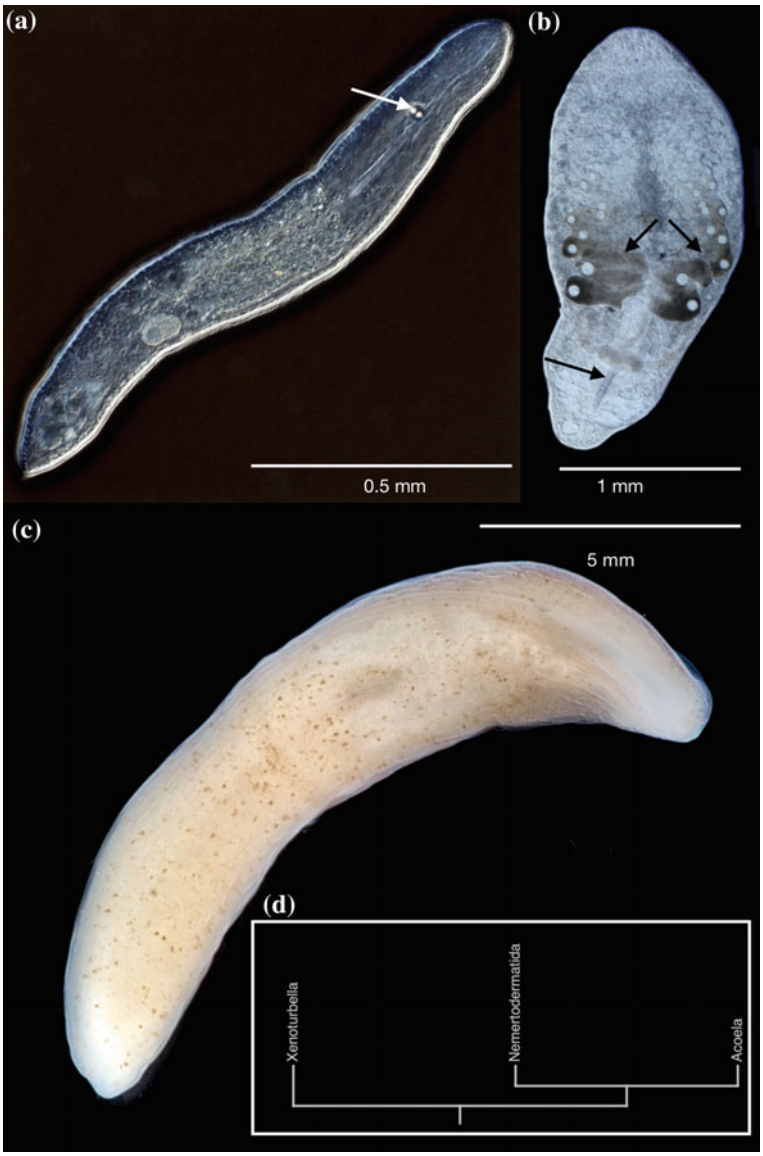


Fig. 14.1 **a**. Nemertodermatida. Live specimen of *Flagellophora apelti*. Charactersitsic double statoliths can be seen in anterior (arrow). **b**. Acoela. In this live specimen of *Childia macroposthium,* the male copulatory organ with proximal seminal vesicle can be seen at the posterior end (arrow) and the two ovaries are visible in the mid-third of the body (arrows). **c**. Swimming specimen of *Xenoturbella bocki* from the Gullmar fjord. **d**. Current phylogenetic hypothesis for Xenacoelomorpha where Xenoturbella is the sister taxon of Acoela and Nemertodermatida, which together form Acoelomorpha

sperm morphology and embryonic development. Even greater confusion pertained to *Xenoturbella*, which differs considerably from a flatworm in its morphology with larger size, rugby ball-like body shape, lack of defined gonads and epidermal nervous system. Its classification within Platyhelminthes was soon questioned; there has been an extraordinary array of different morphology-based phylogenetic hypotheses placing *Xenoturbella* close to Enteropneusta due to their epidermal histology and the morphology of the statocyst (Reisinger 1960), within Acoelomorpha on account of specialisations of its ciliary ultrastructure (Franzen and Afzelius 1987), as sister group to all other Bilateria based on epithelial ultrastructure (Ehlers 1985), and within Mollusca owing to similarities in oocyte morphology (Israelsson 1997).

The first molecular phylogenetic study that tested the position of Acoela in metazoan phylogeny and attempted to control for bias in phylogeny reconstruction potentially caused by among-lineage heterogeneity of substitution rates (so-called long branch attraction) was based on sequences from the 18S rRNA gene. A sister-group relationship between Acoela and Nephrozoa (Bilateria excluding Acoelomorpha and *Xenoturbella*) was supported in this study (Ruiz-Trillo et al. 1999). Subsequent analyses corroborated this result and added data on Nemertodermatida as well as sequences from other genes consistently demonstrating that the two taxa are not part of Platyhelminthes. In early studies, where much of the signal was derived from 18S and 28S rRNA genes, Acoela and Nemertodermatida did not form a monophyletic group but placed at the first and second split within Bilateria (Jondelius et al. 2002; Ruiz-Trillo et al. 2002; Telford et al. 2003), but subsequent phylogenomic studies supported a monophyletic Acoelomorpha, e.g. (Hejnol et al. 2009; Cannon et al. 2016, Rouse et al. 2016).

Xenoturbella initially proved difficult to place phylogenetically also when the first sequence data became available. In the first instance, a close relationship with molluscs was suggested (Noren and Jondelius 1997), but the first sequences from the cytochrome oxidase I mitochondrial gene (COI) were likely a food contamination from protobranch bivalves and analyses of new COI and 18 S sequence data supported a position within deuterostomes as sister group to Ambulacraria (Bourlat et al. 2003).

A comprehensive phylogenomic study of metazoan groups supported a topology where Acoelomorpha and Xenoturbella formed the sister group to Nephrozoa (Hejnol et al. 2009). Re-analysis of data from Hejnol et al. by Philippe et al. (2011) using the CAT substitution model allowing for across sites rate heterogeneity under a Bayesian framework did not recover the Acoelomorpha and *Xenoturbella* as sister group to Nephrozoa, but instead placed the two former close to or within deuterostomes with varying degrees of statistical support depending on the taxonomic composition of the data set. Philippe et al. (2011) suggested that previous studies had been affected by long branch attraction, which, they claimed, was overcome by the use of the CAT model in their study. They also proposed the name Xenacoelomorpha for the group Xenoturbella + Acoelomorpha. Cannon et al. (2016) added more data from acoelomorphs and made efforts to control for long branch attraction through analysis of a number of data sets with differing taxonomic composition under the CAT substitution model. Cannon et al. found strong support for a sister-group relationship between Xenacoelomorpha and Nephrozoa and proposed that missing data

and misleading signal from ribosomal protein genes in Philippe et al. (2011) caused Xenacoelomorpha to associate with deuterostomes in that study (for a discussion about ribosomal protein genes in deep animal phylogeny see: (Bleidorn et al. 2009)). The sister-group relationship between Xenacoelomorpha and Nephrozoa provides a parsimonious explanation for the lack of excretory organs, through gut and body cavity in Xenacoelomorpha; these features evolved within the Nephrozoan stem line. As noted by Philippe et al. (2011), a placing of Xenacoelomorpha within deuterostomes requires numerous ad hoc hypotheses to explain the absence of these features as well as specific deuterostome characters such as gill slits and endostylar tissue.

An early classification of Acoela in the two families Proporidae (acoels with one gonopore) and Convolutidae (acoels with two gonopores) was proposed by von Graff (1905, 1911) who also elevated Acoela to the rank of subclass within Platyhelminthes. Einar Westblad studied the Swedish acoel fauna in a series of papers where he gave detailed accounts of their anatomy and proposed a comprehensive classification of Acoela (Westblad 1940, 1942, 1945, 1946, 1948). Westblad's system, which comprised four additional families, was based on the histology of the gonads, presence or absence of female copulatory organs and the position and anatomy of the male copulatory organ. A major transformation of acoel taxonomy was carried out by Dörjes (1968) who described a large number of new species and introducing six new families and numerous new genera. None of the above classifications were based on explicit phylogenetic hypotheses. In more recent years, Hooge (2001) emphasized the configuration of the body wall musculature as phylogenetically informative and introduced four new families based on features of the body wall musculature and morphology of the male copulatory organ.

Jondelius et al. (2011) used ribosomal and mitochondrial nucleotide sequences in combination with morphological characters to reconstruct phylogeny and character evolution of Acoela and provide a phylogenetic classification of the group down to "family" level. They found a basal split in the acoel tree between Diopisthoporidae with a single hermaphrodite gland and Bitesticulata with paired or follicular testes. The latter group comprises fifteen "family level" taxa. Acoela are relatively easily recognized from their general habitus (cylindrical or drop-shaped, sometimes flattened, ciliated, having no gut lumen and no anus, with obvious statocyst in the anterior and clear copulatory organs in the posterior; all visible in a simple dissecting microscope), and the group is monophyletic in all molecular analyses. In spite of this it is difficult to point to an easily discernible unique acoel feature. The presence of a digestive parenchyma in lieu of an intestinal lumen and a particular type of biflagellate spermatozoa are morphological synapomorphies at the histological and ultrastructural levels. Their position as part of the sister group of Nephrozoa paired with their relative morphological simplicity renders acoels important for understanding the evolution of bilaterian complexity. Of particular interest is the reconstruction of ancestral features of acoels and of Xenacoelomorpha. Among the most studied acoels are several species of *Convolutriloba*, large acoels that are able to reproduce asexually through budding and that harbour symbiotic algae. The morphology of the *Convolutriloba* species was shown by Jondelius et al. (2011) to be highly divergent

from the ancestral acoel. Hence, caution is advised when drawing conclusions about acoel ancestral features based on *Convolutriloba*.

Nemertodermatida comprise 18 nominal species, far fewer than Acoela, and most of them are known from a small number of specimens (Sterrer 1998; Meyer-Wachsmuth et al. 2014). Nemertodermatids are easily recognized by the statocyst that contains two statoliths. Phylogenetic reconstruction based on the ribosomal 18S and 28S genes and the nuclear Histone H3 demonstrated a basal split between Ascopariidae and Nemertodermatidae with the symbiotic *Meara stichopi* as sister to *Sterreria* within Nemertodermatidae (Meyer-Wachsmuth and Jondelius 2016). Species delimitation studies using nucleotide sequence data revealed a high incidence of cryptic species in the nemertodermatid genera *Sterreria* and *Nemertinoides* (Meyer-Wachsmuth et al. 2014). This suggests that global diversity of nemertodermatids may be much higher than previously understood. There is no study analysing a large number of specimens of a nominal acoel species to detect cryptic species, but Kånneby et al. (2015) found three new species of *Archaphanostoma* with limited distribution ranges when sampling on the Swedish west coast where the acoelomorph fauna is comparatively well studied. Data mining of metabarcoding sequences from pelagial and deep benthic habitats demonstrated the presence of new species and even potentially early branching groups of acoels (Arroyo et al. 2016). Thus, there are strong indications of extensive unknown acoel diversity in the oceans.

In summary, knowledge of acoelomorph fauna is fragmentary with a strong bias towards the Scandinavian coasts, the North Sea area, western Mediterranean, Parts of the North American east coast and the Sao Paolo area. The actual number of acoelomorph species is likely to be orders of magnitude larger than the current number of nominal species.

14.3 Morphological Characters of the Xenacoelomorpha

Herein we give a brief review of the main features of xenacoelomorph morphology and discuss the evolution of the various character states. As starting points, we have taken two recent excellent reviews of the current data published by Achatz et al. (2013) and Hazsprunar (2015), which has provided tables of references to phenotypic studies within the Xenacoelomorpha, on gene expression methodologies and molecular data.

14.3.1 Epidermis

Xenacoelomorpha share a multiciliary epidermis without accessory centrioles (Ehlers 1985). Cilia in all three groups of Xenacoelomorpha have unique thinned ciliary tips considered a synapomorphy of the group (for details see Fig. 14.2a and

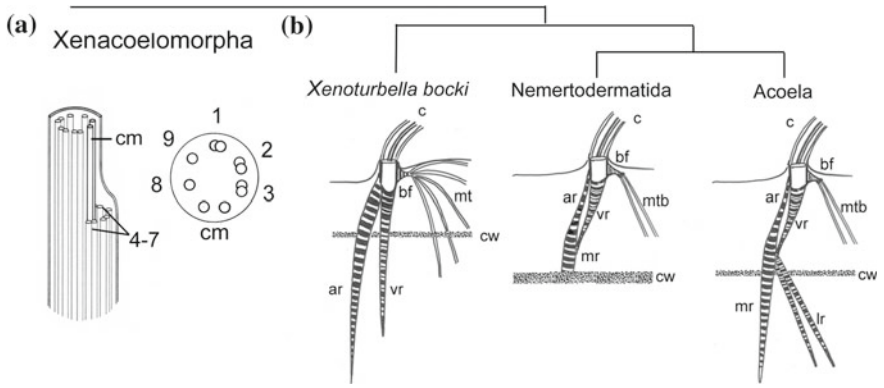


Fig. 14.2 **a** Schematic representation of synapomorphic for Xenacoelomorpha thinned ciliary tips with a distinctive shelf where the doublets 4–7 terminate and central microtubules (cm) continue in a peripheral position (Tyler 1979; Ehlers 1985; Smith and Tyler 1985a, b; Franzen and Afzelius 1987; Lundin 1997, 1998). **b** Hypothesis of rootlet system evolution within Xenacoelomorpha, from Raikova (1991) with modifications. In ***Xenoturbella bocki*** (Franzen and Afzelius 1987) cilia (c) have two rootlets, anterior (ar) and vertical (vr) both running almost vertically down. The basal body has a well-developed basal foot (bf) with microtubules (mt) fanning from it. The rootlets are interconnected on a single level by a horizontal fibrous cell web (cw). In **Nemertodermatida** (Tyler and Rieger 1977) each cilium has an anteriorly directed main rootlet (mr) that forms a knee-like bend and continues vertically down. The main rootlet is bipartite in its upper part, supposedly formed by two fused rootlets: anterior and vertical. The basal foot is linked by microtubule bundles (mtb) with the main rootlets of the two posteriorly adjacent cilia, thus forming an upper interconnection of the rootlet grid. The rootlets are also interconnected by cell web. In **Acoela** (Dorey 1965; Hendelberg and Hedlund 1974; Tyler 1979; Ehlers 1985; Rohde et al. 1988), in addition to the described pattern, there are paired lateral rootlets (lr), linked to the main rootlets of two posteriorly adjacent cilia, forming the lower interconnection of the grid

ref. in the figure caption). Epidermal cilia are anchored within the cytoplasm of epidermal cells by ciliary rootlets (for details see Fig. 14.2b and ref. therein). Cilia in *Xenoturbella bocki* have two almost vertical rootlets interconnected only by fibrous cell web. Acoelomorphs have a single main rootlet, bipartite in its proximal part, likely formed by fusion of the two rootlets found in *X. bocki* (Raikova 1991). In Acoelomorpha, rootlet system reaches great complexity with rootlets interconnected at two (in Nemertodermatida) or at three levels (in Acoela).

14.3.2 Musculature of the Body Wall and Pharynges

Changes in the muscular system pattern were shown to be phylogenetically informative especially in Acoelomorpha (Hooge 2001; Tyler and Hooge 2004; Jondelius et al. 2011; Meyer-Wachsmuth et al. 2013). *Xenoturbella bocki* has only two muscle layers in the body wall: outer circular and inner longitudinal (for details see Fig. 14.3 and ref. in the figure caption). Such orthogonal muscle grid is presumably

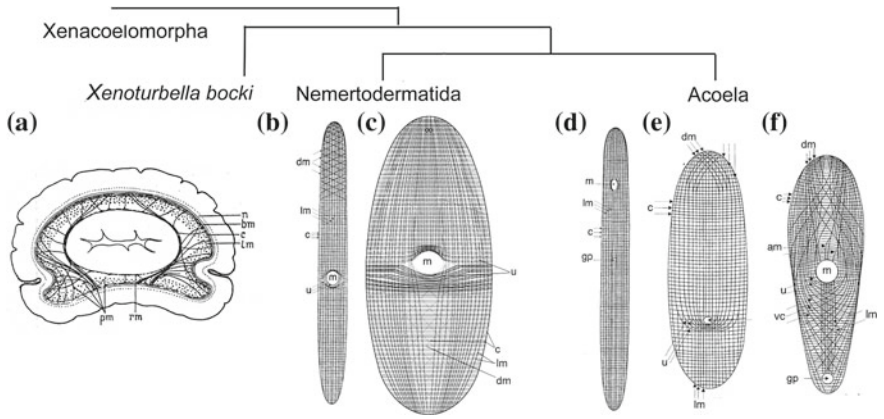


Fig. 14.3 Patterns of body wall musculature in Xenacoelomorpha. **a** *Xenoturbella bocki,* transverse section, from Westblad 1949b. The body wall is composed of an orthogonal grid of outer circular muscles and inner longitudinal muscles (Westblad 1949b; Sopott-Ehlers and Ehlers 1997; Raikova et al. 2000a). **b**–**f**—Schemes of arrangement of ventral body wall musculature from Hooge 2001 (**b**; **d**–**f**) and Meyer-Wachsmuth et al. 2013 **c**. Within Nemertodermatida, *Sterreria psammicola* **b** has mostly orthogonal body wall with few anterior diagonal muscles (Hooge 2001), while *Meara stichopi* **c** has three-layered body wall musculature with diagonal muscles and additional U-shaped muscles around the mouth Among lower acoels, representatives of early emerging clades Solenofilomorphidae and *Proporus* cf. *lonchitus* **d** have only an orthogonal muscle grid (Hooge 2001); Paratomellidae **e** have few anterior diagonal muscles and U-shaped around the mouth (Hooge 2001). The more derived groups of Acoela (Crucimusculata) have very complex musculature with cross-over muscles, especially in the most derived Mecynostomidae **f** (Hooge 2001). Abbreviations: am—accessory muscles; bm—basal lamina; c—circular muscles; dm—diagonal muscles; gp—gonopore; lm—longitudinal muscles; m—mouth; n—nerve net; pm—parenchymal muscles; rm—reticular muscles of the gut; u—U-shaped muscles; vc—ventral cross-over muscles

plesiomorphic for Bilateria (Hooge 2001; Schmidt-Rhaesa 2007). Development of more complex muscle patterns likely took place independently within Acoela and Nemertodermatida, as in each of these taxa cases of plesiomorphic orthogonal grid occur in early emerging clades, while the more derived ones have quite complex muscular patterns (for details see Fig. 14.3 and references therein). There is no evidence of secondary simplification of muscles in some lineages and it seems highly improbable in free-living predators. Parasitic species (like *Meara stichopi*) show even more complex muscle patterns than their free-living relatives (Raikova et al. 2016). Pharynges are absent in *Xenoturbella* and nemertodermatids, but present in many early and mid-level acoel groups (Diopisthoporidae, Proporidae, Hofsteniidae, Solenofilomorphidae, Isodiametridae), where their presence compensates for the absence of cross-over muscles (Hooge 2001; Jondelius et al. 2011). Todt (2009) has demonstrated that different versions of pharynx simplex differ considerably in muscle patterns and speculated that they likely had developed independently.

