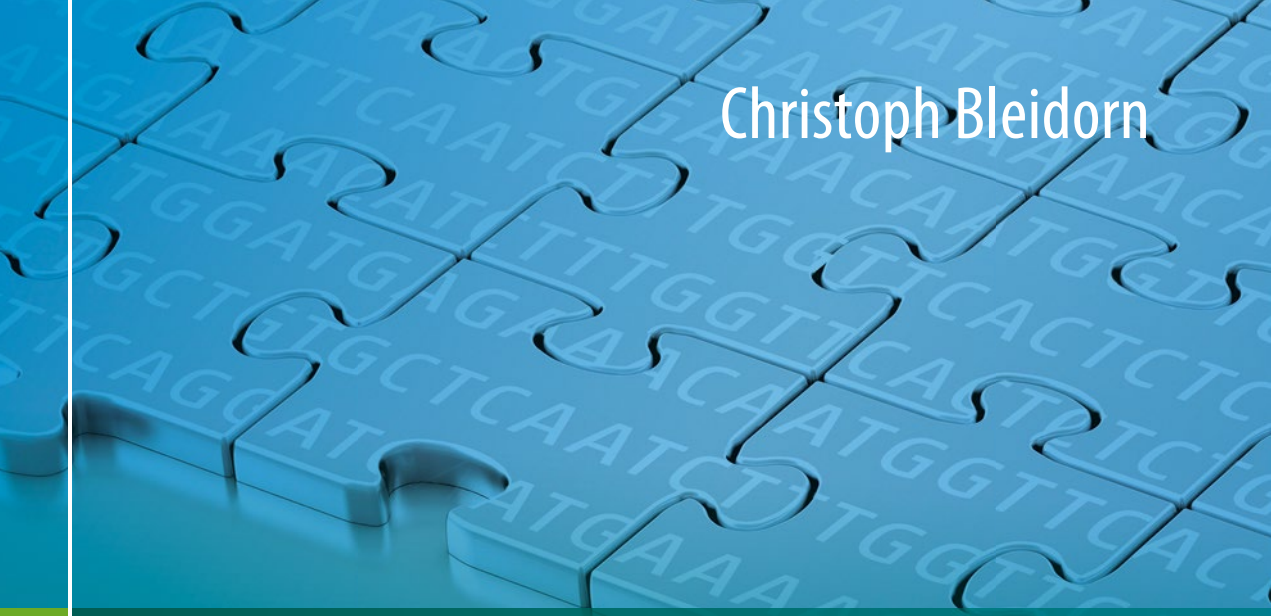


Christoph Bleidorn

Phylo-genomics

An Introduction

Springer

Phylogenomics

Christoph Bleidorn

Phylogenomics

An Introduction

Springer

Christoph Bleidorn
Museo Nacional de Ciencias Naturales
Spanish National Research Council (CSIC)
Madrid
Spain

ISBN 978-3-319-54062-7 ISBN 978-3-319-54064-1 (eBook)
DOI 10.1007/978-3-319-54064-1

Library of Congress Control Number: 2017942964

Printed on acid-free paper

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

Preface

All life on earth shares a common ancestor, and the aim of phylogenetic systematics is to reconstruct the tree or network of life. Shortly after the availability of the first protein sequences, molecular phylogenetic approaches were developed to understand the evolutionary relationships between proteins (or genes). It became clear that gene trees will also help to unravel the phylogeny of species. The introduction of Sanger sequencing and polymerase chain reaction (PCR) paved the way that genetic approaches became available across the scientific community and contributed to the rise of molecular phylogenetics. At the end of the 1990s, results from single-gene studies challenged the century-old textbook view of evolutionary relationships of many groups (e.g. animals, plants). Fierce discussions regarding the validity of these results led to important methodological advances, and, nowadays, molecular phylogenies are broadly accepted to represent organismal relationships in textbooks. In the mid-2000s, the way of sequencing has been revolutionized, leading to a huge drop in its costs, and unprecedented amounts of sequence data became affordable for every type of study and also for non-model organisms. This development transformed the field of molecular phylogenetics to phylogenomics, where genome-scale data (genomes, transcriptomes) can be exploited. The term phylogenomics was already coined in 1998 by Jonathan Eisen (also known under his twitter handle @phylogenomics), who outlined the importance of phylogenetic methods for the annotation of genes without relying on direct (time consuming) functional studies. This underlines how deeply embedded phylogenetic methods are in the field of genomics. The theoretical background for reconstructing gene trees (functional annotations) and species trees (reconstruction of the tree of life) is broadly overlapping. In this book I will introduce the major steps of phylogenomic analyses in general. The first two chapters briefly introduce the field of genomics (▶ Chap. 1, «Genomes») and the evolution and peculiarities of organellar genomes (▶ Chap. 2, «Organellar Genomes and Endosymbionts»). In ▶ Chap. 3 («Sequencing Techniques»), I review the most widely used sequencing platforms, which is difficult in a print format, as the field advances so fast that many numbers describing the output of these machines might be already out of date when you read this chapter. ▶ Chapter 4 («Sequencing Strategies») gives an overview of different strategies to sequence complete or partial genomes and transcriptomes. The outputs of every sequencing platform are sequences which are considerably shorter than chromosomes and in the case of short-read sequencing also shorter than most genes. In ▶ Chap. 5 («Assembly and Data Quality»), ways to puzzle these small pieces into more complete representations of genomes and genes (called assembly) are introduced. Fundamental steps for every phylogenomic study are alignments, read mapping and finding homologous genes, which are explained in ▶ Chaps. 6 («Alignment and Mapping») and 7 («Finding Genes»). Based on a sequence alignment, it is possible to reconstruct phylogenetic trees, and the methods are briefly reviewed in ▶ Chap. 8 («Phylogenetic Analyses»). I kept this chapter on purpose rather brief, as many excellent textbooks describing these methods (and its underlying algorithms) in detail are available (see references in ▶ Chap. 8). Moreover, the basic theory underlying these methods did not change much in the last decade. Surprisingly, even with this vast amount of data, many phylogenetic

questions remain still difficult to resolve. Some problems of phylogenetic reconstruction get even amplified when using hundreds or thousands of genes due to the presence of systematic error. ▶ Chapter 9 («Sources of Error and Incongruence in Phylogenomic Analyses») gives an overview of possible sources of error, as well as recommendations on how to deal with them. Moreover, the differences in analysing gene trees and species trees and possible sources of incongruence between those are outlined. Finally, in ▶ Chap. 10 («Rare Genomic Changes»), I introduce further phylogenetic markers apart from plain sequence data (e.g. integrations of mobile elements, gene order) and give an overview on how these rare genomic changes are utilized for phylogenetic systematics.

During my time at German universities, I was heavily involved in teaching bachelor and master level students. This included lectures, seminars and practical courses. While the field of molecular phylogenetics changed while moving into the postgenomic era, so did my courses. Besides the introduction of phylogenetic methods (e.g. maximum parsimony, maximum likelihood), I realized that more and more background knowledge became of major importance to carry out phylogenetic analyses. This includes knowledge about genomics, sequencing techniques as well as bioinformatic approaches to handle sequence data before the actual phylogenetic analysis starts. With this book I want to give a concise overview of all major steps of a phylogenomic analyses, as well as some insights into recent advantages in the field of genomics. This book is mainly addressed to undergraduate and graduate biology students, but also postdocs newly moving to the field of phylogenomics might use it as a first overview. The chapters are written in a concise way and focus more on explaining the idea behind methods, instead of deeply digging into the algorithmic or technical background. However, I tried always to refer to the appropriate specific literature to get deeper insights into any method (or study) of interest. Furthermore, I specified widely used and important software for every step of the phylogenetic analysis. When possible, I mention several alternatives. The name of software or scripts is always written in all caps, irrespective of the original way a name is written. This book does not include instructions on how to use this software, as in most cases detailed descriptions are available in the manual. As already noted, this book is mainly addressed to biology students. Working in the field of phylogenomics needs good to excellent (bio)informatic skills. Unfortunately, in the curriculum of many bachelor and master programmes, bioinformatics are not taught. However, several international courses teaching programming skills for (evolutionary) biologists take place regularly (e.g. Cold Spring Harbor Course «Programming for Biology»; Programming for Evolutionary Biology in Leipzig), and many excellent online tutorials are available. As such I can only strongly suggest to any student interested in this field to get used to work with Linux/Unix command lines and to acquire at least basic knowledge into (scripting) languages like Python, Perl or R.