14.3.3 Glands

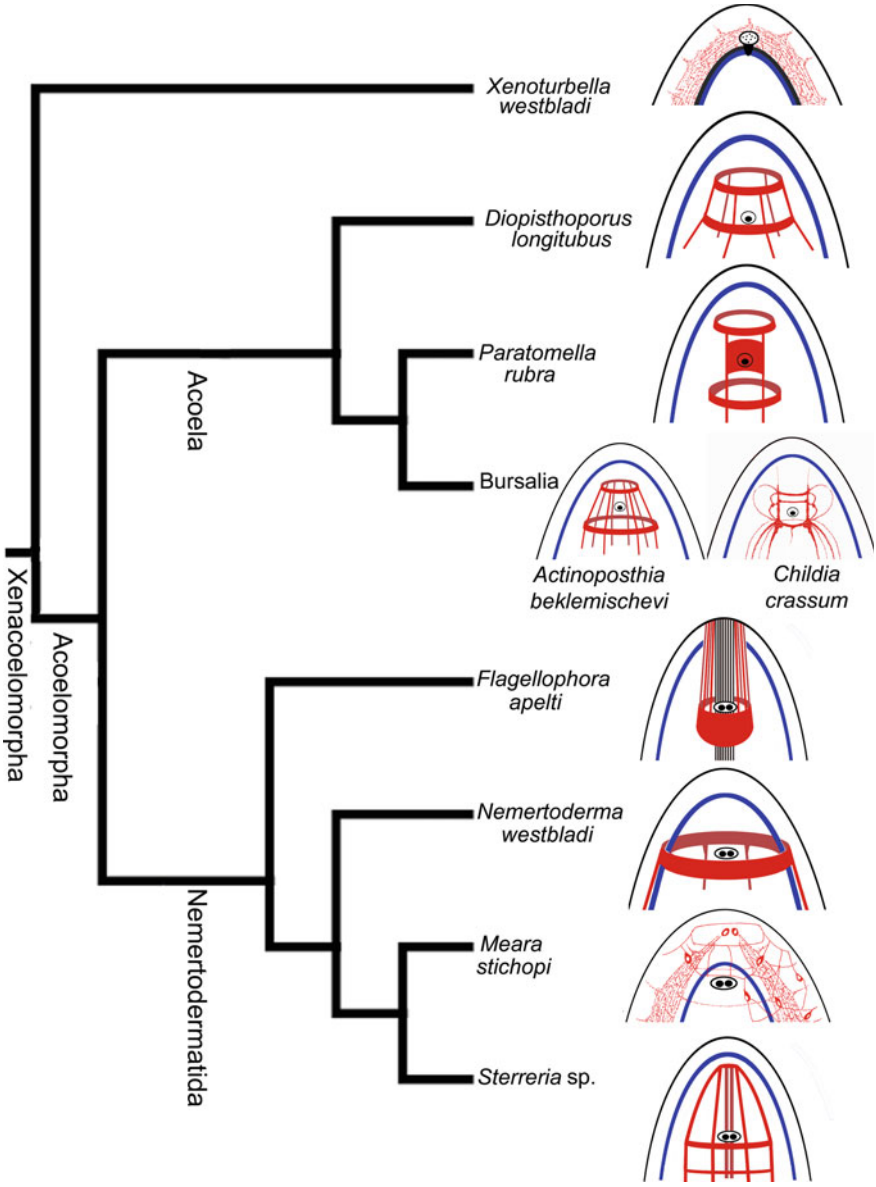
Glands in Xenacoelomorpha are diverse and abundant, especially those producing mucus for mucociliary gliding. Mucous glands form the newly described ventral glandular network in *Xenoturbella* species (Rouse et al. 2016; Nakano et al. 2017). In Nemertodermatida numerous mucous glands are concentrated on the frontal pole of the animal opening separately, which is interpreted as a plesiomorphic condition (Smith and Tyler 1988; Ehlers 1992c). The same condition occurs in the early emerging taxa of Acoela: *Diopisthoporus*, (Smith and Tyler 1985a, b), *Paratomella* (Ehlers 1992a, b, c) and *Afronta* (Proporidae) (Rieger et al. 1991). In the rest of acoels, the glandular ducts of the frontal glands discharge together through a common pore at the anterior pole of the body forming the "frontal organ" (Smith and Tyler 1985a, b; Smith et al. 1986; Klauser et al. 1986; Ehlers 1992b, c).

14.3.4 Receptors

Sensillae are bipolar or multipolar nerve cells with dendrites bearing one or rarely 2 and more cilia and sometimes, in "collar receptors", a collar of thickened microvilli, called "stereocilia". Sensillae were studied by electron microscopy mostly in acoels and classified on morphological basis (Bedini et al. 1973; Todt and Tyler 2006; Bery et al. 2010; Zabotin 2019). In acoels up to five different types of sensillae are usually present in various combinations. The simplest receptor type, likely plesiomorphic, is that with a long and slender ciliary rootlet—and only this type was so far described in *Xenoturbella* (Raikova et al. 2000a). It is also present in Nemertodermatida (Smith and Tyler 1988; Ehlers 1992c). Collar receptors are completely lacking in Xenoturbellida, Nemertodermatida or in lower acoels, suggesting an independent origin of collar receptors within the acoels and in other invertebrates (Todt and Tyler 2006).

Eyes. Xenoturbellida and Nemertodermatida are devoid of eyes. Among acoels, epidermal eyespots occur in some Isodiametridae (Lanfranchi 1990), and very simple eyes are described in Convolutidae (Popova and Mamkaev 1985; Yamasu 1991). These structures show quite different organization and likely evolved independently in originally eyeless Acoela.

Statocysts. Representatives of all three groups of Xenacoelomorpha have anteriorly situated statocysts thoroughly described and discussed by Ehlers (1991). In *Xenoturbella bocki,* the statocyst is intraepidermal; it lies at the anterior pole anchored in the basal lamina and surrounded by thick basiepidermal nerve net. Acoelomorph statocysts are submuscular and lie within the brain, more specifically surrounded by statocyst ganglion. All xenacoelomorph statocysts have a capsule composed of extracellular matrix and inlaid on the inside by parietal cells. Inside the capsule, there are freely movable lithocyte cells with intracellular statoliths. In *X. bocki,* the lithocytes are numerous and have flagella, as do the parietal cells (Ehlers 1991). In Nemertodermatida, there are always two lithocytes (Fig. 14.1a), and a few parietal


Xenacoelomorpha
Acoelomorpha
Acoela
Nemertodermatida
Xenoturbella westbladi
Diopisthoporus longitubus
Paratomella rubra
Bursalia
Actinoposthia beklemischevi
Childia crassum
Flagellophora apelti
Nemertoderma westbladi
Meara stichopi
Sterreria sp.

◀**Fig. 14.4** Schematic drawings of the organization of the nervous system (NS) and its position with respect to the body wall musculature (BWM) in Xenacoelomorpha. For all species only the anterior part is shown, statocysts and broom organ (*Flagellophora*) are shown for morphological context. The NS is shown in red, the layers of BWM in blue. *Xenoturbella* possesses the hypothesized ancestral basiepidermal nerve net without anterior centralization. In Nemertodermatida, the NS is diverse and does not show a trend from an epidermal nerve net towards an insunk centralized lobed brain. It lies outside the BWM in *N. westbladi* and *M. stichopi* and inside the BWM in *F. cf. apelti* and *Sterreria* sp. *Meara stichopi* has no brain, but only a pair of basiepidermal lateral neurite bundles. *Nemertoderma westbladi* has a basiepidermal brain ring and a pair of ventrolateral neurite bundles. *Sterreria* sp. shows a commissural-like brain similar to that observed in Acoela and several neurite bundles going in frontal and caudal directions. *Flagellophora apelti* has a neuropile, located at the level of the statocyst, providing innervation for the broom organ. In Acoela, the NS is mostly submuscular with anterior centralisations of different degrees ranging from two rings in basal taxa to more complex structures with several commissures and paired peptidergic marker neurons. Figure adapted from Achatz and Martinez (2012). Nemertodermatid phylogeny is based on Meyer-Wachsmuth and Jondelius (2016), that of Acoela on Jondelius et al. (2011). Morphological information based on the following studies: *Xenoturbella westbladi* (Westblad 1949b; Raikova et al. 2000a); *Diopisthoporus longitubus* (Raikova 2004); *Paratomella rubra* (Crezée 1978); *Actinoposthia beklemischevi*: (Raikova et al. 1998; Raikova 2004)*; Childia crassum* (Reuter et al. 2001a, b, Raikova et al. 2004b); *Nemertoderma westbladi (*Raikova et al. 2004a and Raikova et al. 2016); *Meara stichopi* (Raikova et al. 2000b; Børve and Hejnol 2014), *Flagellophora cf. apelti* and *Sterreria* spp. (Raikova et al. 2016)

cells with ventral nuclei; while in Acoela there is always one single lithocyte and two parietal cells with dorsal nuclei. No flagella are encountered within the statocysts of Acoelomorpha (Ferrero 1973; Ehlers 1985, 1991). Ehlers (1991) proves that statocysts of Xenacoelomorpha are non-homologous neither between each other, nor with those found in Platyhelminthes.

14.3.5 Nervous System

The nervous system (NS) patterns in xenacoelomorphs are extensively studied (see list of references in Hazsprunar (2015), or Martinez et al. (2017) and in the more recent of Dittmann et al. (2018). Here we give a brief summary of the currently available data (for details see Fig. 14.4 and ref. therein).

Xenoturbella bocki has a very simple NS, consisting solely of a basiepidermal nerve net lying above the thick basal lamina with no brain detected. Nemertodermatids have a very plastic NS, either basiepidermal or submuscular with or without a brain (for details see Fig. 14.4). Acoels show a considerable variety in the details of NS patterns (for details see Fig. 14.4), but some common features can be found. According to Raikova (2004), acoel brain has two parts: (1) a 5-HT- or RF-amide immunoreactive "commissural brain" (Raikova et al. 1998) more often composed of one to several rings, with paired longitudinal cords (or neurite bundles according to terminology proposed by Richter et al. 2010) starting caudal from it and (2) a

ganglion surrounding the statocyst. Unlike the commissural brain, the statocyst ganglion shows no 5-HT- or RF-amide like immunoreactivity; therefore, it is considered non-homologous with "true" cerebral ganglia of Nephrozoa and a synapomorphy of Acoelomorpha. Martinez et al (2017) disputed the use of "commissural brain" terminology, at least for the most recent clades of Acoela, due to the presence of a compact brain in convolutids, with a periphery occupied by the cell bodies and a dense internal neuropile. The acoel NS is currently being extensively studied in species of Convolutidae (Gaerber et al. 2007; Semmler et al. 2010; Bery et al. 2010; Bery and Martínez 2011; Gavilán et al. 2016; Martinez et al. 2017) or Isodiametridae (Achatz and Martinez 2012; Dittmann et al. 2018) as model animals. Moreover, the acoel *S. roscoffensis* is nowadays used to test the functionality of the NS given the fact that adult worms are able to regenerate a whole head from scratch (Sprecher et al. 2015; Arboleda et al. 2018).

The earliest bilaterian NS is thought to consist of a diffuse subepidermal nerve net similar to that of cnidarians (i.e. Gröger and Schmid 2001). Reisinger (1925) proposed that in the course of evolution nerve elements become more concentrated and the NS sinks below the muscle layers. According to this hypothesis, the basiepidermal NS of *Xenoturbella*, *Meara* and *Nemertoderma* would represent the ancestral state, while the submuscular brains of nemertodermatids and acoels would be apomorphic (Westblad 1948; Rieger et al. 1991). Under this hypothesis, a submuscular nervous system evolved multiple times independently in Acoelomorpha (Raikova et al. 2016).

A common for most animals stomatogastric nervous system is absent in all xenacoelomorphs (Westblad 1937, 1949b; Raikova et al. 2000a, b; Raikova 2004). It seems to be a symplesiomorphy of these groups, as a loss of stomatogastric system in primitively predatory organisms is unlikely (Haszprunar 2015; Gavilán et al. 2019).

14.3.6 Digestive System

Mouth is situated mid-ventrally in *Xenoturbella* and nemertodermatids while in acoels mouth position shows great variability. Most commonly it is situated mid-ventrally, sometimes subfrontally, but in the basalmost acoel taxon Diopisthoporidae the mouth opens at the posterior end. Beklemischev (1963) pointed out that the mouth in acoels was homologous to blastopore which makes its posterior position primitive (plesiomorphic). *Xenoturbella* and nemertodermatids have a saccular blind gut with amoeboid gastrodermis (Westblad 1949b, 1949a) that seems to be plesiomorphic for Xenacoelomorpha. Some acoels have cellular amoeboid digestive parenchyma (Smith and Tyler 1985a, b; Ehlers 1992b), but most often there is a central digestive syncytium (Rieger et al. 1991). For a recent review on the digestive system of xenacoelomorphs, check Gavilán and collaborators recent review (Gavilán et al. 2019).

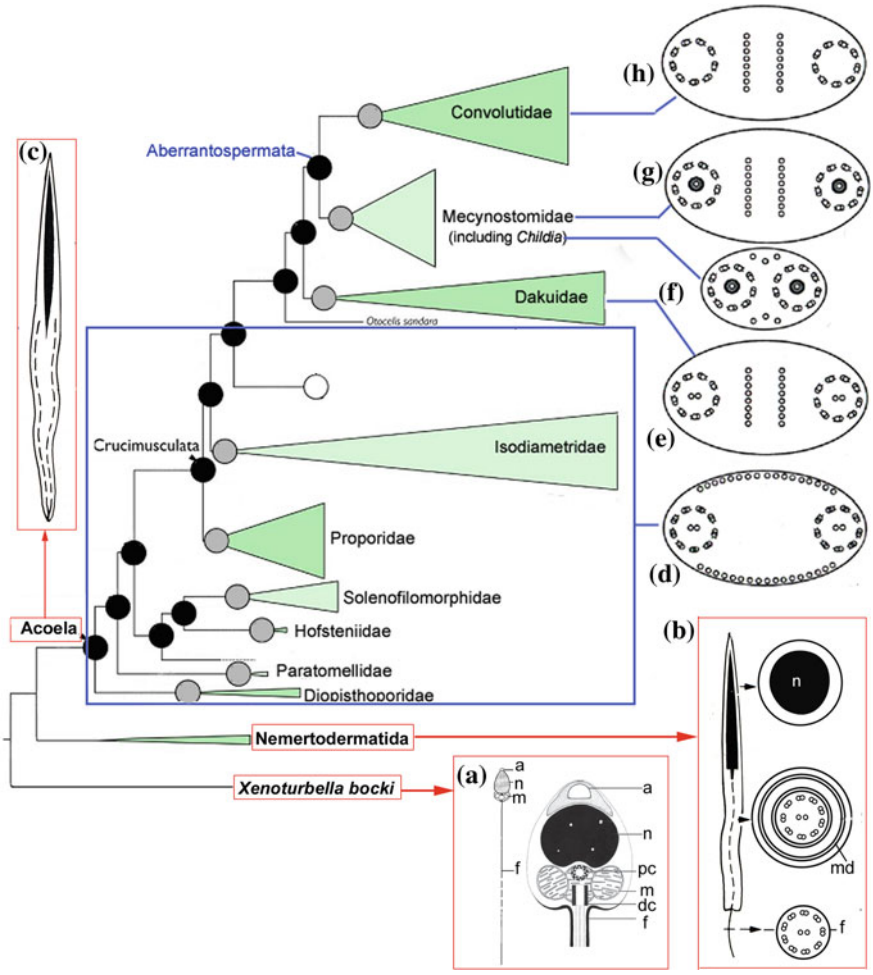
14.3.7 Reproductive Organs

The Acoelomorpha are simultaneous or protandric hermaphrodites suggesting hermaphroditic Urbilateria as likely (Haszprunar 2015). *Xenoturbella* lacks copulatory organs entirely (Westblad 1949b). Nemertodermatids have only a seminal vesicle opening subterminally. Acoels display a striking variety of copulatory organs. Most are not yet sufficiently investigated to be properly classified, especially the copulatory stylets (Tekle et al. 2007a, b; Achatz et al. 2013).

The agonadial condition with free eggs and sperm found in *Xenoturbella* and in several taxa of Acoelomorpha is presumably plesiomorphic (Rieger et al. 1991). The oocytes in *Xenoturbella* and Acoelomorpha are entolecithal with non-sclerotized egg envelope which seem to be plesiomorphic features (Falleni et al. 1995; Raikova et al. 1995 and ref. therein).