I would like to thank several colleagues who commented on earlier versions of the here published chapters. In alphabetical order, they are Maite Aguado, Marie-Theres Gansauge, Michael Gerth, Iker Irisarri, Lars Podsiadlowski and Alexander Suh. I am grateful that Eva Nowack provided a picture of the enigmatic *Paulinella*. Moreover, I want to thank Lars Vogt, Christoph Held and Andreas Schmidt-Rhaesa for introducing

me into the theoretical and practical world of molecular phylogenetics. The above-mentioned university courses, which helped me to develop the outline and content of this book, were taught at the Free University of Berlin, University of Potsdam and University of Leipzig (in collaboration with Matthias Meyer from the Max Planck Institute for Evolutionary Anthropology). I would like to thank the department heads Thomas Bartolomaeus, Ralph Tiedemann and Martin Schlegel who gave me complete freedom in filling these courses with life.

Christoph Bleidorn
Madrid, Spain, January 2017

Contents

Abbreviations

µm	Micrometre
A	Adenine
AIC	Akaike information criterion
ATP	Adenosine triphosphate
BAC	Bacterial artificial chromosome
BI	Bayesian inference
BIC	Bayesian information criterion
BLAST	Basic Local Alignment Search Tool
bp	Base pairs
C	Cytosine
cDNA	Complementary DNA
CI	Cytoplasmatic incompatibility
CMOS	Complementary metal-oxide semiconductor
CNV	Copy number variation
CRISPR	Clustered regularly interspaced short palindromic repeat
ddNTP	Dideoxynucleoside triphosphate
DNA	Deoxyribonucleic acid
dNTP	Deoxynucleoside triphosphate
DUI	Doubly uniparental inheritance
G	Guanine
Gb	Giga base pairs
GBS	Genotyping by sequencing
GTR	General time-reversible model
GWAS	Genome-wide association study
HGT	Horizontal gene transfer
ICE	Integrative conjugative element
ILS	Incomplete lineage sorting
ISFET	Ion-sensitive field-effective transistor
Kb	Kilo base pairs
LBA	Long-branch attraction
LD	Linkage disequilibrium
LINE	Long interspersed element
LRT	Likelihood ratio test
LTR	Long terminal repeat
Mb	Mega base pairs
MCMC	Markov chain Monte Carlo method
MCMCMC	Metropolis-coupled Markov chain Monte Carlo method
MITE	Miniature inverted-repeat transposable element
ML	Maximum likelihood
MP	Maximum parsimony
mRNA	Messenger RNA
mya	Million years ago
NCBI	National Center for Biotechnology Information
NGS	Next-generation sequencing
NIP	Near intron pair
NJ	Neighbour joining
NNI	Nearest neighbour interchange
OTU	Operational taxonomic unit
PAM	Point accepted mutations
PCR	Polymerase chain reaction
PE	Paired-end sequencing
pH	Power of hydrogen
QTL	Quantitative trait loci
RNA	Ribonucleic acid
SINE	Short interspersed element
SMRT	Single-molecule real-time
SNP	Single nucleotide polymorphism
SPR	Subtree pruning and regrafting
T	Thymine
Tb	Tera base pairs
TBR	Tree bisection and reconnection
TE	Transposable element
TPRT	Target-primed reverse transcription
tRNA	Transfer RNA
UCE	Ultraconserved element
wgs	Whole-genome shotgun
ZMW	Zero-mode waveguide

Genomes

C. Bleidorn, *Phylogenomics*, DOI 10.1007/978-3-319-54064-1_1

1

- Life on earth can be largely classified into Bacteria, Archaea and Eukaryota.
- Eukaryotes likely arose by symbiogenic origin due to the fusion of an archaean with a bacterium.
- Bacteria and Archaea have compact genomes with uninterrupted genes, contained by a single, circular DNA molecule, located in the nucleoid.
- Eukaryote genomes are linearly organized into separate chromosomes, located in the nucleus, and contain genes interrupted by introns.
- Eukaryotes bear substantially larger genomes than archaeans and bacteria, but within eukaryotes there is no correlation between complexity and genome size.
- The human genome is around 3.3 Gb in size, but protein-coding genes and other functional DNA only make up a small proportion (<10%), whereas transposable elements are dominating (>44%).
- High-throughput sequencing of ancient human DNA allowed the reconstruction of archaic human genomes and led to the discovery of a hitherto unknown lineage, called Denisovan.

1.1 The Ring of Life

Life on earth was for a long time classified into two major groups, prokaryotes and eukaryotes (Stanier and van Niel 1962; Cavalier-Smith 2010). Prokaryotic cells are characterized by the lack of a true nucleus, absence of cell organelles and the genome is (usually) organized as a circular DNA molecule. Prokaryotic cells are usually small (<10 µm) and mostly unicellular, even though some photosynthetic bacteria form true multicellular chains (Flores and Herrero 2010). Besides the characterization due to all these absences of features, only prokaryotes show a coupling of translation and transcription. In this case, the translation of mRNA starts before transcription has been finished (Martin and Koonin 2006). In contrast, eukaryotic cells have their DNA organized on chromosomes located in a membrane-bound nucleus. With the exception of a few secondary losses, eukaryotes harbour (at least) mitochondria as cell organelles. Cell division is achieved due to mitosis, and meiosis, the prerequisite for sexual reproduction, likely was already present in the last common ancestor of eukaryotes (Ramesh et al. 2005). Eukaryotic cells are usually considerably bigger (>10 µm) than prokaryotic ones, and multicellularity evolved convergently in several major eukaryotic taxa. A strong increase in the number of investigated organisms recovered many exceptions to the here-mentioned features, blurring a clear distinction of «prokaryote-like» and «eukaryote-like» properties (Gregory and DeSalle 2005).

Distinguishing life into two major groups was challenged by a series of publications from the group of the American evolutionary microbiologist Carl Woese. Investigating ribosomal sequence data, they found profound distances between two prokaryote groups, now usually referred to as Bacteria and Archaea (Woese and Fox 1977; Fox et al. 1977; Balch et al. 1977). Being firstly predominantly discovered in extreme environments, Archaea have been since then found in virtually all environments and seem to be dominant in some forms of marine plankton. Moreover, they are the only organisms capable of methanogenesis (Gribaldo and Brochier-Armanet 2006). Fundamental differences between Bacteria and Archaea were confirmed in subsequent studies, leading to a new

classification of life into three domains, where Eukaryota represent the third one (Woese et al. 1990).