The three taxa of Xenacoelomorpha radically differ by their sperm structure. *Xenoturbella bocki* (Fig. 14.5a) has a primitive monoflagellate sperm with a roundish head and a free 9 + 2 flagellum that reflects ectaquatic fertilization (Obst et al. 2011). Nemertodermatida (Fig. 14.5b) have sperm with a single 9 + 2 flagellum starting from the nucleus, a condition clearly plesiomorphic for Acoelomorpha (Tyler and Rieger 1975). However, nemertodermatid sperm is filiform and highly modified in order to penetrate tissues (introsperm). Acoels (Fig. 14.5c) have aberrant filiform biflagellate sperm with incorporated flagella in inverted position and supporting microtubules (cortical or axial) (Hendelberg 1969, 1977, 1986). In Acoela, three configurations of sperm axonemes are described: 9 + 2; 9 + 0 and 9 + 1 (for details see Fig. 14.5d–h and ref. therein). It should be emphasized that the aberrant 9 + 1 configuration where only one central microtubule remains surrounded by a halo of electron dense material (Tekle et al. 2007a, b), found in acoel family Mecynostomidae, is non-homologous with 9 + "1" configuration in flatworm taxon Trepaxonemata, as the central cylinder ("1") in Trepaxonemata does not contain tubulin (Iomini et al. 1995). Sperm characters in Acoela proved to be those most concordant with molecular phylogeny (Petrov et al. 2004; Jondelius et al. 2011); thus, they were extensively studied (see references in Haszprunar (2015).

Recently data on the sperm of Xenacoelomorpha have been re-evaluated by Buckland-Nicks et al. (2018). The authors propose a very interesting hypothesis that, assuming Xenacoelomorph monophyly and ancestral introsperm in this taxon, the re-expression of the aquasperm form could have happened in *Xenoturbella* as a secondarily derived state through "progenetic spermiogenesis" with the precocious development of round spermatids to maturity.


(h)
Convolutidae
(g)
Aberrantospermata
Mecynostomidae
(including Childia)
(f)
Dakuidae
Otocelis sondgra
(e)
Crucimusculata
Isodiametridae
(d)
Proporidae
Solenofilomorphidae
Hofsteniidae
Paratomellidae
Diopisthoporidae
(c)
Acoela
(b)
Nemertodermatida
Xenoturbella bocki
(a)
a
n
m
f
pc
dc
md

◀**Fig. 14.5** Schematic drawings of the sperm morphology in Xenacoelomorpha. **a** *Xenoturbella bocki*. Primitive monoflagellate sperm in with a roundish head and a free 9 + 2 flagellum after Westblad (1949a) and Obst et al. (2011). **b** Nemertodermatida modified filiform monoflagellate sperm with long proximal nucleus, mitochondrial derivatives and supporting microtubules, after Hendelberg (1986), Lundin and Hendelberg (1998), Boone et al. (2011), Buckland-Nicks et al. (2018). **c**. **Acoela**. Aberrant filiform biflagellate sperm with incorporated flagella in inverted position (with basal bodies at the distal end of the cell), after Hendelberg (1986). **d–h** Cross sections of the sperm shaft in different acoel groups showing axoneme configuration and the arrangement of cytoplasmic microtubules. **d** Plesiomorphic axoneme configuration with nine peripheral microtubule doublets and two central microtubules (9 + 2), characteristic for all basal and mid-level taxa of the Acoela (Petrov et al. 2004; Jondelius et al. 2011). Cortical (positioned under the plasma membrane) cytoplasmic microtubules are also plesiomorphic. **e**. Sperm with 9 + 2 axonemes and axial (running along the main axis of the sperm) cytoplasmic microtubules, described in *Philactinoposthia saliens* (Raikova and Justine 1999), *Philocelis brueggemanni* and *Daku woorimensis* (Petrov et al. 2004). **f**. Sperm in *Childia* spp. (Mecynostomidae) with 9 + 1 axonemes and distal microtubules found only in the distal region of the sperm (Tekle et al. 2007a, b). **g**. Sperm in other Mecynostomidae with 9 + 1 axonemes and axial microtubules (Petrov et al. 2004; Tekle et al. 2007a, b). **h**. Sperm in Convolutidae with 9 + 0 axonemes and axial microtubules. Note that Mecynostomidae and Convolutidae due to their aberrant sperm axonemes are united into Aberrantospermata group. The phylogenetic tree used is based on Jondelius et al. (2011) and Hejnol et al. (2009), but simplified. Abbreviations: a—acrosome; a—flagellum; dc—distal centriole; m—mitochondrion; md—mitochondrial derivative; n—nucleus; pc—proximal centriole

14.4 Molecular Regulation in Xenacoelomorphs

The characterization of xenacoelomorph genomes is still in its infancy. Our current knowledge derives, fundamentally, from the isolation of single genes using classical PCR screens. However, and mostly owing to the needs of having large data sets for phylogenomic analysis, several transcriptomes have been obtained over the last number of years. This provides a very rich source of data for downstream characterization of specific genes/gene families, though at present, the data is, with this purpose, very seldom used. Moreover, we have generated data that represents a wide array of species, most of which are unsuitable for developmental studies given the paucity with which they are found and the undeveloped systems for their culture. Most developmental studies have been focused on just a few species, for which we are starting to gather some critical data on the use of regulatory genes and certain families of differentiation markers.

14.4.1 Xenacolomorph Genomes and Transcriptomes

Before proceeding further, let us recapitulate what has been learned in terms of the molecular composition of different xenacoelomorph genomes and transcriptomes. Table 14.1 provides a comprehensive list of our current knowledge.

Table 14.1 Summary of the current genomic and transcriptomic resources available for different xenacoelomorph species

Species	Family	Data source	References
Acoela			
Childia submaculatum	Mecynostomidae	Transcriptome	Cannon et al. (2016)
Convolutriloba macropyga	Convolutidae	Transcriptome	Cannon et al. (2016)
Diopisthoporus gymnopharyngeus	Diopisthoporidae	Transcriptome	Cannon et al. (2016)
Diopisthoporus longitubus	Diopisthoporidae	Transcriptome	Cannon et al. (2016)
Eumecynostomum macrobursalium	Mecynostomidae	Transcriptome	Cannon et al. (2016)
Hofstenia miamia	Hofsteniidae	Transcriptome	Cannon et al. (2016)
Isodiametra pulchra	Isodiametridae	Transcriptome	Cannon et al. (2016)
Symsagittifera roscoffensis	Convolutidae	Mitochondrial genome	Mwinyi et al. (2010)
Isodiametra pulchra	Isodiametridae	Mitochondrial genome	Robertson et al. (2017)
Paratomella rubra	Paratomellidae	Mitochondrial genome (partial)	Ruiz-Trillo et al. (2004)
Paratomella rubra	Paratomellidae	Mitochondrial genome	Robertson et al. (2017)
Archaphanostoma ylvae	Isodiametridae	Mitochondrial genome	Robertson et al. (2017)
Symsagittifera roscoffensis	Convolutidae	Transcriptome (mix stage embryos)	Philippe et al. (2019), Perea-Atienza et al. (2015)
Symsagittifera roscoffensis	Convolutidae	Draft genome	Philippe et al. (2019), Perea-Atienza et al. (2015)
Hofstenia miamia	Hofsteniidae	Transcriptome (embryos + reg. animals)	Srivastava et al. (2014)
Isodiametra pulchra	Isodiametridae	Transcriptome (adults)	Brauchle et al. (2018)
Symsagittifera roscoffensis	Convolutidae	Transcriptome (juveniles)	Brauchle et al. (2018)
Hofstenia miamia	Hofsteniidae	Genome (complete)	Gehrke et al. (2019)
Hofstenia miamia	Hofsteniidae	Transcriptome (adults)	Gehrke et al. (2019)
Praesagittifera naikaiensis	Convolutidae	Genome (complete)	Arimoto et al. (2019)

(continued)

Table 14.1 (continued)

Species	Family	Data source	References
Nemertodermatida			
Ascoparia sp.	Nemertodermatidae	Transcriptome	Cannon et al. (2016)
Meara stichopi	Nemertodermatidae	Transcriptome	Cannon et al. (2016)
Nemertoderma westbladi	Nemertodermatidae	Transcriptome	Cannon et al. (2016)
Sterria sp.	Nemertodermatidae	Transcriptome	Cannon et al. (2016)
Nemertoderma westbladi	Nemertodermatidae	Mitochondrial genome (partial)	Ruiz-Trillo et al. (2004)
Xenoturbellida			
Xenoturbella bocki	Xenoturbellidae	Transcriptome (adults)	Cannon et al. (2016)
Xenoturbella profunda	Xenoturbellidae	Transcriptome (adults)	Rouse et al. (2016)
Xenoturbella monstrosa	Xenoturbellidae	Mitochondrial genome	Rouse et al. (2016)
Xenoturbella hollandorum	Xenoturbellidae	Mitochondrial genome	Rouse et al. (2016)
Xenoturbella churro	Xenoturbellidae	Mitochondrial genome	Rouse et al. (2016)
Xenoturbella bocki	Xenoturbellidae	Draft genome	Philippe et al. (2019); Perea-Atienza et al. (2015)
Xenoturbella bocki	Xenoturbellidae	Transcriptome (adults)	Brauchle et al. (2018)

As mentioned before, the first characterization of xenacoelomorph genes was derived from the need for obtaining molecular markers to infer phylogenetic relationships, whether they were for understanding the position of different Xenacoelomorpha clades within the Metazoa or inferring the internal phylogeny of those clades. In this context, it was, historically, natural that cloning ribosomal (i.e. 18S) and mitochondrial genes were taken as a priority (see the earlier sections for details). The advent of powerful phylogenomic methodologies prompted the use of transcriptomic and genomic data for the inference of clade relationships. With this in mind, the studies of Hejnol et al. (2009), Philippe et al. (2011) and Cannon et al. (2016) were pioneers (see Sect. 2.1 for more details). They used, mostly Cannon and collaborators, large sets of xenacoelomorph transcriptomes, including a wide range of species from various xenacoelomorph families. The species used in these studies are all found in Table 14.1, which summarizes all the species for which there is genomic or transcriptomic data.

Very little is known about the structure and composition of xenacoelomorph genomes. Only two complete genomes, with long scaffolds, have been produced,

that of the early divergent acoel *Hofstenia miamia* (Gehrke et al. 2019) and that of the more recently diverged species *Praesagittifera naikaiensis* (Arimoto et al. 2019). *H. miamia* has a genome with a size of 950 Mb, encoding for (predicted and supported by transcriptome data) 22,632 genes in an environment of 53% of repetitive sequences, mostly derived from transposons. *P. naikaiensis* genome is a bit smaller, approx. 656 Mb, with 70% of the sequence being repetitive elements, though encoding for 22,143 genes (a number supported also by transcriptome data). Other genomic data, in relatively small scaffolds has been obtained for the genomes of the xenoturbellid, *Xenoturbella bocki*, and the acoel, *Symsagittifera roscoffensis* (Philippe et al. 2019). The complete descriptions of those last genomes are, however, still missing. Of the general characteristics, only the sizes (~150 Mb for *X. bocki* and ~1.4 Gb for *S. roscoffensis*) are known, plus the fact that those genomes are full of repetitive elements, accounting for the large size of the *S. roscoffensis* genome, which is nearly half the size of the human genome. Several features of the genome organization of acoels were known, through the use of technologies that include the development of BAC libraries and chromosomal mapping. These were instrumental in understanding that the Hox genes in acoels were dispersed, located on different chromosomes (Moreno et al. 2009). Unfortunately, the potential of these technologies was not explored any further.

In the absence of xenacoelomorph genomes, large-scale analysis of long fragments has been conducted in a few acoelomorphs targeting mitochondrial genomes, shorter gene complexes easier to sequence. Complete or substantial fractions of mitochondrial genomes are known for one nemertodermatid and four acoels (Table 14.1 and: (Ruiz-Trillo et al. 2004; Mwinyi et al. 2010; Robertson et al. 2017). Strikingly, all these reports indicate that the arrangement of genes within these mitochondrial genomes is very different from what has been observed in other metazoans (and with each other), pointing towards a clade-specific set of rearrangements.

Returning to draft genomes, we need to stress that independent of the completeness they represent, in the current state, they have facilitated the characterization of the size and diversity of certain gene families in these two species, for instance, those of the transcriptional regulators, bHLH or Hox, the signalling Wnt family of ligands and the complete set of GPCR receptors (Perea-Atienza et al. 2015; Gavilán et al. 2016). Other families have been characterized through the deep analysis of transcriptomes. In addition to the previous ones (for which we have genomic and transcriptomic data), the whole complement of homeobox-containing families (11 in total) have been identified in *X. bocki* and the two acoels, *S. roscoffensis* and *Isodiametra pulchra* (Brauchle et al. 2018), or the GPCR complements of different xenacoelomorphs (Thiel et al. 2018). In addition, the latter study of transcriptomes carried out by Thiel and colleagues has permitted a comprehensive description of neuropeptide evolution in this phylum.

Before ending, this section is important to emphasize that the phylogenetic position of Xenacoelomorpha allows us to study with detail those genomic novelties associated to the origin and diversification of bilaterian animals and how the activities of those new or lost genes affect the evolutionary transformations of body plans. The important paper of Paps and Holland (2018) describes the particularities of the

birth and loss of gene groups in the metazoans (see their Fig. 14.1 for a graphical view). These authors have shown that the emergence of multicellular animals was accompanied by the incorporation of many genomic novelties (1189 gene homology groups; HG). Other, more recent phylogenetic nodes, such as that representing the last common ancestor of Cnidaria + Bilateria (Planulozoa) and the Bilateria, are also characterized by the incorporation of 1201 and 1580 new HG, respectively. In fact, these nodes are also characterized by some specific HG loses, here 72 and 520. Many of these genes encode for regulatory functions (transcription factors or signalling molecules). Since no comprehensive expression analysis has been done for those specific genes in most metazoan clades, it becomes quite obvious that the genomic underpinnings of the emergence of bilaterian animals (including the early diversification of xenacoelomorphs) are, mostly, a matter of speculation. The study of those regulatory functions should become a major future focus of analysis in the Xenacoelomorpha because they will illuminate many aspects of the origin of bilaterians.

14.4.2 Gene Activity in Xenacoelomorphs

Transcriptomes are ideal sources for the identification of particular genes, avoiding the cumbersome procedure of PCR screens. They can provide sequence information necessary to generate probes for in situ hybridization, thus providing us with entry to viewing particular developmental or physiological processes. The exploitation of transcriptomes for this purpose has been instrumental in identifying genes involved in dorsoventral patterning (Srivastava et al. 2014), mesoderm specification (Chiodin et al. 2013) and stem cell biology (De Mulder et al. 2009) in different acoels or the study of excretion processes in acoelomorphs (Andrikou et al. 2019). All these studies have provided the necessary tools to understand the molecular underpinnings of those processes at a level that was impossible previously using more conventional, gene-by-gene cloning methods (i.e. Ramachandra et al. 2002; Cook et al. 2004; Moreno et al. 2009).

The identification of all those genes plus the development of colorimetric or fluorescent in situ methodologies has supplied a basic understanding of developmental processes. It is true that, in comparison with well-established model systems, our knowledge of xenacoelomorph development, from the molecular point of view, is rather limited. The paucity of functional tools (see the following) and a detailed description of embryogenesis have hampered progress in this area. Nevertheless, it is fair to stress here that in some areas, our knowledge is more detailed than that we have for many other phyla. A clear case would be our characterization of the patterning mechanisms employed along the two major body axes (AP and DV). The genes involved in both processes were of primary interest as it was revealed that acoelomorphs were the first offshoot of Bilateria. It was clear that in order to determine how bilateral animals acquired their specific symmetry, we needed to know

how acoelomorphs specified their orthogonal axis. Hox genes were the primary candidates as they were used as a vectorial system to specify the position of structures along the AP axis across all bilaterians. The study of different acoels uncovered early on that the complement of Hox genes was quite reduced, numbering at only three, one orthologous to each of the so-called Anterior, Central and Posterior classes. Studies of *Convolutriloba longifissura* (Hejnol and Martindale 2009) and *Symsagittifera roscoffensis* (Moreno et al. 2009) exhibited a pattern of staggered domains for the three genes along the major body axis, in agreement with what is observed in other bilaterians. The expression of those genes was initiated during embryogenesis, just after gastrulation (Hejnol and Martindale 2009) and resolved in the staggered pattern seen in hatchlings and juveniles. In budding specimens of *Convolutriloba retrogemma*, it has also been shown that during the budding process (that involves a reversal of polarity in new animals), expression along the bud also conforms to the expected AP Hox pattern (Sikes and Bely 2010). Interestingly, the staggered pattern is not the result of chromosomal clustering (Moreno et al. 2009), just as in other bilaterian animals. However, it seems clear clustering is not necessary for the establishment of a staggered expression pattern (i.e. Seo et al. 2004). Other regulators of the AP axis are the Wnt ligands, and in the acoel *Hofstenia miamia*, they are used in the same process (Srivastava et al. 2014). In regenerating *Convolutriloba retrogemma*, the chemical inhibition of the Wnt pathway also affects the polarity of the regenerates (Sikes and Bely 2010). The role of Hox and Wnt ligands in the specification of posterior structures has been further explored through the use of RNAi technologies in the acoels, *Isodiametra pulchra* (Moreno et al. 2010) and *Hofstenia miamia* (Srivastava et al. 2014). These represent two of the very few functional studies carried out for any xenacoelormoph; in fact, we only have data obtained for acoels. The DV axis of bilaterian animals is specified through the BMP pathway. In the case of *H. miamia*, the interplay between the ligands, BMP and ADMP, seems to specify the DV axis, as probed through the use of RNAi technology (Srivastava et al. 2014). Additional data on the expression of BMP-ADMP genes has been obtained from the study of the nemertodermatid, *Meara stichopi*, and the acoel, *Isodiametra pulchra* (Martín-Durán et al. 2018). These ligands are also present in other xenacoelomorphs, such as *S. roscoffensis* or in *X. bocki* (Chang et al. 2015), though their expression or function has not been evaluated.

A cautionary note before proceeding all the functional experiments using RNAi carried out in acoels have been conducted in regenerating adults. No embryo RNAi methods have been developed to date.