One of the defining features of eukaryotes is the possession of mitochondria. The primary function of these organelles is ATP synthesis through the oxidative electron transport chain, but also other functions are described (e.g. intracellular signalling). Similarities in the physiology and biochemistry of mitochondria with bacterial cells led to the endosymbiotic theory. According to this theory, mitochondria are of bacterial origin, an idea that dates back to a proposal from Ivan E. Wallin (1927). This hypothesis was later strongly advocated by Lynn Margulis (1970). Mitochondria still bear their own, circular genome, but massive transfer of mitochondrial genes to the host genome led to a strong size reduction. Phylogenetic analyses of mitochondrial genes recovered a close relationship with Alphaproteobacteria, thereby strongly supporting the endosymbiotic theory. The initial role of mitochondria in a symbiosis with its host and its environmental circumstances remains debated (Martin and Muller 1998; Wang and Wu 2014).

The three-domain hypothesis suggests the respective monophyly of Bacteria, Archaea and Eukaryota. In this case, these groups should include all descendent lineages of a common ancestor and only these. Phylogenomic analyses were used to investigate this question, and analyses based on a small set of core genes, which are present in all three groups and which are regarded as not been transferred horizontally between groups, recovered the three-domain tree (Ciccarelli et al. 2006). However, eukaryotic genomes contain genes with different origins (Williams et al. 2013). Analyses of gene families group eukaryotic genes either with Cyanobacteria, Alphaproteobacteria or within Archaea (Pisani et al. 2007). These results reflect the symbiotic origin of plastids from Cyanobacteria and the origin of mitochondria from Alphaproteobacteria and further suggest an origin of eukaryotes from an archaeal ancestor. A large-scale phylogenomic analysis including a newly discovered taxon called Lokiarchaeota provides further strong support for the hypothesis that the eukaryotic ancestor evolved from an archaeon (Spang et al. 2015). A subsequent study discovered several so far undescribed archaeans (named Asgard archaea), which group with eukaryotes (Zaremba-Niedzwiedzka et al. 2017). Furthermore, these archaeans bear several proteins, which had been regarded as eukaryote-specific, suggesting that the archaeal host contained many key components important for the control of eukaryotic cellular complexity. Considering emerging evidence from molecular phylogenetics, physiology, cell biology and palaeontology, a symbiogenic origin from the merger of an archaean and an alphaproteobacterium becomes obvious (McInerney et al. 2014). Phylogenetic analyses of eukaryote gene families support the symbiogenic origin of eukaryotes (Rochette et al. 2014). Lane and Martin (2012) suggested that mitochondria are a prerequisite for the evolution of complexity as seen in eukaryote cells. And finally, the fossil record suggests with 3.4 billion years (Wacey et al. 2011) a much older age for bacterial (or archaeal) lineages than for eukaryotes. The first fossilized eukaryotic cell dates 1.7–1.8 billion years ago (Rasmussen et al. 2008), which sets a possible time horizon for the merging event (McInerney et al. 2014). The symbiogenic origin of eukaryotes renders two of the domains paraphyletic. Instead, of being strictly bifurcating, the early tree of life seems to be better represented by a network or a ring (Fig. 1.1).

Sequencing of bacterial, archaeal and eukaryote genomes enabled the discovery of many important insights into the evolution, ecology and physiology of these organisms (Fraser et al. 2000; Galagan et al. 2005). However, there is a bias in available genome sequences in these groups. Whereas many taxa including model organisms, pathogens or

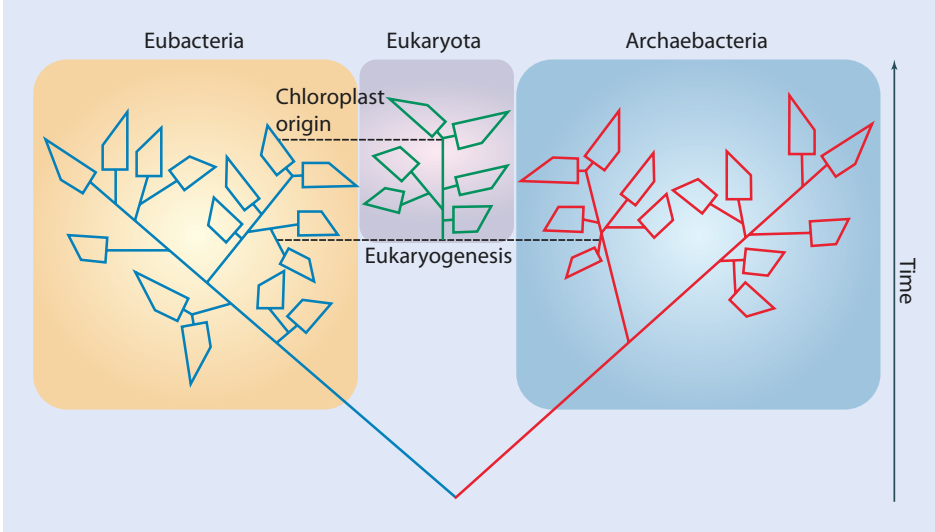


Fig. 1.1 The ring of life hypothesis (Reprinted by permission from Macmillan Publishers Ltd: Nature (McInerney et al. 2014), Copyright 2014)

organisms with economic importance are well investigated, other taxa are completely neglected. Consequently, a phylogeny-driven approach to cover genome sequencing across the whole tree of life has been proposed to fill these gaps (Wu et al. 2009; del Campo et al. 2014). Currently, major initiatives organize collaborative efforts in taxon-specific genome sequencing projects. Especially for animals, large-scale sequencing projects aim to sequence hundreds to thousands of nematode, arthropod, invertebrate and vertebrate genomes (Robinson et al. 2011; Genome 10K Community of Scientists 2009; Kumar et al. 2012; GIGA Community of Scientists 2014). Phylogenetic analyses of whole-genome or transcriptome data greatly improved our understanding of bacterial, archaeal and eukaryotic relationships. Backbone trees of bacterial and archaeal phylogenies are available and have been used to study the influence of horizontal gene transfer on the evolution of these groups (Nelson-Sathi et al. 2015; Lang et al. 2013; Wu et al. 2009; Groussin et al. 2016). Phylogenomic analyses of eukaryotes recover five major clades comprising their vast diversity (Fig. 1.2): (I) Archaeplastida (plants and green algae, red algae, glaucophytes); (II) the SAR clade representing stramenopiles, alveolates and Rhizaria; (III) Excavata; (IV) Amoebozoa; and (V) Opisthokonta, which unites fungi, choanoflagellates and animals (Katz and Grant 2014).

1.2 Genome Structure

There are profound differences between prokaryotes and eukaryotes in the structure and organization of their genomes, which in turn strongly influence the way to work with them in phylogenomic studies. Generally, prokaryote genomes are smaller and more compact than those of eukaryotes, clearly reducing the effort of sequencing and assembling them. However, due to the endosymbiotic origin of eukaryotes, it is obvious that a mosaic-like distribution for many of the features discussed below is found. Most

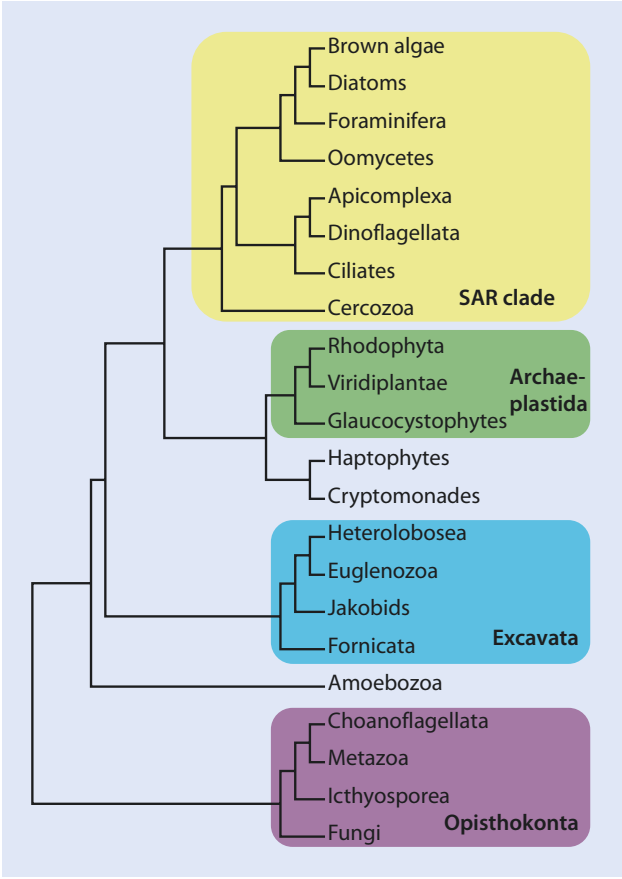


Fig. 1.2 Phylogenetic relationships of eukaryotes based on the phylogenomic analyses of Katz and Grant (2014)

genomes of bacteria and archaeans are contained by a single, circular DNA molecule, located in the nucleoid. For packaging, the double-stranded DNA molecule is supercoiled, which is facilitated by DNA-binding proteins. Whereas in bacteria the supercoiling is achieved by proteins like DNA gyrase, DNA topoisomerase I and HU proteins, archaeans have proteins for packaging that are similar to the histones of eukaryotes (White and Bell 2002). Exceptions from these general patterns exist, and, e.g. some members of the bacterial taxa spirochaetes and actinomycetes show linearly organized genomes (Hinnebusch and Tilly 1993). Multipartite genomes are not unusual across prokaryotes as well (Harrison et al. 2010). Eukaryote genomes are linearly organized into separate chromosomes. Within chromosomes the DNA forms nucleosomes due to association with histone proteins for packaging. Further on, chromosomes bear centromeres and telomeres. Centromeres are characterized by a special set of proteins which form the attachment point for microtubules during cell division. Telomeres are the cap of the chromosome ends and are characterized by the presence of repetitive DNA motifs (Brown 2007).

Prokaryotes often have a high potential for horizontal gene transfer (HGT) by mobile genetic elements. Movement of DNA can be facilitated by transformation, conjugation or transduction. In the case of transformation, cellular DNA is taken up by the recipient due to the presence of special proteins. Conjugation is gene transfer mediated by plasmids

or so-called integrative conjugative elements (ICEs) via contact between donor recipient cells. Finally, transduction is gene transfer by bacteriophages (Frost et al. 2005). The presence of extrachromosomal elements such as plasmids, which usually carry accessory (but not essential) genes, and the frequent occurrence of HGT lead to the phenomenon that within prokaryotic species often large differences in gene content are found. This led to the formulation of the pan-genome concept. A pan-genome is composed of two parts: a «core genome», containing the genes present in all strains of a prokaryotic species, and the «dispensable genome» summarizing the genes which occur in a subset of strains or only one (Medini et al. 2005). Most archaeal and many bacterial genomes bear clustered regularly interspaced short palindromic repeats (CRISPRs). Together with associated proteins (CAS) these repeats constitute an adaptive immune system that can target invading bacteriophages or conjugative plasmids (Horvath and Barrangou 2010; Burstein et al. 2016). Plasmids are also occurring in some eukaryotes, e.g. in yeast and other fungi (Hausner 2003).

Prokaryotic genomes are usually compactly organized, with a small proportion of non-coding intragenic DNA. Consequently, prokaryotic genomes are relatively small, rarely exceeding sizes of 10 Mb. The smallest known genomes are reported for endosymbiotic bacteria, with the betaproteobacterium *Candidatus* Tremblaya princeps as record holder with its only 139 Kb genome. Bacteria with extremely reduced genomes are dependent on genes from their host or from other co-occurring endosymbionts (Husnik et al. 2013; McCutcheon and Moran 2012). Genome sizes of eukaryotes are more variable and can exceed several hundred Gb (see 1.3 for more details). Not only are the genomes of prokaryotes smaller than those of eukaryotes but also their genes. The mean protein length is 40–60% higher in eukaryotes than in prokaryotes, and this holds true across different functional classes of proteins (Zhang 2000; Brocchieri and Karlin 2005). Moreover, prokaryote genes are not interrupted by spliceosomal introns, which are typical for eukaryote nuclear genomes (Roy and Gilbert 2006). For example, human genes are interrupted in average by nine introns, and intronic sequences make up a substantial amount of the complete genome (Venter et al. 2001). Spliceosomal introns exhibit special sequence motifs and are removed before transcription by the spliceosome, which is formed by five small RNAs and over 200 proteins (Irimia and Roy 2014). However, other types of introns can be found in prokaryotes. Group II introns are self-splicing introns that have been reported in ~25% of all sequenced bacterial genomes, but always in low frequency. Moreover, they are also found in eukaryote organelle genomes, but are only known from few archaeal genomes, which likely originate from horizontal transfer from bacteria (Lambowitz and Zimmerly 2011). Other types of introns are more rare and often restricted to certain types of genes (e.g. tRNAs), but can also be found across all organisms (Irimia and Roy 2014).

Eukaryote genomes often carry a huge proportion of interspersed elements and tandem repeats. Both types are usually rare or completely absent in prokaryotic genomes. Tandemly repeated DNA, which is sometimes called satellite DNA, can be found around centromeres or randomly scattered across chromosomes. Tandem repeats with short repetitive motifs are known as mini- and microsatellites (Brown 2007). Interspersed elements have the ability to integrate into new sites of the genome of their origin, often in a random pattern, even though many transposons show the preference for a specific target site. These transposable elements are historically classified according to their mode of transposition into retrotransposons (class I) and DNA transposons (class II) (Finnegan 1989). Such elements altogether often contribute massively to the genome size of eukaryotes (Kazazian

2004). DNA transposons are mobile elements transposed by a cut-and-paste mechanism, where they are excised from one genomic site and integrated into a new one. These elements usually encode a transposase and bear terminal inverted repeats. Ten different superfamilies of eukaryotic cut-and-paste DNA transposons are currently distinguished, which show an enormous variation in their distribution across taxa (Wicker et al. 2007). Two further groups of DNA transposons (*Helitrons*, *Mavericks*) likely use copy-and-paste mechanisms for their spread across genomes (Feschotte and Pritham 2007). In contrast to DNA transposons, retrotransposons are transcribed into RNA and subsequently reverse transcribed and copied into the genome (copy and paste), leading to a duplication of the element. Some autonomous retrotransposons bear long terminal repeats (LTRs) at their ends. These LTR retrotranposons encode for several specific genes including a reverse transcriptase and integrase, and they are generally similar to retroviruses, with which they share their replication mechanism (Kazazian 2004). It should be mentioned that there is no real distinction between LTR retrotransposons and retroviruses, as exogenous retroviruses can easily become endogenous by losing their *env* gene, which produces the protein on the surface of the viral particle that is responsible for cell entry (Magiorkinis et al. 2012). Other autonomous retrotransposons lack the LTRs and use a different copy-and-paste mechanism than LTR retrotransposons, namely, target-primed reverse transcription (TPRT) (Luan et al. 1993). Autonomous non-LTR retrotransposons, which are also called LINEs (long interspersed elements), such as L1 elements, constitute a high proportion of the human genome (see below). In contrast, nonautonomous non-LTR retrotransposons lack coding capacity for genes needed for their retrotransposition. These elements are commonly referred to as SINEs (short interspersed elements) and mostly range in length between 100 and 500 bp. SINEs are transcribed by RNA polymerase III, for which they contain a promoter in their sequence. For reverse transcription, they have to be bound by the reverse transcriptase of a LINE, and they are subsequently integrated into a new genomic location via TPRT (Kramerov and Vassetzky 2011). SINEs classified as *Alu* elements show the highest copy number of all transposable elements in humans (Batzer and Deininger 2002). DNA transposons are frequently found in both eukaryotes and prokaryotes and are frequently transferred horizontally (Gilbert et al. 2010). Retrotransposons are usually restricted to eukaryotes, and their horizontal transfer is less frequent, except for the RTE superfamily of LINEs (Suh et al. 2016).