During animal development, the establishment of the major body axis is followed by a regionalization process followed (or in concert) with the specification of the major tissues. With acoels, for the most part, much effort has been devoted to characterizing the molecular players involved in the specification of tissues, with a major focus placed on understanding mesodermal and neural developmental mechanisms. The mesodermal derivatives in acoels have been studied, fundamentally, in two species—*Isodiametra pulchra* and, to a lesser degree, *Symsagittifera roscoffensis*. While in the latter, the study of the musculature, and its structural components, has been assessed (Chiodin et al. 2011), in the case of *I. pulchra*, there is a relatively

complete understanding of how different mesodermal derivatives are molecularly regulated (Chiodin et al. 2013). The study of Chiodin and collaborators demonstrated that the mesoderm in acoels is composed of a few different cell types and that the gonads and stem cells are all derived, most probably, from the endomesoderm. The specific role of the stem cell system has been explored in the past (Gschwentner et al. 2001), but it is through the more detailed studies of De Mulder et al. (2009) that we came to understand various molecular players. In particular, we learned that the piwi orthologues (there are two in *I. pulchra*) are involved in specifying/maintaining the stem cell state. This has been proven through the use of RNAi methods in regenerating animals.

The development of the nervous system in acoels has been evaluated through the identification and characterization of candidate genes belonging to different transcription factor families. A major focus has been allocated to the identification of the so-called neurogenic genes of the bHLH family. The characterization of their expression patterns has been carried out in embryos and hatchlings of the species, *Symsagittifera roscoffensis* (Perea-Atienza et al. 2018). Interestingly, the different members of this family seem to have quite different spatio-temporal patterns of expression, with some of them being specifically restricted to neurogenic areas (from early on in embryogenesis). Their sometimes overlapping domains hint at the possibility of a combinatorial code of genes responsible for the specification of different neural domains or phenotypes. In addition to this family, other genes involved in neurogenesis have been characterized, though sometimes in the context of exploring other processes. This is the case of SoxB relatives in *Convolutriloba longifissura* (Hejnol and Martindale 2008) or in *S. roscoffensis* (Semmler et al. 2010). In both cases, the domain of expression of SoxB is compatible with their use in the specification of the neurogenic domain. Other regulatory genes expressed in presumptive neural tissue have been described in *C. longiffisura* (Cdx, Otp and NK2.1; Hejnol and Martindale 2008)) or *Neochildia fusca* (Brn1 and Brn3; (Ramachandra et al. 2002)). Very little is known of the regulation of nemertodermatid nervous system formation. However, the expression domains of certain key regulatory genes have been recently reported by Martin-Duran and colleagues utilizing *Meara stichopi* (Martín-Durán et al. 2018).

In spite of the knowledge gathered over the last number of years on the molecular components that regulate certain basic processes in xenacoelomorphs, it is clear that what needs to be done in order to "understand" both the development and physiology of this group of animals represent still an enormous task. This would be one that should involve the generation of more, and better, sequence data and the incorporation of novel technologies to assess gene function.

14.5 Future Perspectives

The realization that Xenacoelomorpha represent a clade of early divergent bilaterian has prompted a renewed interest in its study. While the collection and initial morphological characterization of different species within this clade date to more than

a century ago, it has been the introduction of molecular technologies that have revolutionized their modern study. Large gene data sets have allowed, on the one side, to reanalyse the phylogenetic affinities of Xenacoelomorpha, and its constitutive clades, and on the other, to initiate the characterization of the regulatory functions of many genes. While the functional analysis of these genes are still in its infancy, the promise of new technologies such as single-cell sequencing, CRISPR, high throughput genomic and proteomic analysis, should bring the study of these animals to a new level. However, it is important to note here that in absence of a careful analysis of the morphology and developmental processes, as well as a detailed understanding of the diversity of this group, molecular data per se will be of little value. Understanding how xenacoelomorphs have diversified over time and how the patterns of divergence are reflected in changes in developmental processes will need of a continuous investigation of the interplay between morphology changes, development and evolutionary diversification (Martinez 2018). If we follow this recommendation, we anticipate that the study of Xenacoelomorpha will become a key research program geared at a better understanding of the origin and diversification of Bilateria.

Acknowledgements Funding from The Swedish Research Council (project 2018-05191) is gratefully acknowledged by Ulf Jondelius. The work of Olga Raikova was supported by the Ministry of Education and Science of the Russian Federation (project no. AAAA-A19-119020690076-7) and the RFBR (project numbers 16-04-00593a and 20-04-01006a). We would also like to thank Dr. Pierre Pontarotti (Marseille) for organizing the yearly "Evolutionary Biology Meeting" in Marseille and for inviting us to submit this chapter.

References

Achatz JG, Martinez P (2012) The nervous system of *Isodiametra pulchra* (Acoela) with a discussion on the neuroanatomy of the Xenacoelomorpha and its evolutionary implications. Frontiers Zool 9. https://doi.org/10.1186/1742-9994-9-27

Achatz JG, Chiodin M, Salvenmoser W, Tyler S, Martinez P (2013) The Acoela: On Their Kind and Kinships, Especially with Nemertodermatids and Xenoturbellids (Bilateria Incertae Sedis). Org Diversity Evol 13(2):267–286. https://doi.org/10.1007/s13127-012-0112-4

Andrikou C, Thiel D, Ruiz-Santiesteban JA, Hejnol A. (2019) Active mode of excretion across digestive tissues predates the origin of excretory organs. PLoS Biol 17(7):e3000408. https://doi.org/10.1371/journal.pbio.3000408

Arboleda E, Hartenstein V, Martinez P, Reichert H, Sen S, Sprecher SG, Bailly X (2018) An emerging system to study photosymbiosis, brain regeneration, chronobiology, and behavior: the marine acoel symsagittifera roscoffensis. BioEssays 40(10). https://doi.org/10.1002/bies.201800107

Arimoto A, Hikosaka-Katayama T, Hikosaka A, Tagawa K, Inoue T, Ueki T, Yoshida M et al (2019) A draft nuclear-genome assembly of the acoel flatworm *Praesagittifera naikaiensis*. GigaScience 8(4):1–8. https://doi.org/10.1093/gigascience/giz023

Arroyo A, López-Escardó D, de Vargas C, Ruiz-Trillo I (2016) Hidden diversity of Acoelomorpha revealed through metabarcoding. Biol Lett 12(9). https://doi.org/10.1098/rsbl.2016.0674

Bedini C, Ferrero E, Lanfranchi A (1973) The ultrastructure of ciliary sensory cells in two Turbellaria Acoela. Tissue Cell 5(3):359–372. https://doi.org/10.1016/S0040-8166(73)80030-8

Beklemischev VN (1963) On the relationship of the Turbellaria to the other groups of the animal kingdom. In: Dougherty EC (ed) The lower Metazoa. University California Press, Berkeley, pp 234–244

Bery A, Martínez P (2011) Acetylcholinesterase activity in the developing and regenerating nervous system of the acoel symsagittifera roscoffensis. Acta Zoologica 92(4):383–392. https://doi.org/10.1111/j.1463-6395.2010.00472.x

Bery A, Cardona A, Martinez P, Hartenstein V (2010) Structure of the central nervous system of a juvenile acoel, *Symsagittifera roscoffensis*. Dev Genes Evol 220(3–4):61–76. https://doi.org/10.1007/s00427-010-0328-2

Bleidorn C, Podsiadlowski L, Zhong M, Eeckhaut I, Hartmann S, Halanych KM, Tiedemann R (2009) On the phylogenetic position of Myzostomida: can 77 genes get it wrong? BMC Evol Biol 9(1):150

Boone M, Bert W, Claeys M, Houthoofd W, Artois T (2011) Spermatogenesis and the structure of the testes in Nemertodermatida. Zoomorphology 130:273–282

Børve A, Hejnol A (2014) Development and juvenile anatomy of the nemertodermatid *Meara stichopi* (Bock) Westblad 1949 (Acoelomorpha). Front Zool 11(January):50. https://doi.org/10.1186/1742-9994-11-50

Bourlat SJ, Nielsen C, Lockyer AE, Littlewood DT, Telford MJ (2003) *Xenoturbella* is a deuterostome that eats molluscs. Nature 530:925–928

Bourlat SJ, Juliusdottir T, Lowe CJ, Freeman R, Aronowicz J, Kirschner M, Lander ES et al (2006) Deuterostome phylogeny reveals monophyletic chordates and the new phylum Xenoturbellida. Nature 444(November):85–88. https://doi.org/10.1038/nature05241

Brauchle M, Bilican A, Eyer C, Bailly X, Martínez P, Ladurner P, Bruggmann R, Sprecher SG (2018) Xenacoelomorpha survey reveals that all 11 animal Homeobox gene classes were present in the first Bilaterians. Genome Biol Evol 10(9):2205–2217. https://doi.org/10.1093/gbe/evy170

Buckland-Nicks J, Lundin K, Wallberg A (2018) The sperm of Xenacoelomorpha revisited: implications for the evolution of early Bilaterians. Zoomorphology. https://doi.org/10.1007/s00435-018-0425-8

Cannon JT, Vellutini B, Smith J, Ronquist F, Jondelius U, Hejnol A (2016) Xenacoelomorpha is the sister group to Nephrozoa. Nature 530(7588):89–93. https://doi.org/10.1038/nature16520

Chang YC, Pai CY, Chen YC, Ting HC, Martinez P, Telford MJ. Yu JK, Su YH (2015) Regulatory circuit rewiring and functional divergence of the duplicate admp genes in dorsoventral axial patterning. https://doi.org/10.1016/j.ydbio.2015.12.015

Chiodin M, Achatz JG, Wanninger A, Martinez P (2011) Molecular architecture of muscles in an acoel and its evolutionary implications. J Exp Zool B Mol Dev Evol 316(6):427–439. https://doi.org/10.1002/jez.b.21416

Chiodin M, Børve A, Berezikov E, Ladurner P, Martinez P, Hejnol A (2013) Mesodermal gene expression in the acoel *Isodiametra pulchra* indicates a low number of mesodermal cell types and the endomesodermal origin of the gonads. PLoS ONE 8(2):e55499. https://doi.org/10.1371/journal.pone.0055499

Conway-Morris S, George JD, Gibson R, Pllatt HM (1985) The origins and relationships of lower invertebrates. Clarendon Press, Oxford

Cook CE, Jimenez-Guri E, Akam M, Salo E (2004) The hox gene complement of acoel flatworms, a basal bilaterian clade. Article Evol Dev 6(3):154–163

Crezée M (1978) *Paratomella rubra* Rieger and Ott, an amphiatlantic acoel turbellarian. Cah Biol Mar 19:1–9

Dittmann IL, Zauchner T, Nevard LM, Telford MJ, Egger B (2018) SALMFamide2 and serotonin immunoreactivity in the nervous system of some acoels (Xenacoelomorpha). J Morphol 279(5):589–597

Dorey AE (1965) The organization and replacement of the epidermis in acoelous turbellarians. Q J Microsc Sci 106:147–172

Dörjes J (1968) Die Acoela (Turbellaria) Der Deutschen Nordseekste Und Ein Neues System Der Ordnung. Z. Zool Syst Evolutionforsch 6:56–452

Dörjes J (1972) *Faerlea echinocardii* Sp. N. Und Diskussion Der Gattungen *Avagina* Leiper Und *Faerlea* Westblad (Turbellaria Acoela). Zoolog Scr 1(3):185–189
Ehlers U (1984) Phylogenetisches system der plathelminthes. Verh Natwiss Ver Hamburg 27:291–294
Ehlers U (1985) Das Phylogenetische System Der Plathelminthes. Gustav Fischer, Stutgart
Ehlers U (1991) Comparative morphology of statocysts in the Plathelminthes and the Xenoturbellida. Hydrobiologia 227:263–271
Ehlers U (1992a) Dermonephridia–modified epidermal cells with a probable excretory function in *Paratomella rubra* (Acoela, Plathelminthes). Microfauna Mar 7:253–64
Ehlers U (1992b) On the fine structure of *Paratomella rubra* Rieger & Ott (Acoela) and the position of the taxon *Paratomella* Dörjes in a phylogenetic system of the Acoelomorpha (Plathelminthes). Microfauna Mar 7:265–293
Ehlers U (1992c) Frontal glandular and sensory structures in *Nemertoderma* (Nemertodermatida) and *Paratomella* (Acoela): ultrastructure and phylogenetic implications for the monophyly of the Euplathelminthes (Plathelminthes). Zoomorphology 112(4):227–236. https://doi.org/10.1007/BF01632820
Falleni A, Raikova O, Gremigni V (1995) Ultrastructural and cytochemical features of the ovary in *Paratomella rubra* (Platyhelminthes, Acoela). J Submicroscopical Cytol Pathol 27:511–523
Ferrero E (1973) A fine structural analysis of the statocyst in Turbellaria Acoela. Zoologica Scripta 2(1):5–16. https://doi.org/10.1111/j.1463-6409.1973.tb00793.x
Franzen A, Afzelius B (1987) The ciliated epidermis of *Xenoturbella bocki* (Platyhelminthes, Xenoturbellida) with some phylogenetic considerations. Zoolog Scr 16(1):9–17. https://doi.org/10.1111/j.1463-6409.1987.tb00046.x
Gaerber CW, Salvenmoser W, Rieger RM, Gschwentner R (2007) The nervous system of *Convolutriloba* (Acoela) and its patterning during regeneration after asexual reproduction. Zoomorphology 126(2):73–87. https://doi.org/10.1007/s00435-007-0039-z
Gavilán B, Sprecher SG, Hartenstein V, Martinez P (2019) The digestive system of xenacoelomorphs. Cell Tissue Res (in press). https://doi.org/10.1007/s00441-019-03038-2
Gavilán B, Perea-Atienza E, Martínez P (2016) Xenacoelomorpha: a case of independent nervous system centralization? Philos Trans R Soc Lond S B Biol Sci 371(1685):20150039. https://doi.org/10.1098/rstb.2015.0039
Gehrke AR, Neverett E, Luo Y-J, Brandt A, Ricci L, Hulett RE, Gompers A et al (2019) Acoel genome reveals the regulatory landscape of whole-body regeneration. Science 363 (6432):eaau6173. https://doi.org/10.1126/science.aau6173
Graff LV (1905) Turbellaria I. Acoela. In: F Schulze (ed) Das Tierreich, Eine Zusammenstellung Und Kennzeichnung Der Rezenten Tierformen Heft 23, pp 23–34. Königl. Preuss. Akademie der Wissenschaften zu Berlin
Graff LV (1911) Acoela, Rhabdocoela Und Alloeocoela Des Ostens Der Vereinigten Staaten von Amerika. Z Wiss Zool. 99:321–428
Gröger H, Schmid V (2001) Larval development in Cnidaria: a connection to Bilateria? Genesis 29:110–114
Gschwentner R, Ladurner P, Nimeth K, Rieger R (2001) Stem cells in a basal bilaterian. S-phase and mitotic cells in *Convolutriloba longifissura* (Acoela, Platyhelminthes). Cell Tissue Res 304(3):401–408
Haszprunar G (2015) Review of data for a morphological look on Xenacoelomorpha (Bilateria Incertae Sedis). Org Diversity Evol 16(2):363–389. https://doi.org/10.1007/s13127-015-0249-z
Hejnol A, Martindale MQ (2008) Acoel development indicates the independent evolution of the bilaterian mouth and anus. Nature 456(7220):382–386. https://doi.org/10.1038/nature07309
Hejnol A, Martindale MQ (2009) Coordinated spatial and temporal expression of Hox genes during embryogenesis in the acoel Convolutriloba Longifissura. BMC Biol 1(7):65
Hejnol A, Obst M, Stamatakis A, Ott M, Rouse GW, Edgecombe GD, Martinez P et al (2009) Assessing the root of bilaterian animals with scalable phylogenomic methods. https://doi.org/10.1098/rspb.2009.0896

Hendelberg J (1969) On the development of different types of spermatozoa from spermatids with two flagella in the Turbellaria with remarks on the ultrastructure of the flagella. Zoologiska Bidrag Uppsala 38:1–52

Hendelberg J (1977) Comparative morphology of turbellarian spermatozoa studied by electron microscopy. Acta Zoologica Fennica 154:149–162

Hendelberg J (1986) The phylogenetic significance of sperm morphology in the Platyhelminthes. In: Advances in the biology of turbellarians and related platyhelminthes, pp 53–58. Springer, Dordrecht

Hendelberg J, Hedlund K-O (1974) On the morphology of the epidermal ciliary rootlet system of the acoelous turbellarian *Childia groenlandica*. Zoon 2:13–24

Hooge MD (2001) Evolution of the body-wall musculature in the Platyhelminthes (Acoelomorpha, Catenulida, Rhabditophora). J Morph 249:171–194

Iomini C, Raikova OI, Noury-Sraïri N, Justine J-L (1995) Immunocytochemistry of tubulin in spermatozoa of Platyhelminthes. Adv Spermatozoal Phylogeny Taxonomy 166:97–110

Israelsson O (1997) *Xenoturbella*'s Molluscan Relatives…] and Molluscan Embryogenesis. Nature 390:32

Jondelius U, Ruiz-Trillo I, Baguñà J, Riutort M (2002) The Nemertodermatida are basal bilaterians and not members of the Platyhelminthes. Zool Scr 31(2):201–215. https://doi.org/10.1046/j.1463-6409.2002.00090.x

Jondelius U, Wallberg A, Hooge M, Raikova OI (2011) How the worm got its pharynx: phylogeny, classification and Bayesian assessment of character evolution in Acoela. Syst Biol 60(6):845–871. https://doi.org/10.1093/sysbio/syr073

Kånneby T, Bernvi DC, Jondelius U (2015) Distribution, delimitation and description of species of *Archaphanostoma* (Acoela). Zool Scr 44(2):218–231. https://doi.org/10.1111/zsc.12092

Karling TG (1940) Zur Morphologie Und Systematik Der Alloeocoela Cumulata and Rhabditophora Lecithophora (Turbellaria). Acta Zool Fennica 26:1–160