1.3 Genome Size

The genome size of an organism can be measured by the c-value, which describes the mass of DNA content of a haploid cell in picogram (pg). A c-value of 1 pg equals ~978 Mb (Dolezel et al. 2003). Bacterial and archaeal genomes are usually rather small, but within eukaryotes genome size shows huge variations with differences that can exceed 10,000–100,000 folds in pairwise comparisons (Fig. 1.3). However, it seems that there is no relation between the complexity of an organism (e.g. defined by the number of different cell types) and its genome size, a conundrum which is known as the «c-value paradox» (Thomas 1971; Gregory 2001). For example, the canopy plant *Paris japonica* has a c-value of ~133 pg, more than 35× bigger than that of humans (~3.5 pg) (Pellicer et al. 2010). As it has been shown by genome sequencing projects, eukaryotic genomes often contain only small amounts of coding or functional DNA, and the large genome size in eukaryotes is usually due to huge amounts of mobile elements (Lynch 2007).

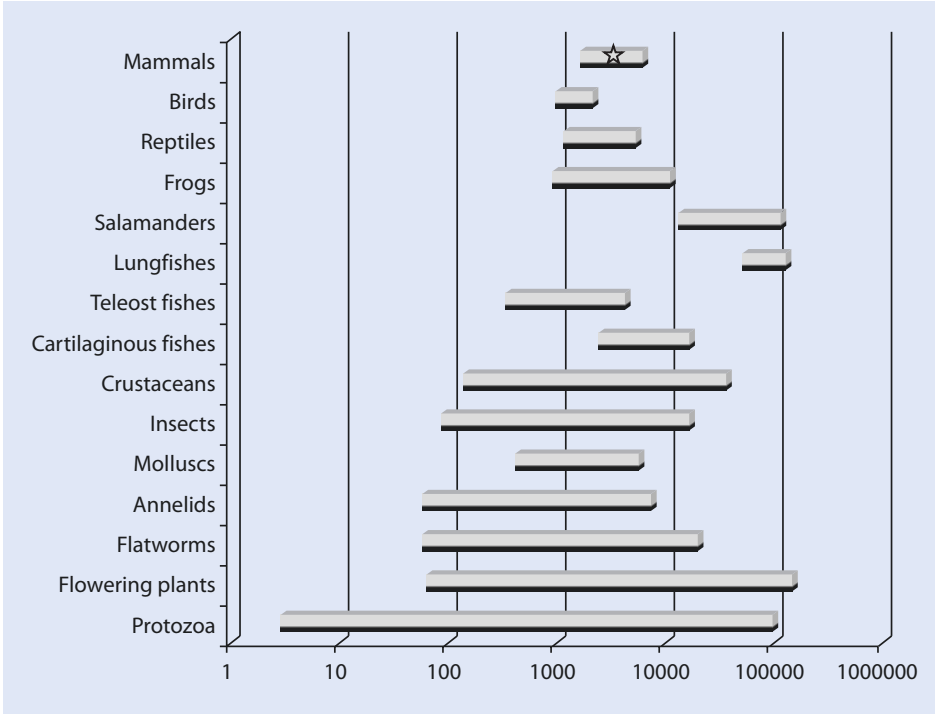


Fig. 1.3 Variation of genome size (given in Kb) across eukaryotes (Reprinted from Palazzo and Gregory (2014))

Several evolutionary hypotheses have been formulated to explain the huge differences in genome size between organisms. The selfish DNA hypothesis states that non-coding DNA is a by-product of «selfish» transposable elements (Orgel and Crick 1980; Doolittle and Sapienza 1980). The «bulk DNA» hypothesis assumes that total DNA content is a direct product of natural selection (Cavalier-Smith 1978). In contrast, a non-adaptive view is favoured by the «mutational burden» hypothesis (Lynch 2006). According to this view, excessive DNA is regarded as mutational burden, where purifying selection will eliminate deleterious genomic elements from populations. As the efficiency of selection is strongest in large populations, this hypothesis aims to explain why prokaryotes, which usually occur in much larger (long-term) populations than eukaryotes, have more compact genomes than eukaryotes (Lynch 2007). Inversely, the lack of expansion and restructuring of prokaryote genomes could also explain the absencve of complex morphologies among them (Lynch and Conery 2003). However, in several cases, differences in genome size of eukaryotes show a correlation with body size, metabolism or development (Gregory 2013).

Besides a lack of correlation between genome size and complexity, there seems also to be no relationship between complexity and gene number, sometimes termed «g-value paradox» (Hahn and Wray 2002). While the definition of a gene remains controversial, comparisons of the amount of protein-coding base pairs with organismic complexity

similarly show no correlation. However, it seems that the amount of non-protein-coding sequences (e.g. various RNAs; ► see Infobox 1.1) increases consistently in more complex organisms (Taft et al. 2007). Consequently, differences in gene regulation, interaction of genes, alternative splicing and differential expression contribute to explain the g-value paradox (Gregory 2013).

Infobox 1.1

The Variety of Non-coding RNAs

Non-coding RNAs comprise RNAs that do not encode proteins. Well known are ribosomal RNAs and tRNAs, which all play a vital role in protein biosynthesis. Many other classes of RNAs are involved in the regulation of gene expression, transcription, splicing or editing (Mattick and Makunin 2006). Several classes of such RNAs are recent discoveries, with some of them incompletely characterized in their biological role. An overview of some important RNAs is given here:

microRNAs microRNAs are short (~22 bp) non-coding RNAs found in animals and plants which are involved in the regulation of gene expression (Ambros 2004). Mature microRNAs were shown to be highly conserved across animal taxa, and several hundred distinct microRNA families have been reported for Metazoa (Kozomara and Griffiths-Jones 2011). A typical role of microRNAs is that they guide molecules involved in post-transcriptional gene silencing by pairing them with target mRNAs, leading to their cleavage or repression. The expression of many microRNAs is known to be tissue-specific, and, additionally, the disparity of microRNAs of a given animal taxon can often be linked to its morphological complexity (Sempere et al. 2006).

piRNAs piRNAs are small non-coding RNAs that interact with Piwi proteins (Aravin et al. 2006). In contrast to microRNAs, piRNAs are slightly longer (24–31 bp) and are derived from single-stranded precursors originating from repetitive sequences in the genome. So-called piRNA-induced silencing complexes are able to repress transposon activity, thereby maintaining the genome integrity of the germ line (Iwasaki et al. 2015). Additionally, in some organisms piRNAs also function in the regulation of cellular genes.

snoRNAs Small nucleolar (sno) RNAs are an abundant class of RNAs present in the nucleolus of eukaryotes of approximately 60–300 bp length. According to their secondary structure and the presence of specific sequence motifs, snoRNAs can be classified into two major groups: C/D and H/ACA snoRNAs (Kiss 2002). Usually, snoRNAs are components of ribonucleoprotein complexes where they provide a scaffold to assemble partner proteins. Moreover, they guide for the recognition of target DNAs and sites of post-transcriptional modification (Bratkovič and Rogelj 2014). Modifications include methylation of DNAs and pseudouridylation of RNAs, and this system is found in eukaryotes and archaeans (Reichow et al. 2007).

lncRNAs RNA transcripts of >200 nt size which lack an open reading frame are summarized as long non-coding (lnc) RNAs. Especially multicellular organisms seem to pervasively transcribe different types of this heterogeneous class of RNAs, for which a specific function is often not understood. According to the place of expression, cytoplasmic and nuclear lncRNAs can be distinguished (Fatica and Bozzoni 2014). Important roles in the control of gene expression during developmental processes are known for some lncRNAs, e.g. dosage compensation, epigenetic imprinting or cell differentiation. Thousands of tissue-specific lncRNAs are catalogued, and RNA-RNA, RNA-DNA as well as RNA-protein interactions have been reported (Quinn et al. 2014). In vertebrates, transposable elements are found in a large proportion of lncRNAs and also make up a substantial part of their sequence length (Kapusta et al. 2013).

1

1.4 The Genomes of Modern and Archaic Humans

In 1990 an ambitious collaborative project was launched to sequence the human genome (Watson 1990). After finishing the mapping of the genome, sequencing of organisms with smaller genomes was conducted as proof of principle for the method. The final sequencing of the human genome was carried out by the International Human Genome Sequencing Consortium (IHGSC) involving 20 major institutions in six countries (International Human Genome Sequencing Consortium 2004). In the mid-1990s, a team around Craig Venter simultaneously started sequencing the human genome using whole-genome shotgun sequencing coupled with a high-throughput Sanger sequencing approach. Both groups published draft genomes for an initial view of the human genome in 2001 (International Human Genome Sequencing Consortium 2001; Venter et al. 2001). These drafts still lacked ~10% of the euchromatic regions, and contigs (contiguous segments of DNA) were separated by a huge number of gaps. A more complete human genome sequence assembly covering 99% of the euchromatic regions and including less gaps was published in 2004 (International Human Genome Sequencing Consortium 2004). However, several repeat-rich regions still remained difficult to assemble, and it needed long-read sequencing to close at least half of the remaining gaps (Chaisson et al. 2015). Several new assemblies and annotations of the human genome were published since its official final completion. The size of the human genome varies in its estimations between 3.1 and 3.3 Gb. According to Gencode v25 (► www.gencodegenes.org), the genome contains 19,950 protein-coding genes; 15,767 long non-coding RNAs; and 7258 small RNAs. Besides this, a small fraction of the genome contains regulatory regions controlling gene expression, replication origins, telomeres and centromeres. This means that exonic DNA of protein-coding genes represent only around 1.5% of the human DNA and in total essential/functional DNA does not exceed 10% of the genomic DNA. The majority of the human genome comprises intron sequences and transposable elements (TE), the latter make up by far the largest part of the genome, including SINEs, LINEs, endogenous retroviruses and DNA transposons. Most of these TEs are not active anymore and therefore often highly degenerated, making it difficult to estimate the proportion of TE-derived sequences. Recent approximations vary between 45% (Cordaux and Batzer 2009) and 75% (de Koning et al. 2011). The most abundant TEs are L1 elements and *Alu* elements, with the latter exceeding more than one million copies (Cordaux and Batzer 2009). Finally, altogether, 14,650 pseudogenes were recognized, with the majority of them being processed (Gencode v25), indicative of originating via L1-mediated reverse transcription (Esnault et al. 2000).

Two large projects building on the finalized sequence data were initiated to investigate the genetic diversity (HapMap) and functionality of the human genome (ENCODE). The goal of the HapMap project was to determine common patterns of sequence variation among different human populations (The International HapMap Consortium 2003), and a first haploid diversity map was published in 2005 (The International HapMap Consortium 2005). Fuelled by the availability of new and more powerful sequencing techniques, the cataloguing of single-nucleotide polymorphisms was extended to the analysis of more than 2500 human genomes from 26 populations (The 1000 Genomes Project Consortium 2015). Additionally, copy number variations (CNVs) of larger DNA segments, which can alter the diploid status of the DNA, have been compiled (Zarrei et al. 2015). Human genomes have been found more variable than initially thought, exceeding over 1% of differences in cross-comparisons (Pang et al. 2010). The availability of this data significantly

changed the way to investigate the origin of diseases and complex traits by conducting genome-wide association studies (GWAS), where phenotypes are correlated with genetic variation (Naidoo et al. 2011). Generally, GWAS try to identify SNPs which exhibit linkage disequilibrium (LD), meaning alleles at two or more loci show a non-random association (Slatkin 2008). The HapMap project confirmed that many chromosomal regions of the human genome consist of nonoverlapping sets of loci with strong LD, called haplotype blocks, which can exceed sizes of more than 100 Kb. Consequently, using SNPs which are overrepresented in correlation with the investigated phenotype and show strong LD makes it possible to detect larger genomic regions associated with a phenotype (Visscher et al. 2012). It is possible to calculate the probability that a single SNP is correlated with a phenotype (e.g. a disease) called the odds ratio. GWAS gained huge popularity, and «human genetic variation» was elected as breakthrough of the year by the journal *Science* in 2007 (Pennisi 2007), a year in which nearly 100 studies using this approach were published.

However, GWAS became strongly criticized in the scientific community, and some researchers questioned their relevance at all (Visscher et al. 2012). In case of diseases, the rationale of GWAS is that common diseases are partly (and additively) and sizeably attributable to SNPs which are represented in more than 1–5% of the population. This hypothesis is called «common disease-common variant» (CD/CV) model. Using this background the observed phenotypic variation is associated with a set of SNPs in GWAS. However, the validity of basic assumptions of GWAS became questioned when researchers found that most of the phenotypic variability seems to remain unexplained in these studies. For example, 32 loci have been identified to affect Crohn's disease risk using GWAS, but they seem to explain only 20% of the heritability of the disease (Barrett et al. 2008). Obviously, still 80% of the variance of the phenotype remains unexplained. This phenomenon has been called missing heritability (Maher 2008). The missing heritability is even more pronounced for most published GWAS. Possible reasons for the failure of GWAS include sampling errors (investigation of too few SNPs) or model misspecifications in the subsequent statistical analysis (Marjoram et al. 2014). As such, the assumption of most GWAS of additive genetic variance, which basically means that each SNP contributes part of the heritability and can be added together, ignores evidence that gene-gene interactions can be highly complex (epistasis) and non-additive (McKinney and Pajewski 2012). Moreover, environmental influences on the transcription of genes have often been neglected, too. However, even inclusion of such data in the statistical framework of GWAS seems not to significantly improve their explanatory power (Aschard et al. 2012). Therefore, for future studies, a shift from the discovery of SNPs or genes associated with a given phenotype to functional assays investigating the biological mechanisms of these genotype-phenotype associations seems important (Shendure 2014).

A massive project to categorize all functional elements in the human genome is ENCODE, the encyclopaedia of DNA elements project. An initial pilot project investigated the functionality of 1% of the human genome and was followed up by the main study covering most of the genome (Consortium TEp 2007; ENCODE 2012). Using a huge array of methodologies, focussing on gene annotation, transcriptome analyses, chromatin analyses, transcription factor binding, methylation and protein conformation, the biochemical functionality of the human genome was documented. From this study it became obvious that the organization of the human genome is even more complex than previously anticipated. For example, it was found that genes and their regulatory elements can form complex networks and are engaged in interactions over a long genomic range