Klauser MD, Smith JPS, Tyler S (1986) Ultrastructure of the frontal organ in *Convoluta* and *Macrostomum* spp.: significance for models of the turbellarian archetype. Hydrobiologia 132(1):47–52. https://doi.org/10.1007/BF00046227

Lanfranchi A (1990) Ultrastructure of the epidermal eyespots of an acoel platyhelminth. Tissue Cell 22 (4):541–46. http://www.ncbi.nlm.nih.gov/pubmed/18620320

Lundin K (1997) Comparative ultrastructure of the epidermal ciliary rootlets and associated structures in species of the Nemertodermatida and Acoela (Plathelminthes). Zoomorphology 117 (2):81–92. https://doi.org/10.1007/s004350050033

Lundin K (1998) The epidermal ciliary rootlets of *Xenoturbella bocki* (Xenoturbellida) revisited: new support for a possible kinship with the Acoelomorpha (Platyhelminthes). Zool Scr 27(3):263–270. https://doi.org/10.1111/j.1463-6409.1998.tb00440.x

Lundin K, Hendelberg J (1998) Is the sperm type of the Nemertodermatida close to that of the ancestral Platyhelminthes? Hydrobiologia 383:197–205

Martín-Durán JM, Pang K, Børve A, Lê HS, Furu A, Cannon JT, Jondelius U, Hejnol A (2018) Convergent evolution of Bilaterian nerve cords. Nature 553(7686):45–50. https://doi.org/10.1038/nature25030

Martinez P (2018) The Comparative method in biology and the essentialist trap. Frontiers in ecology and evolution 6 (AUG). https://doi.org/10.3389/fevo.2018.00130

Martinez P, Hartenstein V, Sprecher SG (2017) Xenacoelomorpha Nervous Systems. In: SM Sherman (ed) Oxford encyclopaedia of neurosciences. Oxford University Press

Meyer-Wachsmuth I, Jondelius U (2016) Interrelationships of Nemertodermatida. Org Div Evol. https://doi.org/10.1007/s13127-015-0240-8

Meyer-Wachsmuth I, Raikova OI, Jondelius U (2013) The muscular system of *Nemertoderma westbladi* and *Meara stichopi* (Nemertodermatida, Acoelomorpha). Zoomorphology 132(3):239–252. https://doi.org/10.1007/s00435-013-0191-6

Meyer-Wachsmuth I, Curini-Galletti M, Jondelius U (2014) Hyper-cryptic marine meiofauna: species complexes in Nemertodermatida. PLoS ONE 9(9):e107688. https://doi.org/10.1371/journal.pone.0107688

Moreno E, De Mulder K, Salvenmoser W, Ladurner P, Martínez P (2010) Inferring the ancestral function of the posterior hox gene within the *Bilateria*: controlling the maintenance of reproductive structures, the musculature and the nervous system in the acoel flatworm *Isodiametra pulchra*. Evol Dev May. https://doi.org/10.1111/j.1525-142x.2010.00411.x

Moreno E, Nadal M, Baguñà, J, Martínez P (2009) Tracking the origins of the *Bilaterian* hox patterning system: insights from the acoel flatworm *Symsagittifera roscoffensis*. Evol Dev 11 (5). https://doi.org/10.1111/j.1525-142X.2009.00363.x

Mulder K De, Kuales G, Pfister D, Willems M, Egger B, Salvenmoser W, Thaler M et al (2009) Characterization of the stem cell system of the acoel *Isodiametra pulchra*. BMC Dev Biol 9:69. https://doi.org/10.1186/1471-213X-9-69

Mwinyi A, Bailly X, Bourlat SJ, Jondelius U, Littlewood TJ, Podsiadlowski L (2010) The phylogenetic position of Acoela as revealed by the complete mitochondrial genome of *Symsagittifera roscoffensis*. BMC Evol Biol 10:309. https://doi.org/10.1186/1471-2148-10-309

Nakano H, Miyazawa H, Maeno A, Shiroishi T, Kakui K, Koyanagi R, Kanda M, Satoh N, Omori A, Kohtsuka H (2017) A new species of *Xenoturbella* from the western Pacific Ocean and the evolution of *Xenoturbella*. BMC Evol Biol 17(1):245

Noren M, Jondelius U (1997) *Xenoturbella*'s molluscan relatives. Nature 390:31–32. https://doi.org/10.1038/36242

Obst M, Nakano H, Bourlat SJ, Thorndyke MC, Telford MJ, Nyengaard JR, Funch P (2011) Spermatozoon ultrastructure of *Xenoturbella bocki* (Westblad 1949). Acta Zool 92(2):109–115. https://doi.org/10.1111/j.1463-6395.2010.00496.x

Paps J, Holland PWH (2018) Reconstruction of the ancestral metazoan genome reveals an increase in genomic novelty. Nat Commun 9(1):1730. https://doi.org/10.1038/s41467-018-04136-5

Perea-Atienza E, Gavilán B, Chiodin M, Abril J-F, Hoff KJ, Poustka AJ, Martinez P (2015) The nervous system of Xenacoelomorpha: a genomic perspective. J Exp Biol 218(Pt 4):618–628. https://doi.org/10.1242/jeb.110379

Perea-Atienza E, Sprecher SG, Martínez P (2018) Characterization of the bHLH family of transcriptional regulators in the acoel *S. roscoffensis* and their putative role in neurogenesis. Evodevo 9:8

Petrov AA, Hooge M, Tyler S (2004) Ultrastructure of sperms in Acoela (Acoelomorpha) and its concordance with molecular systematics. Invertebr Biol 123:183–197

Philippe H, Brinkmann H, Copley RR, Moroz LL, Nakano H, Poustka AJ, Wallberg A, Peterson KJ, Telford MJ (2011) Acoelomorph flatworms are deuterostomes related to *Xenoturbella*. Nature 470(7333):255–258. https://doi.org/10.1038/nature09676

Philippe H, PoustkaAJ, Chiodin M, Hoff KJ, Dessimoz C, Tomiczek B, Schiffer PH, Müller S, Domman D, Horn M, Kuhl H, Timmermann B, Satoh N, Hikosaka-Katayama T, Nakano H, Rowe ML, Elphick MR, Thomas-Chollier M, Ha MJ (2019) Mitigating anticipated effects of systematic errors supports sister-group relationship between Xenacoelomorpha and Ambulacraria. Curr Biol (in press)

Popova NV, Mamkaev YV (1985) Ultrastructure and primitive features of the eyes of *Convoluta convoluta* (Turbellaria Acoela). Dokl Akad Nauk SSSR 283:756–759

Raikova OI (1991) On phylogenetic significance of ultrastructural characters in Turbellaria. In: Proceedings of the Zoological Institute of the Academy of Sciences of the USSR, pp 26–52

Raikova OI (2004) Neuroanatomy of basal bilaterians (Xenoturbellida, Nemertodermatida, Acoela) and its phylogenetic implications (Ph.D. Thesis). Åbo Akademi University. Åbo, Finland

Raikova OI, Justine J-L (1999) Microtubular system during spermiogenesis and in the spermatozoon of *Convoluta saliens* (Platyhelminthes, Acoela): tubulin immunocytochemistry and electron microscopy. Mol Reprod Dev 52:74–85

Raikova OI, Falleni A, Gremigni V (1995) Oogenesis in *Actinoposthia beklemischevi* (Platyhelminthes, Acoela): an ultrastructural and cytochemical study. Tissue Cell 27:621–633

Raikova OI, Reuter M, Kotikova EA, Gustafsson MKS (1998) A commissural brain! The pattern of 5-HT immunoreactivity in Acoela (Plathelminthes). Zoomorphology 118(2):69–77. https://doi.org/10.1007/s004350050058

Raikova OI, Reuter M, Jondelius U, Gustafsson MKS (2000a) An immunocytochemical and ultrastructural study of the nervous and muscular systems of *Xenoturbella westbladi* (Bilateria Inc. Sed.). Zoomorphology 120(2):107–18. https://doi.org/10.1007/s004350000028

Raikova OI, Reuter M, Jondelius U, Gustafsson MKS (2000b) The brain of the Nemertodermatida (Platyhelminthes) as revealed by anti-5HT and anti-FMRFamide immunostainings. Tissue Cell 32(5):358–365. https://doi.org/10.1054/tice.2000.0121

Raikova OI, Reuter M, Gustafsson MKS, Maule AG, Halton DW, Jondelius U (2004a) Basiepidermal nervous system in *Nemertoderma westbladi* (Nemertodermatida): GYIRFamide immunoreactivity. Zoology 107(1):75–86. https://doi.org/10.1016/j.zool.2003.12.002

Raikova OI, Reuter M, Gustafsson MKS, Maule AG, Halton DW, Jondelius U (2004b) Evolution of the nervous system in *Paraphanostoma* (Acoela). Zool Scr 33:71–88

Raikova OI, Meyer-Wachsmuth I, Jondelius U (2016) The plastic nervous system of Nemertodermatida. Org Div Evol 16(1):85–104. https://doi.org/10.1007/s13127-015-0248-0

Ramachandra NB, Gates RD, Ladurner P, Jacobs DK, Hartenstein V (2002) Embryonic development in the primitive bilaterian *Neochildia fusca*: normal morphogenesis and isolation of POU genes Brn-1 and Brn-3. Dev Genes Evol 212(2):55–69. https://doi.org/10.1007/s00427-001-0207-y

Reisinger E (1925) Untersuchungen Am Nervensystem Der *Bothrioplana semperi* Braun. (Zugleich Ein Beitrag Zur Technik Der Vitalen Nervenfaerbung Und Zur Vergleichenden Anatomie Des Plathelminthennervensystem). Z Morphol Okol Tiere 5:119–149

Reisinger E (1960) Was ist *Xenoturbella*. Z Wiss Zool 164:188–198

Reuter M, Raikova OI, Jondelius U, Gustafsson MKS, Maule AG, Halton, DV (2001) Organisation of the nervous system in the Acoela: an immunocytochemical study. Tissue Cell 33(2):119–28. http://www.ncbi.nlm.nih.gov/pubmed/11392663

Reuter M, Raikova OI, Gustafsson MKS (2001) Patterns in the nervous and muscle systems in lower flatworms. Belgian J Zool 131 (Suppl:47–53

Richter S, Loesel R, Purschke G, Schmidt-Rhaesa A, Scholtz G, Stach T, Vogt L, Wanninger A, Brenneis G, Döring C, Faller S, Fritsch M, Grobe P, Heuer CM, Kaul S, Møller OS, Müller CH, Rieger V, Rothe BH, Stegner ME, Harzsch S (2010) Invertebrate neurophylogeny: suggested terms and definitions for a neuroanatomical glossary. Front Zool. 7:29. https://doi.org/10.1186/1742-9994-7-29

Rieger RM, Tyler S, Smith JPS III, Rieger GE (1991) Platyhelminthes: Turbellaria. In: Bogitsh BJ, Harrison FW (eds) Microscopic anatomy of invertebrates. Wiley-Liss, New York

Robertson HE, Lapraz F, Egger B, Telford MJ, Schiffer PH (2017) The mitochondrial genomes of the acoelomorph worms *Paratomella rubra*, *Isodiametra pulchra* and *Archaphanostoma ylvae*. Sci R 7(1):1–16. https://doi.org/10.1038/s41598-017-01608-4

Rohde K, Watson NA, Cannon LRG (1988) Ultrastructure of epidermal cilia of *Pseudactinoposthia* sp. (Platyhelminthes, Acoela); implications for the phylogenetic status of the Xenoturbellida and Acoelomorpha. J Submicroscopical Cytol Pathol 20:759–767

Rouse GW, Wilson NG, Carvajal JI, Vrijenhoek RC (2016) New deep-sea species of *Xenoturbella* and the position of Xenacoelomorpha. Nature. https://doi.org/10.1038/nature16545

Ruiz-Trillo I, Riutort M, Littlewood DT, Herniou EA, Baguña J (1999) Acoel flatworms: earliest extant bilaterian metazoans, not members of Platyhelminthes. Science 283 (5409):1919–23. http://www.ncbi.nlm.nih.gov/pubmed/10082465

Ruiz-Trillo I, Paps J, Loukota M, Ribera C, Jondelius U, Baguñà J, Riutort M (2002) A phylogenetic analysis of myosin heavy chain type II sequences corroborates that Acoela and Nemertodermatida are basal bilaterians. Proc Natl Acad Sci 99(17):11246–11251. https://doi.org/10.1073/pnas.172390199

Ruiz-Trillo I, Riutort M, Fourcade HM, Baguñà J, Boore JL (2004) Mitochondrial genome data support the basal position of Acoelomorpha and the polyphyly of the Platyhelminthes. Mol Phylogenet Evol 33(2):321–332. https://doi.org/10.1016/j.ympev.2004.06.002

Schmidt-Rhaesa, A. (2007). The evolution of organ systems. Oxford University Press
Semmler H, Chiodin M, Bailly X, Martinez P, Wanninger A (2010) Steps towards a centralized nervous system in basal bilaterians: insights from neurogenesis of the acoel *Symsagittifera roscoffensis*. Dev Growth Differ 52(8):701–713. https://doi.org/10.1111/j.1440-169X.2010.01207.x
Seo HC, Edvardsen RB, Maeland AD, Bjordal M, Jensen MF, Hansen A, Flaat M, Weissenbach J, Lehrach H, Wincker P, Reinhardt R, Chourrout D (2004) Hox cluster disintegration with persistent anteroposterior order of expression in *Oikopleura dioica*. Nature 431(7004):67–71
Sikes JM, Bely AE (2010) Making heads from tails: development of a reversed anterior–posterior axis during budding in an acoel. Dev Biol 338(1):86–97. https://doi.org/10.1016/j.ydbio.2009.10.033
Smith III JPS, Tyler S (1988) Frontal organs in the Nemertodermatida (Turbellaria). Am Zool 28(4):140A, #747
Smith III JPS, Tyler S, Rieger RM (1986) Is the Turbellaria Polyphyletic? Hydrobiologia 132:13–21
Smith JPS, Tyler S (1985) Fine-structure and evolutionary implications of the frontal organ in Turbellaria Acoela. 1 *Diopisthoporus gymnopharyngeus* sp.n. Zool Scr 14(2):91–102. https://doi.org/10.1111/j.1463-6409.1985.tb00180.x
Smith JPS, Tyler S (1985) The acoel turbellarians: kingpins of metazoan evolution or a specialized offshoot?" In: Morris SC, George JD, Gibson R, Platt HM (eds) The origins and relationships of lower invertebrates, 123142. Oxford: Clarendon Press
Sopott-Ehlers B, Ehlers U (1997) Ultrastructure of the subepidermal musculature of *Xenoturbella bocki*, the adelphotaxon of the Bilateria. Zoomorphology 117:71–79. https://doi.org/10.1007/s004350050032
Sprecher SG, Bernardo-Garcia F-J, van Giesen L, Hartenstein V, Reichert H, Neves R, Bailly X et al. (2015) Functional brain regeneration in the acoel worm *Symsagittifera roscoffensis*. Biology Open 4 (12):1688–95. https://doi.org/10.1242/bio.014266
Srivastava M, Mazza-Curll KL, van Wolfswinkel JC, Reddien PW (2014) Whole-body acoel regeneration is controlled by wnt and bmp-admp signaling. Curr Biol CB 24(10):1107–1113. https://doi.org/10.1016/j.cub.2014.03.042
Sterrer W (1998) New and known Nemertodermatida (Platyhelminthes-Acoelomorpha): a revision. Belgian J Zool 128(1):55–92. https://www.researchgate.net/publication/265923621_New_and_known_Nemertodermatida_Platyhelminthes-Acoelomorpha_A_revision
Tekle YI, Raikova OI, Justine J-L, Jondelius U (2007a) Ultrastructure and tubulin immunocytochemistry of the copulatory stylet-like structure in *Childia* species (Acoela). J Morphol 268:166–180
Tekle YI, Raikova OI, Justine J-L, Hendelberg J, Jondelius U (2007b) Ultrastructural and immunocytochemical investigation of acoel sperms with 9+ 1 axoneme structure: new sperm characters for unraveling phylogeny in Acoela. Zoomorphology 126:1–16
Telford MJ, Lockyer AE, Cartwright-Finch C, Littlewood TJ (2003) Combined large and small subunit ribosomal RNA phylogenies support a basal position of the acoelomorph flatworms. Proc Biol Sci 270 (1519):1077–83. https://doi.org/10.1098/rspb.2003.2342
Thiel D, Franz-Wachtel M, Aguilera F, Hejnol A (2018) Xenacoelomorph neuropeptidomes reveal a major expansion of neuropeptide systems during early bilaterian evolution. Mol Biol Evol 35(10):2528–2543. https://doi.org/10.1093/molbev/msy160
Todt C (2009) Structure and evolution of the pharynx simplex in acoel flatworms (Acoela). J Morphol 270(3):271–290. https://doi.org/10.1002/jmor.10682
Todt C, Tyler S (2006) Ciliary receptors associated with the mouth and pharynx of Acoela (Acoelomorpha): a comparative ultrastructural study. Acta Zoologica 88(1):41–58. https://doi.org/10.1111/j.1463-6395.2007.00246.x
Tyler S (1979) Distinctive features of cilia in metazoans and their significance for systematics. Tissue Cell 11:385–400
Tyler S, Hooge MD (2004) Comparative morphology of the body wall in flatworms (Platyhelminthes). Can J Zool 82:194–210
Tyler S, Rieger RM (1975) Uniflagellate spermatozoa in Nemertoderma (Turbellaria) and their phylogenetic significance. Science 188:730–732

Tyler S, Rieger RM (1977) Ultrastructural evidence for the systematic position of the Nemertodermatida (Turbellaria). Acta Zool Fennica 54:193–207
Uljanin WN (1870) Die Turbellarien Der Bucht von Sebastopol. Arbeiten Der 2.Versammlung Russischer Naturforscher Zu Moskau. 1869:1–96
Westblad E (1937) Die Turbellarien-Gattung *Nemertoderma* Steinböck. Acta Societatis pro Fauna et Flora Fennica 60:45–89
Westblad E (1940) Studien Über Skandinavische Turbellaria Acoela. I. Arkiv För Zoologi 32A(20):1–28
Westblad E (1942) Studien Über Skandinavische Turbellaria Acoela. II. Arkiv För Zoologi 33A(14):1–48
Westblad E (1945) Studien Über Skandinavische Turbellaria Acoela. III. Ark Zool. 36A(5):1–56
Westblad E (1946) Studien Über Skandinavische Turbellaria Acoela. IV. Ark Zool 38A(1):1–56
Westblad E (1948) Studien Über Skandinavische Turbellaria Acoela. V. Ark Zool 41:191–273
Westblad E (1949a) On *Meara stichopi* (Bock) Westblad, a new representative of Turbellaria archoophora. Ark Zool Ser 2 1(5):43–57
Westblad E (1949b) *Xenoturbella bocki* N.g., N. Sp., a peculiar, primitive Turbellarian type. Ark. Zool 1:3–29
Yamasu T (1991) Fine structure and function of ocelli and sagittocysts of acoel flatworms. Hydrobiologia 227(1):273–282. https://doi.org/10.1007/BF00027612
Zabotin YI (2019) Ultrastructure of epidermal sensillae in three species of Acoela. Invertebr Zool 6(1):71–77

Chapter 15
Cooperation and Competition in Mammalian Evolution

Gene Domestication from LTR Retrotransposons

Tomoko Kaneko-Ishino and Fumitoshi Ishino

Abstract Mammalian genomes have had to endure the integration of exogenous DNA sequences over the course of time. In most cases, such events have proven harmful to individuals thus afflicted, but in the long-term gene domestication of exogenous DNA sequences, such as LTR retrotransposons, has also served as a driving mechanism in biological evolution. This is especially the case in eutherian mammals, in which two lines of domesticated genes increased in number in a common eutherian ancestor, eleven sushi-ichi-related retrotransposon homologs (*SIRH*)/retrotransposon Gag-like (*RTL*) genes and more than fifteen paraneoplastic Ma antigen (*PNMA*) genes. It is clear that these *SIRH*/*RTL* and *PNMA* genes were positively selected due to the advantage conferred on eutherian reproductive success. Thus, the principle of "competition among individuals within the same species" in the Darwinian theory of evolution is effectively at work in the domestication process. However, when the number of domestication events is taken into account, how could the common eutherian ancestor have acquired that many domesticated genes? We suggest that sexual mating across multiple generations of individuals with one or a small number of such domesticated genes may have been critically important for accumulating all of them into a single line, thus leading to the common eutherian ancestor. Then, we would like to propose that "cooperation among individuals within the same species" in the form of interactive behaviors of the individuals within a tightly delimited species is also at work in this process.

15.1 Introduction

The acquisition of vivipary in therian mammals and subsequent divergence into viviparous marsupial and eutherian subgroups having a different placental system

T. Kaneko-Ishino
Tokai University, Isehara City, Kanagawa 259-1193, Japan
e-mail: tkanekoi@is.icc.u-tokai.ac.jp

F. Ishino (✉)
Tokyo Medical and Dental University, Yushima, Bunkyo-ku, Tokyo 113-8510, Japan
e-mail: fishino.epgn@mri.tmd.ac.jp

P. Pontarotti (ed.), *Evolution, Origin of Life, Concepts and Methods*,
https://doi.org/10.1007/978-3-030-30363-1_15

and reproductive strategy are critical events in the course of mammalian evolution. What underlies these events from the viewpoint of genomic function? It is probable that genomic imprinting and gene domestication from long terminal repeat (LTR) retrotransposons/retroviruses played important roles in these processes by promoting dramatic change in gene function of host organisms. This is because two paternally expressed imprinted genes, paternally expressed 10 (*PEG10*) and *PEG11*/retrotransposon Gag-like 1 (*RTL1*), are essential placental genes, even though they were actually derived from a suchi-ichi-related LTR retrotransposon (Ono et al. 2001, 2006; Charlier et al. 2001; Sekita et al. 2008). Moreover, at least some of the nine additional domesticated genes from the same retrotransposon, i.e., sushi-ichi-related retrotransposon homologs (*SIRH*)/*RTL* genes, have important roles not only in the placenta, but also in brain in the current developmental system (Naruse et al. 2014; Irie et al. 2015, 2016; Kaneko-Ishino and Ishino 2015). Interestingly, both the genomic imprinting and gene domestication mechanisms are related to the insertion of exogenous DNA. The former was apparently established as a defense mechanism against the insertion of exogenous cis-regulatory units (Barlow 1993; Kaneko-Ishino and Ishino 2019), while the latter is exaptation mechanism for creating genes of novel function from the inserted LTR retrotransposons/retroviruses (Gould and Vrba 1982; Brosius and Gould 1992).

In terms of genomic imprinting that regulates parent-of-origin-specific monoallelic expression of imprinted genes, a critical regulatory sequence in each imprinted region, the so-called differentially methylated region (DMR), is derived from exogenous DNA having cis-regulatory function, such as insulators, enhancers, and promoters of antisense RNA. The genomic imprinting mechanism then reorganized the gene expression profile around the insertion site so as to ensure the expression of all of the genes by regulating paternal- or maternal-specific monoallelic expression. It is probable that this mechanism is advantageous because additional imprinted regions repeatedly emerged independently in several different eutherian lineages (Kaneko-Ishino and Ishino 2019).

It is known that during biological evolution, certain essential genes were acquired by domestication (exaptation or cooption) of LTR retrotransposons/retroviruses and/or DNA transposons (Kaneko-Ishino and Ishino 2012), such as telomerase in eukaryotes (Nakamura and Cech 1998), recombination activating genes (*RAG1* and *RAG2*) for immunoglobulin class switching by somatic recombination in vertebrates (Agrawal et al 1998; Hiomi et al. 1998), far-red elongated hypocotyls 3 (*FHY3*), and far-red-impaired response (*FAR1*) genes for photosynthesis in plants (Lisch et al. 2001; Hudson et al. 2003; Lin et al. 2007). Another famous example in mammals is syncytin gene that is essential for placental syncytiotrophoblast formation, but interestingly different syncytin genes were domesticated from different retroviruses in a lineage-specific manner (Blond et al. 2000; Mi et al. 2000; Dupressoir et al. 2005; Heidemann et al. 2009). Interestingly, comprehensive genome analysis in mammals revealed that the number of domesticated genes from another LTR retrotransposon, paraneoplastic Ma antigen (*PNMA*) genes, also increased in a common eutherian ancestor as well as *SIRH*/*RTL* genes (Brandt et al. 2005; Schüller et al. 2005; Youngson et al. 2005; Campillos et al. 2006; Ono et al. 2006). Accumulating evidence

has shown that these *PNMA* genes also play important roles in the current eutherian developmental systems (Cho et al. 2008a, b, 2011; Hanson et al. 2011a, b), indicating that *SIRH/RTL* and *PNMA* gene domestication was a driving force in evolution of marsupial and eutherians (Ono et al. 2006; Sekita et al. 2008; Cho et al. 2008a, b, 2011; Hanson et al. 2011a, b; Naruse et al. 2014; Irie et al. 2015, 2016).

These data suggest that the defense mechanism against the insertion of exogenous DNAs, such as DNA transposons and LTR retrotransposons/retroviruses, as well as the capacity for gene domestication from such inserted DNAs, is important for biological evolution of organisms. As we have already articulated the importance of genomic imprinting in mammalian evolution in a recent review (Kaneko-Ishino and Ishino 2019), in this chapter we focus on the gene domestication mechanism in mammalian evolution and consider how the expansion of the *SIRH/RTL* genes could have occurred in the common eutherian ancestor. We propose that both "cooperation" and "competition" among individuals within a species are important principles in mammalian evolution. The same idea may also apply to the domestication of other selfish genetic elements such as LINEs, SINEs, and retroviruses as well as non-retroviral RNA viruses such as bornavirus and ebola/marburgviruses (Horie et al. 2010; Belyi et al. 2010).

15.2 LTR Retrotransposon-Derived *SIRH/RTL* and *PNMA* Genes in Eutherians and Marsupials

PEG10 and *PEG11/RTL1* were the first two domesticated genes derived from an LTR retrotransposon to be recognized in mammals (Ono et al. 2001; Charlier et al. 2001). They were first identified as paternally expressed imprinted genes in humans/mice and cows, respectively, exhibiting significant homology to a sushi-ichi retrotransposon isolated from fugu fish (Fig. 15.1). *PEG10* is a therian-specific gene, i.e., conserved in both the marsupials and eutherians (Suzuki et al. 2007), while *PEG11/RTL1* is a eutherian-specific gene (Edwards et al. 2008) (Fig. 15.2). The fact that both *PEG10* and *PEG11/RTL1* are essential placental genes in mice (Ono et al. 2006; Sekita et al. 2008) promoted the screening of related domesticated genes in humans and mice. Importantly, eutherians have nine additional genes domesticated from the same LTR retrotransposon family, which accordingly were named sushi-ichi-related retrotransposon homologs (*SIRH3–11*) [alternatively called *MART* or *SUSHI* genes and recently renamed retrotransposon Gag-like (*RTL*)] (Brandt et al. 2005; Youngson et al. 2005; Ono et al. 2006) (Figs. 15.1 and 15.3). Importantly, they are "eutherian-specific," like *PEG11/RTL1*, because there are no orthologous genes present in marsupials (Ono et al. 2011). On the other hand, marsupials have another *SIRH12* that is absent from eutherians, suggesting that the expansion of the *SIRH/RTL* genes occurred independently in the marsupial and eutherian lineages except for *PEG10.* The tammar wallaby, an Australian marsupial species, has *SIRH12*, while opossum, a South American marsupial species, has *pseudoSIRH12* along with many

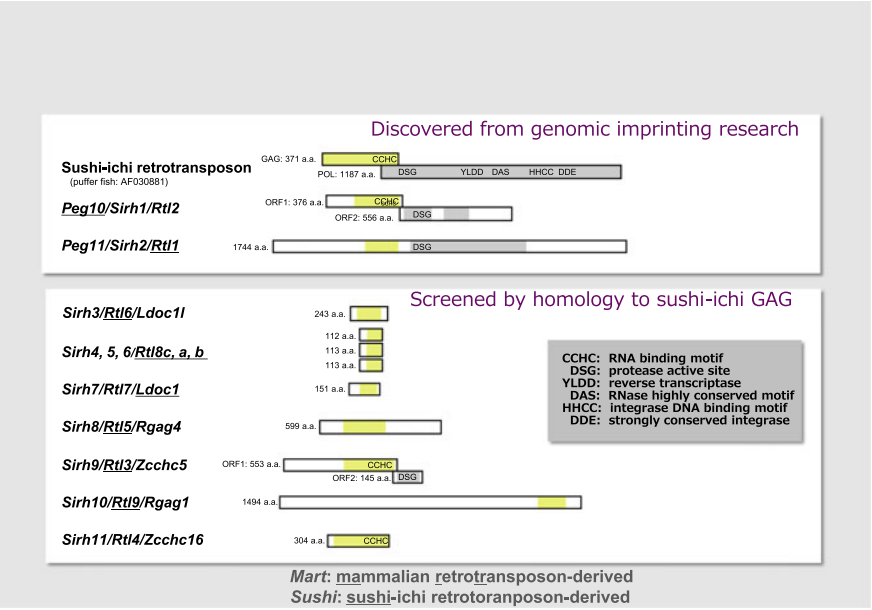


Fig. 15.1 *SIRH*/*RTL* genes in eutherians. Mouse 11 *SIRH*/*RTL* genes are shown as examples. The formal name of each *SIRH*/*RTL* gene is underlined. The yellow part indicates homology to the sushi-ichi Gag protein and the gray part the sushi-ichi Pol protein

nonsense mutations in its ORF, suggesting the insertion of *SIRH12* occurred in a common marsupial ancestor, but was only domesticated in the australidelphian and not the ameridelphian lineages (Ono et al. 2011) (Fig. 15.2). Even in eutherians, some *SIRH*/*RTL* genes have been lost in various species: *SIRH8*/*RTL5* became a pseudogene in most afrotherian and xenarthran species, and *SIRH9*/*RTL3* has been lost in many eutherian species (Kaneko-Ishino et al. 2017). No functional *SIRH11*/*RTL4* is present in xenarthran species (Irie et al. 2015), and the lack of an N-terminal or a CCHC RNA-binding domain in the C-terminal has been observed to be characteristics of the boreotherians in a species- or lineage-specific manner (Irie et al. 2016).

The situation is the same for the paraneoplastic Ma antigen (*PNMA*) genes, genes domesticated from a Gypsy_12DR-related LTR retrotransposon (Schüller et al. 2005; Campillos et al. 2006; Kokosar and Kordiš 2013; Iwasaki et al. 2013): Eutherians have more than 15 of these genes while marsupials have two genes that are different from the eutherians, *PNMA-MS1* and *PNMA-MS2,* but also have many remnants of the Gypsy_12DR-related LTR retrotransposon. These facts suggest that eutherians have a much stronger capacity for exapting new genes from the inserted LTR retrotransposon sequences than do marsupials (Ono et al. 2011; Iwasaki et al. 2013). It is highly likely that these *SIRH*/*RTL* and *PNMA* genes contributed to the evolution and diversification of marsupials and eutherians, presumably via their specific functional

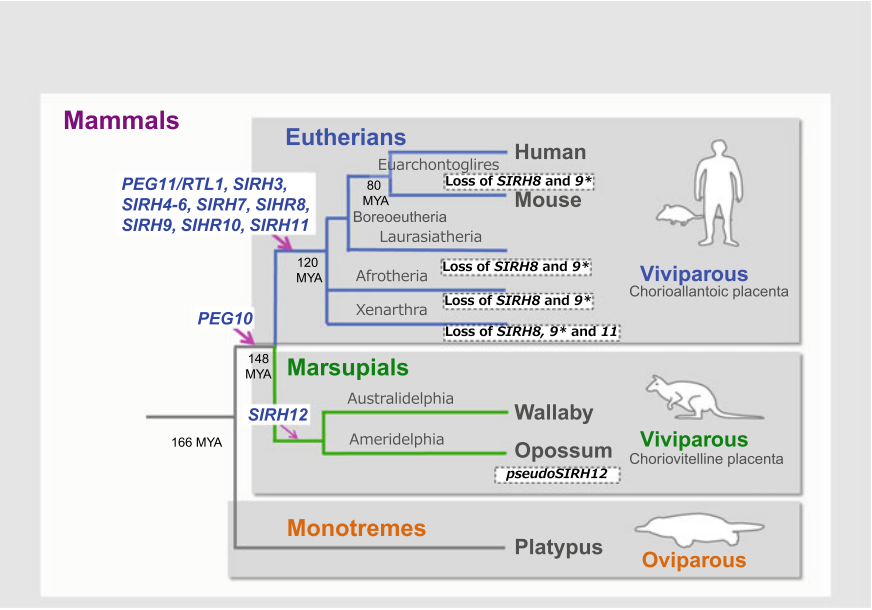


Fig. 15.2 Emergence of the *SIRH* family of genes in mammals. Shown are groups of mammals in which each *SIRH* gene is present

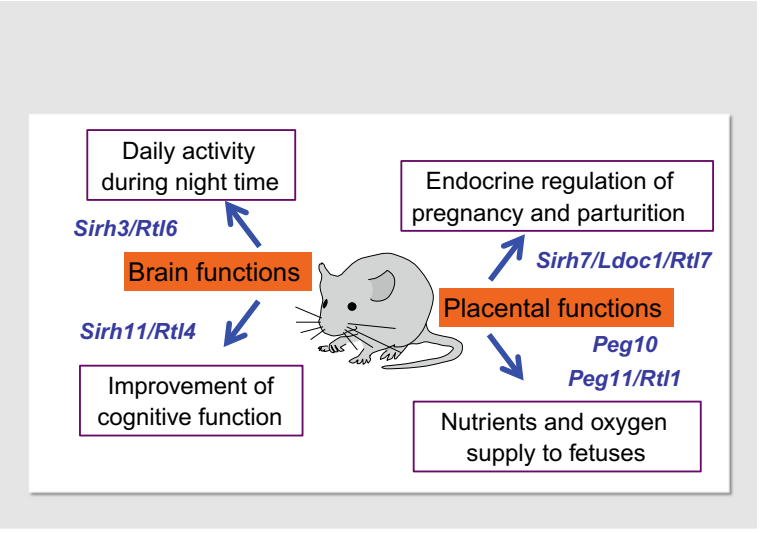


Fig. 15.3 Biological functions of *SIRH*/*RTL* genes in mice. The biological functions are deduced from each of the KO mouse studies. They have independent placenta- and brain-related functions, but only the major phenotypes are presented

roles in the reproductive and developmental systems (see Sects. 3 and 4) (Fig. 15.3). Thus, eutherians and marsupials constitute distinct animal groups with regard to genes domesticated from LTR retrotransposons.

15.3 *PEG10* Domestication in a Common Therian Ancestor

The emergence of vivipary in therian mammals is one of the truly major events in the course of mammalian evolution. It is reasonable to conclude that *PEG10* has greatly contributed to this process because *PEG10* is essential for placenta formation as a therian-specific gene. *PEG10* encodes two open reading frames (ORFs), PEG10-ORF1 and ORF2, that exhibit homologies to the Gag and Pol proteins of a sushi-ichi retrotransposon, respectively (Ono et al. 2001, 2006; Shigemoto et al. 2001) (Fig. 15.1). Importantly, the ORF1-2 fusion protein of the current *PEG10* is translated via a −1 frameshift mechanism that is only functional in LTR retrotransposons and retroviruses, and even then, not in all of them (Ono et al. 2001, 2006; Shigemoto et al. 2001; Manktelow et al. 2005; Clark et al. 2007). *PEG10* is expressed in both the embryo and placenta (Ono et al. 2001), while *PEG10* knockout (KO) mice exhibit early embryonic lethality due to a poorly developed placenta (Ono et al. 2006). Their placentas lack labyrinths and spongiotrophoblast layers, essential components for nutrient and gas exchange between the fetus and mother. Given that ectoplacental cone (EPC) growth was markedly affected in the KO placenta, it is highly probable that *PEG10* plays an important role in the differentiation of trophoblast cells in the labyrinth and spongiotrophoblast layers from the EPC cells. It is of considerable interest to elucidate whether *PEG10* also plays an essential role in the choriovitelline (yolk sac) placenta of marsupials, which is of a different type than the chorioallantoic placenta of eutherians (Suzuki et al. 2007; Renfree et al. 2013).