(Sanyal et al. 2012). Re-annotation of the genome discovered many new small RNAs (e.g. microRNAs, snoRNAs, etc.; ▶ see Infobox 1.1), and many of these RNAs overlap with coding transcripts (Djebali et al. 2012). Confirming previous studies (Kapranov et al. 2007), a pervasive transcription of the genome has been recorded, which means that most of the DNA is at least found in one transcript (Djebali et al. 2012). This shows that the transcriptome is not only derived from protein-coding genes and short non-coding RNAs and such a pattern seems to be common for eukaryote genomes in general (Berretta and Morillon 2009). The numbers of how much of the human genome is transcribed vary between studies, are strongly dependent on the investigated cell type and exceed 85% at the higher end (Hangauer et al. 2013). These studies uncovered a high number of previously undetected long non-coding RNAs (▶ see also Infobox 1.1). Biochemical activity of most part of the genome was also found using other types of experiments, leading to the suggestion that indeed ~80% of the genome is functional (ENCODE 2012). DNA elements classified as functional include those which are either transcribed, associated with modified histones, bind to a transcription factor, show signs of CpG methylation, or are found in open-chromatin areas. This result came as a big surprise, as it was considered that only ~10% of the human genome is functional and the rest of the DNA was classified as «junk». Without surprise, this bold claim led to a huge controversy focussing on problems with the methodology and the definition of the term function (Graur et al. 2013; Doolittle 2013; Kellis et al. 2014; Palazzo and Gregory 2014). The term «junk DNA» goes back to Ohno (1972) who recognized the small proportion of DNA coding for genes in the human genome. Some researches prefer to use a less polarizing description and favour to use «non-functional DNA» which has no or little selective advantage for the organism (Eddy 2012). Obviously, TEs and intron sequences, which make up a huge percentage of the biochemical activity detected in the ENCODE study, would qualify as non-functional DNA under this evolutionary definition. Moreover, as most TEs contain promotor region, it lies in the nature of these elements to be transcribed, which often is achieved in a random fashion. Fittingly, comparative genomic studies conclude that only 5–15% of the human genome can be regarded as functional regarding a criterion of evolutionary conservation (Lindblad-Toh et al. 2011; Meader et al. 2010).

The sequenced human genome became also an important source of data to understand human evolution. Humans are closely related to chimpanzees and bonobos, and this group together forms the sister clade of gorillas. According to time calibrations using molecular data, the human-chimpanzee split dates back ~6.5–9.3 mya, which is in line with the fossil record suggesting ~6.5–10 mya (Moorjani et al. 2016). In contrast to their primate relatives, humans are able to manufacture complex tools and use a complex language for information transfer (Pääbo 2014). Anatomically modern humans (*Homo sapiens*) appeared ~200,000 years ago in Africa, from where according to the well-supported out-of-Africa hypothesis they colonized all continents. In line with this hypothesis, African populations show higher genetic diversity then non-African populations (Henn et al. 2012). Thanks to the advent of ancient DNA techniques and high-throughput sequencing techniques, the field of palaeogenomics flourished. Ancient DNA analyses allowed studying the change of genetic diversity through time and to clarify evolutionary hypothesis based on fossils. Initial ancient DNA studies were mostly limited to high copy number genes as, e.g. derived from mitochondria (Shapiro and Hofreiter 2014). However, improved sequencing library construction methods and the massive output of Illumina short-read sequencers made it possible to sequence genomes of archaic humans in a coverage and quality of modern DNA (Meyer et al.

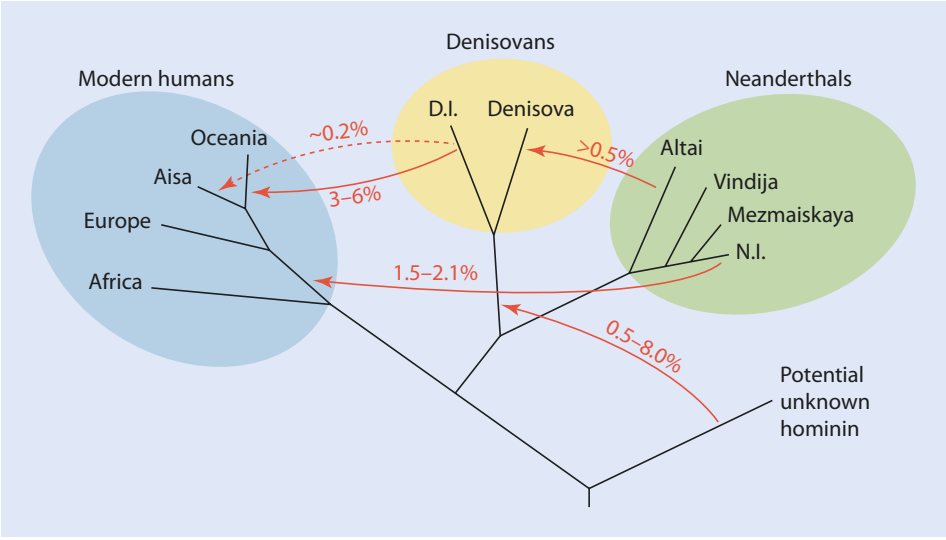


Fig. 1.4 Human relationships and possible model of gene flow. Note that the age of archaic genomes does not allow detection of gene flow of modern humans towards them (Reprinted by permission from Macmillan Publishers Ltd: Nature (Prüfer et al. 2014), Copyright 2014)

2012; Prüfer et al. 2014). Nuclear genomes of two lineages of archaic humans which lived contemporary with modern humans have been sequenced: Neanderthal and Denisovan (Pääbo 2014).

Neanderthals are well known from the fossil record, which shows them appearing ~300,000 years ago and getting extinct ~30,000 years ago. In contrast, the Denisovan lineage has not been recognized by analysing fossils and has been firstly described based on its divergent genomic data (Krause et al. 2010; Reich et al. 2010). Both lineages seem to represent sister groups (Fig. 1.4), and comparative genomic analyses recovered DNA segments stemming from these lineages in modern humans. Interestingly, Neanderthal-derived DNA is only found in non-African populations, and recent analyses suggest at least two hybridization events with modern humans. Moreover, introgressed DNA supports a hybridization event of Denisovan with modern human populations colonizing Papua New Guinea and Australia (Fig. 1.3) (Prüfer et al. 2014). Additionally, they were able to sequence complete mitochondrial genomes and some nuclear sequence from human fossils found in the Sima de los Huesos near Burgos in Spain (Meyer et al. 2014, 2016). This site is known to harbour the oldest European hominin fossils. Many of them date back more than ~300,000 years ago and are affiliated with *Homo heidelbergensis*. The sequenced fossil dates back ~400,000 years ago, making it the oldest sequenced hominin ancient DNA, opening a complete new window to understand human evolutionary history. Interestingly, this mitochondrial genome is closest to the one of Denisovan in a phylogenetic analysis, even though analyses of nuclear genes show a close relationship to Neanderthals (Meyer et al. 2016).

Comparative and population genetic studies of all sequenced human and archaic genomes allow many interesting insights into our evolutionary history. Analysing diploid genome data with Bayesian approaches helps to infer the population size change over time. Such analyses find a severe decrease in the size of all human populations around

200,000 years ago. Whereas both Denisovan and Neanderthal went extinct in the last 30,000 years, a huge increase of population sizes of the modern human can be observed for the same time (Prüfer et al. 2014). Comparative analyses show that recent human genomes of non-African populations carry around 2% DNA with Neanderthal ancestry. Genome-wide searches in hundreds of modern human genomes enable recovering ~20% of the Neanderthal genome (Vernot and Akey 2014). This analysis shows that Neanderthal-derived DNA contributed to loci adaptive for skin phenotypes. Moreover, Neanderthal alleles related to the immune system of modern humans seem to be positively selected and can rise to high frequencies in some populations (Abi-Rached et al. 2011). Similarly, positively selected haplotypes related to altitude adaptation in Tibetans likely stem from introgression of Denisovan-like DNA (Huerta-Sanchez et al. 2014). Comparative analyses also allow identifying those positions in the modern human genome which changed since the split from Neanderthal and Denisovan. More than 30,000 SNPs specific for modern humans have been identified so far, of which ~10% are found in putatively regulatory regions. In the future, functional studies investigating these genetic variants will help to find those changes which might be functionally significant (Pääbo 2014).