The genomic record shows that *PEG10* is present only in eutherian and marsupial mammals among vertebrates, indicating that *PEG10* emerged from an original LTR retrotransposon after diversification from monotremes (166 MYA) and before the split of marsupials and eutherians (148 MYA) (Suzuki et al. 2007) (Figs. 15.2 and 15.4: A. Common therian ancestor 1). Assuming that the sushi-ichi retrotransposon is the direct ancestor of *PEG10*, the fact that amino acid sequence identity between the sushi-ichi retrotransposon Gag/Pol and the mouse PEG10 ORF1-2 protein is only 26.8–33.0% suggests that numerous mutations were required to form the current *PEG10* (Fig. 15.4: B. Common therian ancestor 2 and orange dotted lines between A and B). However, it might be also possible that certain currently unknown LTR retrotransposon with higher homology to *PEG10* is the direct ancestor. In any case, it is clear that *PEG10* spreads throughout all the species in therians, a subclass of mammals, because of the developmental advantage it confers, as proposed by the Darwinian theory of evolution (Kaneko-Ishino and Ishino 2010, 2015). As mentioned, it remains unknown whether marsupial *PEG10* is essential for yolk sac placental function, although marsupial *PEG10* is highly expressed in the yolk sac.

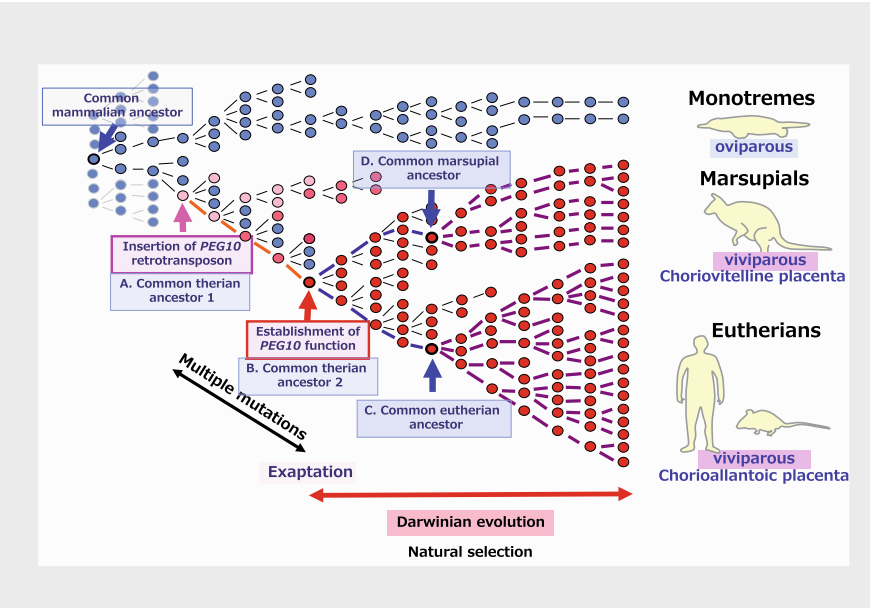


Fig. 15.4 Acquisition and propagation of *PEG10*. Integration of the original PEG10 retrotransposon (pink arrow) occurred after the divergence of monotremes and therian mammals, and its domestication (red arrow) might have been completed before the split of the eutherians and marsupials. Multiple mutations were required between these two events

Therefore, it is necessary to elucidate *PEG10's* role in the marsupial placenta to conclude when the current form of *PEG10* was established in mammalian history.

15.4 Subsequent Acquisition of *PEG11/RTL1* and Other *SIRH/RTL* Genes in a Common Eutherian Ancestor

The divergence of the two viviparous mammalian groups, marsupials and eutherians, with each having its own, distinct reproductive strategy using a different placental system, is another major event in mammalian evolution (Renfree 2010; Renfree et al. 2013). It is reasonable to think that *PEG11/RTL1* and *SIRH7*/Leucine zipper, down-regulated in cancer 1 (*LDOC1*, also called *RTL7*), have powerfully contributed to this process, because both are eutherian-specific and essential for placental function: the former in the maintenance of fetal capillaries in the eutherian type chorioallantoic placenta (Sekita et al. 2008; Kitazawa et al. 2017) and the latter in the endocrinological regulation of a variety of placental hormones (Naruse et al. 2014).

The PEG11/RTL1 protein also possesses considerable homology to both the Gag and Pol proteins of the sushi-ichi retrotransposon, like the PEG10 ORF1-2

fusion protein (Fig. 15.1), although no frameshift is necessary for the current form of PEG11/RTL1 (Charlier et al. 2001). The amino acid sequence identity between the Gag- and Pol-like regions of the mouse PEG11/RTL1 protein and the sushi-ichi retrotransposon Gag and Pol is 25.0 and 21.8%, respectively. Then, the requirement of multiple mutations to establish current *PEG11/RTL1* is the same as *PEG10*. However, it should be noted that the amino acid sequence identity between the PEG10 and PEG11/RTL1 proteins is approximately 20–30%, and that they have completely different functions in development. *PEG11/RTL1* KO mice clearly demonstrated that *PEG11/RTL1* is essential for the maintenance of the placental fetal capillary network in the mid-to-late fetal stages (Sekita et al. 2008; Kitazawa et al. 2017). *PEG11/RTL1* is expressed in endothelial cells of the fetal capillaries in the labyrinth layer where important feto-maternal interactions occur, such as gas and nutrient exchange. It is likely that the PEG11/RTL1 protein protects the endothelial cells against attack from surrounding placental trophoblast cells having a highly invasive and hence dangerous nature, because the loss of *PEG11/RTL1* causes clogging in many of the fetal capillaries in the labyrinth layer by extensive phagocytosis of endothelial cells, a process carried out by the trophoblast cells.

As mentioned above, marsupials have a choriovitelline (yolk sac) placenta and give birth to their young after a very short gestation period compared to the eutherians (Renfree 2010). *PEG11/RTL1* function is necessary for the eutherian type of chorioallantoic placenta that is required for the longer gestational period in the eutherians. When and how was the original *PEG11/RTL1* inserted into a eutherian ancestor? One simple idea is that it was acquired after divergence from marsupials (Fig. 15.5). Alternatively, it may have been inserted into a common therian ancestor like *PEG10* (Fig. 15.4) but became functionally lost in the marsupial lineage after their divergence because a similar DNA sequence with very few regions of homology remains at an appropriate position of the marsupial genome (Edwards et al. 2008). It is clear that *PEG11/RTL1* has a pivotal role in the eutherian reproduction system, indicating its critical contribution to the divergence of these two viviparous mammalian groups. Thus, *PEG11/RTL1* provides another good example of macroevolution in mammals, the establishment of eutherian mammals, which was achieved by the domestication of genes from LTR retrotransposons (Kaneko-Ishino and Ishino 2010, 2015). However, it is difficult to conclude the specific when and how this domestication took place from the limited information available at present.

SIRH7/LDOC1 (Nagasaki et al. 1999) is also a eutherian-specific gene (Naruse et al. 2014). It encodes a small Gag-like protein corresponding to the central part of the Gag protein (Fig. 15.1). The mouse SIRH7/LDOC1 protein exhibits 28.3% identity to the sushi-ichi Gag protein. In mice, *SIRH7/LDOC1* is predominantly expressed in the early stages of the placenta and a high level of expression has been reported in all of the placental cells in the three layers of the placenta, i.e., the labyrinth, spongiotrophoblast (SpT) and giant trophoblast (TG) layers (Naruse et al. 2014). In addition to structural problems, endocrinological abnormalities are present in the *SIRH7/LDOC1* KO placenta, such as progesterone (P4) overproduction and a delayed switch from placental lactogen I to II (PL1 to PL2) in the TG cells and up/downregulation of several prolactin-like proteins (PRLs) in the SpT cells. P4 is

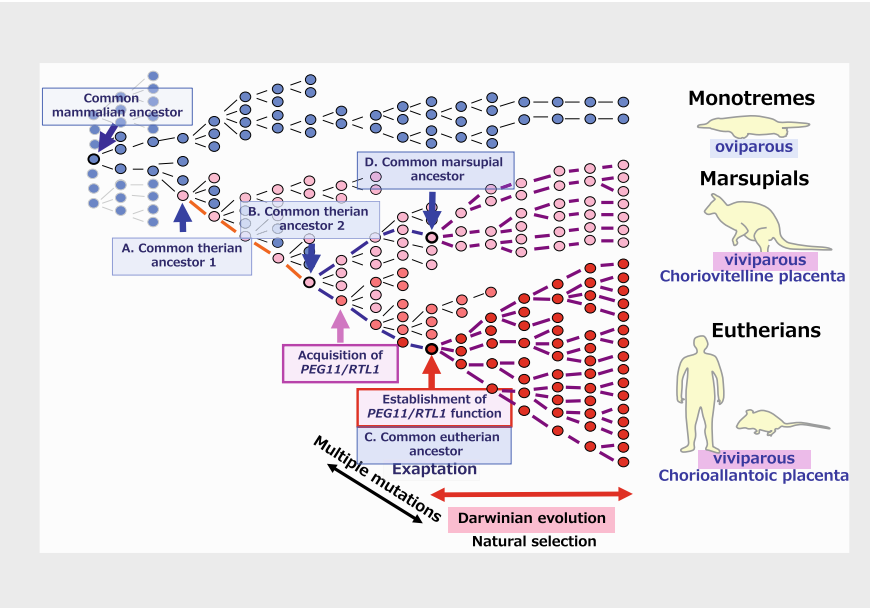


Fig. 15.5 Acquisition and propagation of *PEG11/RTL1*. In this scenario, acquisition of the original *PEG11/RTL1* sequence (pink arrow) occurred after the divergence of eutherians and marsupials, and its domestication (red arrow) was completed before the common eutherian ancestor emerged. However, the timing of the acquisition of the original *PEG11/RTL1* sequence remains to be determined

an essential hormone for maintaining pregnancy in mammals, and the downregulation of P4 to a zero level is necessary one day before of parturition. In the case of *SIRH7/LDOC1* KO animals, females exhibited delayed parturition, leading to a low pup weaning rate (Naruse et al. 2014). Reproductive success is a major driving force in the Darwinian theory of evolution, and *SIRH7/LDOC1* is an essential placental gene for reproductive fitness, thus contributing to the divergence of these two viviparous mammalian groups, like *PEG11/RTL1*.

Interestingly, some *SIRH/RTL* genes play an important role in brain function, such as *SIRH11/RTL4* (also called *ZCCHC16*) (Irie et al. 2015), *SIRH3/RTL6* (also called *LDOC1L*) (Irie et al. submitted), and *SIRH8/RTL5* (also called *RGAG4*) (unpublished data). In addition to the placenta, the highly developed central nerves system (CNS) is one of the most apparent features of the eutherians. Therefore, it is possible that all of the eutherian-specific *SIRH3–11* genes have contributed to the divergence of these two viviparous mammalian groups, even though the functions of *SIRH9/RTL3* and *SIRH10/RTL9* remain unknown to date. Similar to the case for *PEG11/RTL1*, it is difficult to conclude when and the means by which the original *SIRH3–11* genes were inserted into the eutherian ancestor and became functional genes in the current adaptive form, because both scenarios, before and after the divergence of the eutherians from marsupials, are possible, as discussed above (Figs. 15.4 and 15.5).

15.5 How Did a Common Eutherian Ancestor Acquire All of the *SIRH*/*RTL* Genes?

Comprehensive genome analysis in mammals has revealed three different gene groups have been domesticated from LTR retrotransposons/retroviruses in mammals: (1) Mammalian-specific *SASPase* (a retroviral-like aspartic protease) is a single domesticated gene shared by monotremes, marsupials, and eutherians (Bernard et al. 2005; Matsui et al. 2006, 2011), (2) the *SIRH*/*RTL* genes from a sushi-ichi-related LTR retrotransposon, and (3) *PNMA* genes from a presumable Gypsey_12DR-related LTR retrotransposon. Thus, it is clear that at least three independent domestication events from LTR retrotransposons/retroviruses occurred during mammalian evolution. Subsequently, the number of *SIRH*/*RTL* genes as well as *PNMA* genes increased in a common eutherian ancestor (Brandt et al. 2005; Schüller et al. 2005; Youngson et al. 2005; Campillos et al. 2006; Ono et al. 2006), while their total numbers remained small in the marsupial lineage (Ono et al. 2011; Iwasaki et al. 2013).

As mentioned earlier, *PEG10* is therian-specific while *PEG11*/*RTL1* and *SIRH3–11* are eutherian-specific genes, and all these genes exhibit a certain degree of homology (approximately 20–30%) to the sushi-ichi retrotransposon from fugu. Importantly, the amino acid sequence identities among these genes are also approximately 20–30%, indicating that they are functionally divergent and were independently selected during the course of eutherian evolution. Comprehensive phylogenic analysis also supported this notion, except for a triplet of *SIRH4, 5*, *6*/*RTL8abc*, with almost the same amino acid sequences, and *SIRH3*/*RTL6* and *SIRH8*/*RTL5*, which are closely related to each other, because the latter, larger protein covers the entire Gag-derived region of the former, smaller protein.

As discussed in the previous section, reconstruction of the entire domestication process of these genes is very difficult because several major questions remain unaddressed at the moment.

(1) Which is the origin of all the *SIRH* genes, an LTR retrotransposon or retrovirus (Malik et al. 2000; Ribet et al. 2008)?
(2) Were they all acquired directly from some LTR retrotransposon/retrovirus by independent domestication events or did their number increase by expansion from a single or a small number of progenitor DNA sequences (Naville et al. 2016; Kordiš 2017)? If the latter is the case, what is the progenitor DNA sequence(s) of these *SIRH* genes in the genome, if it indeed existed in the past?
(3) How were the original LTR retrotransposon/retrovirus-derived sequences expanded? Which method was used, reinfection of the retroviruses, retrotransposition of the LTR retrotransposons, gene duplication of the integrated chromosomal sites (Suga et al. 1999; Gu et al. 2002; Naville et al. 2016), or cDNA retropositioning of transcripts from the first domesticated gene (Vanin 1985; Brosius 1999; Pavlícek et al. 2002)?

(4) How long did each domestication (exaptation) process take, that is, from the insertion of the original LTR retrotransposon/retrovirus-derived DNA sequence to the current form of each *SIRH*/*RTL* gene?

As all *SIRH*/*RTL* genes exhibit a certain degree of homology to the current sushi-ichi LTR retrotransposon and there were no Env-like sequences observed in any of these genes, it is reasonable to think that they are simply derived from some sushi-ichi-related LTR retrotransposon in the past. However, it is also possible that the Env as well as LTR sequences at both ends were lost during the extremely long time since the beginning of domestication (at least 120 million years before present) if these sequences were functionally dispensable in the new genes. Therefore, a retrovirus with a high homology to the current *SIRH* genes might be the original source of these *SIRH*/*RTL* genes (Malik et al. 2000; Ribet et al. 2008). In such a case, all or some could be derived from independent and sequential infection events that occurred between the split of the therians (148 MYA) and the establishment of a common eutherian ancestor (at least 120 MYA). The subsequently remaining members might have been expanded by either reinfection with the original virus or gene duplication of the inserted loci, as reported in the case of HERV-K (Belshaw et al. 2005).

Assuming the derivation from an LTR retrotransposon, how was the first progenitor able to infect the germline cells of our ancestors? In one scenario, an original LTR retrotransposon may have acted as an infectious agent by using a co-infected retroviral envelops, and the inserted LTR retrotransposon generated progeny lines via sequential retrotransposition events. Alternatively, a mechanism for producing processed pseudogenes might have been at work in this process; i.e., several different lengths of a cDNA may have been synthesized from transcripts of the first domesticated gene using reverse transcriptase of an LTR retrotransposon/retrovirus or by the activity of an L1 retrotransposon ORF2-like protein (Vanin 1985; Brosius 1999; Pavlícek et al. 2002). It is proposed that the latter mechanism contributed to the increment of processed pseudogenes that is associated with the simultaneous burst of *Alu* elements that occurred in the early stage of primate evolution (40–50 MYA) (Ohshima et al. 2003). In this scenario, it might have taken a long-time period for the expanded processed pseudogenes to become functional genes because of the sequential acquisition of the cis-regulatory units required for expression, such as a promoter, appropriate 5′- and 3′ untranslated regions (UTRs) (Kordiš 2017). As all of the *SIRT*/*RTL* genes encode different parts of the Gag and Pol regions of presumably the same original LTR retrotransposon/retrovirus, it is likely that the route via processed pseudogene was used during the expansion process in addition to the subsequent gene duplication for producing the triplet *SIRH4–6*/*RTL8abc* along with *SIRH3*/*RTL6* and *SIRH8*/*RTL5*. But many other combinations of mechanisms are possible, as discussed above.

Which one is the progenitor of the *SIRH*/*RTL* genes? Unfortunately, phylogenic analysis on these *SIRH* genes does not provide useful information because each *SIRH* gene has almost equal phylogenic distance from other *SIRH* genes except the triplet *SIRH4–6*/*RTL8abc* and the couple of *SIRH3*/*RTL6* and *SIRH8*/*RTL5*. However, as *PEG10* is exclusively a therian-specific gene and has both Gag- and Pol-derived

regions, it seems probable that other *SIRH*/*RTL* genes were expanded from *PEG10* via several of the routes discussed above. However, *PEG11*/*RTL1* encodes a longer Pol-like region exhibiting higher homology to the sushi-ichi Pol protein than the PEG10 Pol-like region, suggesting that at least *PEG10* and *PEG11*/*RTL1* were independently domesticated. Another possibility is that *PEG11*/*RTL1* generated *PEG10* in an ancestral line of common therians but was lost in the marsupial lineage because its function was not required for the marsupial yolk sac placenta, as described above. It is also possible that the progenitor of all *SIRH* genes, including *PEG10*, was lost after the domestication of *PEG10*.

15.6 The Principles of Mammalian Evolution from the Viewpoint of Gene Domestication from LTR Retrotransposons

The possibilities discussed above are ultimately speculative and a considerable amount of new evidence will be required to answer the outstanding questions. However, the critically important point is that the common eutherian ancestor must have had all of the 11 *SIRH*/*RTL* genes. Subsequently, some were lost in a species- or lineage-specific manner during the evolutionary radiation of eutherian mammals as described above. In summary, *PEG10*, *PEG11*/*RTL1*, *SIRH3*/*RTL6*, *SIRH4–6*/*SIRH8abc,* and *LDOC1*/*SIRH7*/*RTL7* are highly conserved in the eutherian species, whereas *SIRH8*/*RTL5*, *SIRH9*/*RTL3*, *SIRH10*/*RTL9,* and *SIRH11*/*RTL4* are lost in a species- and/or lineage-specific manner in eutherians. It is likely that the former genes play essential roles in the eutherian development and behavioral systems, whereas the latter genes may serve as critical determinants in the process of diversification among the eutherians via their brain-related functions or other lineage- and/or species-specific characteristics (Kaneko-Ishino and Ishino 2015).

In humans and mice, all the 11 *SIRH*/*RTL* genes are conserved and a series of KO mouse studies clearly demonstrated that at least *PEG10* (Ono et al. 2006), *PEG11*/*RTL1* (Sekita et al. 2008), *SIRH3*/*RTL6* (Irie et al. submitted), *SIRH4–6*/*SIRH8abc* (unpublished data), *LDOC1*/*SIRH7*/*RTL7* (Naruse et al. 2014), *SIRH11*/*RTL4*/*Zcchc16* (Irie et al. 2015), and *SIRH8*/*RTL5* (unpublished data) have independently important roles in development, reproduction, and/or behavior in mice. These facts provide strong evidence that these *SIRH*/*RTL* genes are positively selected in eutherian evolution due to the advantages they confer (Figs. 15.4 and 15.5).

In terms of the multiple acquisitions of the 11 *SIRH*/*RTL* genes, did they occur in a single ancestral line leading to a common eutherian ancestor? Given that viral infection was the source of the *SIRH*/*RTL* genes, it is highly probable that each of their domestication events first occurred independently in many infected individuals (Fig. 15.6a–c). *SIRH*/*RTL* genes may then have accumulated by sexual mating across several generations of host individuals (Fig. 15.6d, e). Finally, all of the *SIRH*/*RTL*

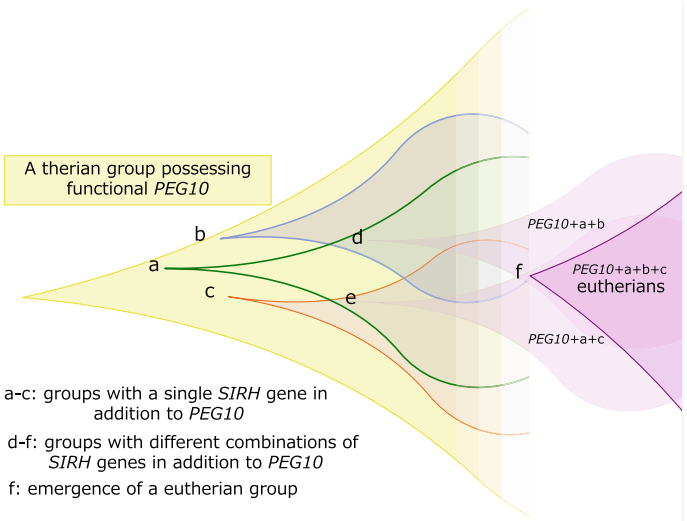


Fig. 15.6 Emergence of the eutherians by the sequential gain of domesticated *SIRH*/*RTL* genes. In this scenario, individuals (**a**–**c**) in a therian group having the domesticated *PEG10* inhabiting a certain limited area are infected with a sushi-ichi-related retrotransposon (or retrovirus) and the original *SIRH* DNA sequences become integrated into their germ cells. After several generations of host individuals, the *SIRH* genes accumulate by sexual mating in a step-by-step manner (**d**, **e**). Finally, all the *SIRH* genes come to be accumulated in a single individual (**f**, the common eutherian ancestor), generating a single eutherian ancestral line. Progeny possessing all of the *SIRH* genes are propagated in this line. It is hypothesized that the coexistence of several independent ancestral lines, each possessing different *SIRH* genes, is required to produce the common eutherian ancestor. Although the precise timing of the domestication of these genes is presently unknown, it should be noted that the expansion of *PNMA* genes occurred simultaneously in the same group, as discussed in the main text

genes may have eventually accumulated in a single individual or a single ancestral line by sexual mating across further generations of host individuals in a step-by-step manner (Fig. 15.6f). Such an individual or line might have emerged as the common eutherian ancestor because of the combined advantages in development, reproduction, and behavior. In this scenario, it is reasonable that many of the ancestral lines would have existed at the same time, rather than a single ancestral line directly led to the common eutherian ancestor.

Given that the propagation from the progenitor(s) was the mechanism of *SIRH*/*RTL* gene expansion, a similar scenario, the existence of a considerable number of ancestral lines may be the likeliest scenario. Thus, when we take individual domesticated genes into account, we clearly see the "competition" principle proposed by Darwinian theory of evolution acting as the major driving force in eutherian evolution (Figs. 15.4 and 15.5). However, when we take multiple domesticated genes into account, the existence of a considerable number of ancestral lines was suggested to generate a single individual of common eutherian ancestor or a single ancestral line starting from the common eutherian ancestor. This means that the "cooperation" or

integrated activities of a number of individuals or lines within the same species might have had to play an essential background role in generating the common eutherian ancestral line. The same may be true for the expansion process of *PNMA* genes. Thus, from the multiple domestications of genes from LTR retrotransposons viewpoint, the cooperative activity of individuals seems another important principle of biological evolution to go along with the competition between individuals within a species.

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

Barlow DP (1993) Methylation and imprinting: from host defense to gene regulation? Science 260:309–310

Belshaw R, Dawson AL, Woolven-Allen J, Redding J, Burt A, Tristem M (2005) Genomewide screening reveals high levels of insertional polymorphism in the human endogenous retrovirus family HERV-K(HML2): implications for present-day activity. J Virol 79:12507–12514

Belyi VA, Levine AJ, Skalka AM (2010) Unexpected inheritance: multiple integrations of ancient bornavirus and ebolavirus/marburgvirus sequences in vertebrate genomes. PLoS Pathog 6:e1001030

Bernard D, Méhul B, Thomas-Collignon A, Delattre C, Donovan M, Schmidt R (2005) Identification and characterization of a novel retroviral-like aspartic protease specifically expressed in human epidermis. J Invest Dermatol 125:278–287

Blond J-L, Lavillette D, Cheynet V, Bouton O, Oriol G, Chapel-Fernandes S, Mandrand B, Mallet F, Cosset FL (2000) An envelope glycoprotein of the human endogenous retrovirus HERV-W is expressed in the human placenta and fuses cells expressing the type D mammalian retrovirus receptor. J Virol 74:3321–3329

Brandt J, Schrauth S, Veith AM, Froschauer A, Haneke T, Schultheis C, Gessler M, Leimeister C, Volff JN (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 (1999) RNAs from all categories generate retrosequences that may be exapted as novel genes or regulatory elements. Gene 238:115–134

Brosius J, Gould SJ (1992) On "genomenclature": a comprehensive (and respectful) taxonomy for pseudogenes 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

Charlier C, Segers K, Wagenaar D, Karim L, Berghmans S, Jaillon O, Shay T, Weissenbach J, Cockett N, Gyapay G, Georges M (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

Cho G, Bhat SS, Gao J, Collins JS, Rogers RC, Simensen RJ, Schwartz CE, Golden JA, Srivastava AK (2008a) Evidence that *SIZN1* is a candidate X-linked mental retardation gene. Am J Med Genet A 146A:2644–2650

Cho G, Lim Y, Zand D, Golden JA (2008b) *Sizn1* is a novel protein that functions as a transcriptional coactivator of bone morphogenic protein signaling. Mol Cell Biol 28:1565–1572

Cho G, Lim Y, Golden JA (2011) XLMR candidate mouse gene, Zcchc12 (Sizn1) is a novel marker of Cajal-Retzius cells. Gene Expr Patterns 11:216–220

Clark MB, Jänicke M, Gottesbühren U, Kleffmann T, Legge M, Poole ES, Tate WP (2007) Mammalian gene PEG10 expresses two reading frames by high efficiency −1 frameshifting in embryonic-associated tissues. J Biol Chem 282:37359–37369

Dupressoir A, Marceau G, Vernochet C, Bénit L, Kanellopoulos C, Sapin V, Heidmann T (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, Mungall AJ, Matthews L, Ryder E, Gray DJ, Pask AJ, Shaw G, Graves JA, Rogers J, SAVOIR Consortium, Dunham I, Renfree MB, Ferguson-Smith AC (2008) The evolution of the *DLK1-DIO3* imprinted domain in mammals. PLoS Biol 6:e135

Gould SJ, Vrba ES (1982) Exaptation; a missing term in the science of form. Paleobiology 8:4–15

Gu X, Wang Y, Gu J (2002) Age distribution of human gene families shows significant roles of both large- and small-scale duplications in vertebrate evolution. Nat Genet 31:205–209

Hanson D, Murray PG, O'Sullivan J, Urquhart J, Daly S, Bhaskar SS, Biesecker LG, Skae M, Smith C, Cole T, Kirk J, Chandler K, Kingston H, Donnai D, Clayton PE, Black GC (2011a) Exome sequencing identifies *CCDC8* mutations in 3-M syndrome, suggesting that *CCDC8* contributes in a pathway with *CUL7* and *OBSL1* to control human growth. Am J Hum Genet 89:148–153

Hanson D, Murray PG, Black GC, Clayton PE (2011b) The genetics of 3-m syndrome: unravelling a potential new regulatory growth pathway. Horm Res Paediatr 76:369–378

Heidmann O, Vernochet C, Dupressoir A, Heidmann T (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

Hiom K, Mele M, Gellert M (1998) DNA transposition by the RAG1 and RAG2 proteins: a possible source of oncogenic translocations. Cell 94:463–470

Horie M, Honda T, Suzuki Y, Kobayashi Y, Daito T, Oshida T, Ikuta K, Jern P, Gojobori T, Coffin JM, Tomonaga K (2010) Endogenous non-retroviral RNA virus elements in mammalian genomes. Nature 463:84–87

Hudson ME, Lisch DR, Quail PH (2003) The *FHY3* and *FAR1* genes encode transposase-related proteins involved in regulation of gene expression by the phytochrome A-signaling pathway. Plant J 34:453–471

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

Irie M, Koga A, Kaneko-Ishino T, Ishino F (2016) An LTR retrotransposon-derived gene displays lineage-specific structural and putative species-specific functional variations in eutherians. Front Chem 4:26

Irie M, Ito J, Furuse T, Ishida S, Yamada I, Wakana S, Kiyonari H, Kihara M, Tachibana S, Kohda T, Tanaka, K Ishino F, Kaneko-Ishino T (submitted) Gene targeting of an LTR retrotransposon-derived *Sirh3/Rtl6* gene leads to a prolonged sleep phenotype in mice

Iwasaki S, Suzuki S, Clark H, Ono R, Shaw G, Renfree MB, Kaneko-Ishino T, Ishino F (2013) Identification of novel *PNMA-MS1* in marsupials suggests LTR retrotransposon-derived *PNMA* genes differently expanded in marsupials and eutherians. DNA Res 20:425–436

Kaneko-Ishino T, Ishino F (2010) Retrotransposon silencing by DNA methylation contributed to the evolution of placentation and genomic imprinting in mammals. Dev 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

Kaneko-Ishino T, Ishino F (2019) Evolution of viviparity in mammals: what genomic imprinting tells us about mammalian placental evolution. Reprod Fert Dev. https://doi.org/10.1071/RD18127

Kaneko-Ishino T, Irie M, Ishino F (2017) Mammalian-specific traits generated by LTR retrotransposon-drived genes. In: Pontarotti P (ed) Evolutionary biology: self/nonself evolution, species and complex traits evolution, methods and concepts. Springer International Publishing, pp 129–145

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 of placental size in *Peg11/Rtl1* KO mice. Genes Cells 22:174–188

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

Kordiš D (2017) The life history of domesticated genes illuminates the evolution of novel mammalian genes. In: Pontarotti P (ed) Evolutionary biology: self/nonself evolution, species and complex traits evolution, methods and concepts. Springer International Publishing, pp 129–145

Lin R, Ding L, Casola C, Ripoll DR, Feschotte C, Wang H (2007) Transposase-derived transcription factors regulate light signaling in *Arabidopsis*. Science 318:1302–1305

Lisch DR, Freeling M, Langham RJ, Choy MY (2001) Mutator transposase is widespread in the grasses. Plant Physiol 125:1293–1303

Malik HS, Henikoff S, Eickbush TH (2000) Poised for contagion: evolutionary origins of the infectious abilities of invertebrate retroviruses. Genome Res 10:1307–1318

Manktelow E, Shigemoto K, Brierley I (2005) Characterization of the frameshift signal of Edr, a mammalian example of programmed −1 ribosomal frameshifting. Nucleic Acids Res 33:1553–1563

Matsui T, Kinoshita-Ida Y, Hayashi-Kisumi F, Hata M, Matsubara K, Chiba M, Katahira-Tayama S, Morita K, Miyachi Y, Tsukita S (2006) Mouse homologue of skin-specific retroviral-like aspartic protease involved in wrinkle formation. J Biol Chem 281:27512–27525

Matsui T, Miyamoto K, Kubo A, Kawasaki H, Ebihara T, Hata K, Tanahashi S, Ichinose S, Imoto I, Inazawa J, Kudoh J, Amagai M (2011) SASPase regulates stratum corneum hydration through profilaggrin-to-filaggrin processing. EMBO Mol Med 3:320–333

Mi S, Lee X, Li X-P, Veldman GM, Finnerty H, Racie L, LaVallie E, Tang X-Y, 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

Nagasaki K, Manabe T, Hanzawa H, Maass N, Tsukada T, Yamaguchi K (1999) Identification of a novel gene, *LDOC1*, down-regulated in cancer cell lines. Cancer Lett 140:227–234

Nakamura TM, Cech TR (1998) Reversing time: origin of telomerase. Cell 92:587–590

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

Ohshima K, Hattori M, Yada T, Gojobori T, Sakaki Y, Okada N (2003) Whole-genome screening indicates a possible burst of formation of processed pseudogenes and Alu repeats by particular L1 subfamilies in ancestral primates. Genome Biol 4:R74

Ono R, Kobayashi S, Wagatsuma H, Aisaka K, Kohda T, Kaneko-Ishino T, Ishino F (2001) A retrotransposon-derived gene, *PEG10*, is a novel imprinted gene located on human chromosome 7q21. Genomics 73:232–237

Ono R, Nakamura K, Inoue K, Naruse M, Usami T, Wakisaka-Saito N, Hino T, Suzuki-Migishima R, Ogonuki N, Miki H, Kohda T, Ogura A, Yokoyama M, Kaneko-Ishino T, Ishino F (2006) Deletion of *PEG10*, an imprinted gene acquired from a retrotransposon, causes early embryonic lethality. Nat Genet 38:101–106

Ono R, Kuroki Y, Naruse M, Ishii M, Iwasaki S, Toyoda A, Fujiyama A, Shaw G, Renfree MB, Kaneko-Ishino T and Ishino F (2011) Identification of tammar wallaby *SIRH12*, derived from a marsupial-specific retrotransposition event. DNA Res 18:211–219

Pavlícek A, Paces J, Elleder D, Hejnar J (2002) Processed pseudogenes of human endogenous retroviruses generated by LINEs: their integration, stability, and distribution. Genome Res 12:391–399

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

Ribet D, Harper F, Dupressoir A, Dewannieux M, Pierron G, Heidmann T (2008) An infectious progenitor for the murine IAP retrotransposon: emergence of an intracellular genetic parasite from an ancient retrovirus. Genome Res 18:597–609

Schüller M, Jenne D and Voltz R (2005) The human PNMA family: novel neuronal proteins implicated in paraneoplastic neurological disease. J Neuroimmunol 169:172–176

Sekita Y, Wagatsuma H, Nakamura K, Ono R, Kagami M, Wakisaka N, Hino T, Suzuki-Migishima R, Kohda T, Ogura A, Ogata T, Yokoyama M, Kaneko-Ishino T, Ishino F (2008) Role of retrotransposon-derived imprinted gene, *Rtl1*, in the feto-maternal interface of mouse placenta. Nat Genet 40:243–248

Shigemoto K, Brennan J, Walls E, Watson CJ, Stott D, Rigby PW, Reith AD (2001) Identification and characterisation of a developmentally regulated mammalian gene that utilises-1 programmed ribosomal frameshifting. Nucleic Acids Res 29:4079–4088

Suga H, Koyanagi M, Hoshiyama D, Ono K, Iwabe N, Kuma K, Miyata T (1999) Extensive gene duplication in the early evolution of animals before the parazoan-eumetazoan split demonstrated by G proteins and protein tyrosine kinases from sponge and hydra. Mol Evol 48:646–653

Suzuki S, Ono R, Narita T, Pask AJ, Shaw G, Wang C, Kohda T, Alsop AE, Graves MJA, Kohara Y, Ishino F, Renfree MB, Kaneko-Ishino T (2007) Retrotransposon silencing by DNA methylation can drive mammalian genomic imprinting. PLoS Genet 3:e55

Vanin EF (1985) Processed pseudogenes: characteristics and evolution. Annu Rev Genet 19:253–272

Youngson NA, Kocialkowski S, Peel N, Ferguson-Smith AC (2005) A small family of sushi-class retrotransposon-derived genes in mammals and their relation to genomic imprinting. J Mol Evol 61:481–490