References

Abi-Rached L, Jobin MJ, Kulkarni S, McWhinnie A, Dalva K, Gragert L, Babrzadeh F, Gharizadeh B, Luo M, Plummer FA, Kimani J, Carrington M, Middleton D, Rajalingam R, Beksac M, Marsh SGE, Maiers M, Guethlein LA, Tavoularis S, Little A-M, Green RE, Norman PJ, Parham P (2011) The shaping of modern human immune systems by multiregional admixture with archaic humans. Science 334:89–94

Ambros V (2004) The functions of animal microRNAs. Nature 431:350–355

Aravin A, Gaidatzis D, Pfeffer S, Lagos-Quintana M, Landgraf P, Iovino N, Morris P, Brownstein MJ, Kuramochi-Miyagawa S, Nakano T, Chien M, Russo JJ, Ju J, Sheridan R, Sander C, Zavolan M, Tuschl T (2006) A novel class of small RNAs bind to MILI protein in mouse testes. Nature 442:203–207

Aschard H, Chen J, Cornelis Marilyn C, Chibnik Lori B, Karlson Elizabeth W, Kraft P (2012) Inclusion of gene-gene and gene-environment interactions unlikely to dramatically improve risk prediction for complex diseases. Am J Hum Genet 90:962–972

Balch W, Magrum L, Fox G, Wolfe R, Woese C (1977) An ancient divergence among the bacteria. J Mol Evol 9:305–311

Barrett JC, Hansoul S, Nicolae DL, Cho JH, Duerr RH, Rioux JD, Brant SR, Silverberg MS, Taylor KD, Barmada MM, Bitton A, Dassopoulos T, Datta LW, Green T, Griffiths AM, Kistner EO, Murtha MT, Regueiro MD, Rotter JI, Schumm LP, Steinhart AH, Targan SR, Xavier RJ, Libioulle C, Sandor C, Lathrop M, Belaiche J, Dewit O, Gut I, Heath S, Laukens D, Mni M, Rutgeerts P, Van Gossum A, Zelenika D, Franchimont D, Hugot J-P, de Vos M, Vermeire S, Louis E, Cardon LR, Anderson CA, Drummond H, Nimmo E, Ahmad T, Prescott NJ, Onnie CM, Fisher SA, Marchini J, Ghori J, Bumpstead S, Gwilliam R, Tremelling M, Deloukas P, Mansfield J, Jewell D, Satsangi J, Mathew CG, Parkes M, Georges M, Daly MJ (2008) Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. Nat Genet 40:955–962

Batzer MA, Deininger PL (2002) Alu repeats and human genomic diversity. Nat Rev Genet 3:370–379

Berretta J, Morillon A (2009) Pervasive transcription constitutes a new level of eukaryotic genome regulation. EMBO Rep 10:973–982

Bratkovič T, Rogelj B (2014) The many faces of small nucleolar RNAs. Biochim Biophys Acta Gene Regul Mech 1839:438–443

Brocchieri L, Karlin S (2005) Protein length in eukaryotic and prokaryotic proteomes. Nucleic Acids Res 33:3390–3400

Brown T (2007) Genomes 3. Garland Science Publisher, New York

Burstein D, Harrington LB, Strutt SC, Probst AJ, Anantharaman K, Thomas BC, Doudna JA, Banfield JF (2016) New CRISPR–cas systems from uncultivated microbes. Nature advance online publication

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 34:247–278

References

Cavalier-Smith T (2010) Deep phylogeny, ancestral groups and the four ages of life. Philos Trans R Soc Lond Ser B Biol Sci 365:111–132

Chaisson MJP, Huddleston J, Dennis MY, Sudmant PH, Malig M, Hormozdiari F, Antonacci F, Surti U, Sandstrom R, Boitano M, Landolin JM, Stamatoyannopoulos JA, Hunkapiller MW, Korlach J, Eichler EE (2015) Resolving the complexity of the human genome using single-molecule sequencing. Nature 517:608–611

Ciccarelli FD, Doerks T, von Mering C, Creevey CJ, Snel B, Bork P (2006) Toward automatic reconstruction of a highly resolved tree of life. Science 311:1283–1287

Consortium TEp (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature 447:799–816

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

de Koning APJ, Gu W, Castoe TA, Batzer MA, Pollock DD (2011) Repetitive elements may comprise over two-thirds of the human genome. PLoS Genet 7:e1002384

del Campo J, Sieracki ME, Molestina R, Keeling P, Massana R, Ruiz-Trillo I (2014) The others: our biased perspective of eukaryotic genomes. Trends Ecol Evol 29:252–259

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, Roder 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 L-H, 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, Guigo R, Gingeras TR (2012) Landscape of transcription in human cells. Nature 489:101–108

Dolezel J, Bartos J, Voglmayr H, Greilhuber J (2003) Nuclear DNA content and genome size of trout and human. Cytometry 51A:127–128

Doolittle WF (2013) Is junk DNA bunk? A critique of ENCODE. Proc Natl Acad Sci U S A 110:5294–5300

Doolittle WF, Sapienza C (1980) Selfish genes, the phenotype paradigm and genome evolution. Nature 284:601–603

Eddy SR (2012) The C-value paradox, junk DNA and ENCODE. Curr Biol 22:R898–R899

ENCODE (2012) An integrated encyclopedia of DNA elements in the human genome. Nature 489:57–74

Esnault C, Maestre J, Heidmann T (2000) Human LINE retrotransposons generate processed pseudogenes. Nat Genet 24(4):363–367

Fatica A, Bozzoni I (2014) Long non-coding RNAs: new players in cell differentiation and development. Nat Rev Genet 15:7–21

Feschotte C, Pritham EJ (2007) DNA transposons and the evolution of eukaryotic genomes. Annu Rev Genet 41:331–368

Finnegan DJ (1989) Eukaryotic transposable elements and genome evolution. Trends Genet 5:103–107

Flores E, Herrero A (2010) Compartmentalized function through cell differentiation in filamentous cyanobacteria. Nat Rev Microbiol 8:39–50

Fox GE, Magrum LJ, Balch WE, Wolfe RS, Woese CR (1977) Classification of methanogenic bacteria by 16S ribosomal RNA characterization. Proc Natl Acad Sci U S A 74:4537–4541

Fraser CM, Eisen JA, Salzberg SL (2000) Microbial genome sequencing. Nature 406:799–803

Frost LS, Leplae R, Summers AO, Toussaint A (2005) Mobile genetic elements: the agents of open source evolution. Nat Rev Microbiol 3:722–732

Galagan JE, Henn MR, Ma L-J, Cuomo CA, Birren B (2005) Genomics of the fungal kingdom: insights into eukaryotic biology. Genome Res 15:1620–1631

Genome 10K Community of Scientists (2009) Genome 10 K: a proposal to obtain whole-genome sequence for 10 000 vertebrate species. J Hered 100:659–674

GIGA Community of Scientists (2014) The global invertebrate genomics alliance (GIGA): developing community resources to study diverse invertebrate genomes. J Hered 105:1–18

Gilbert C, Schaack S, Pace Ii JK, Brindley PJ, Feschotte C (2010) A role for host-parasite interactions in the horizontal transfer of transposons across phyla. Nature 464:1347–1350

Graur D, Zheng Y, Price N, Azevedo RBR, Zufall RA, Elhaik E (2013) On the immortality of television sets: «function» in the human genome according to the evolution-free gospel of ENCODE. Genome Biol Evol 5:578–590

Gregory R (2013) Molecules and macroevolution: a Gouldian view of the genome. In: Danieli G, Minelli A, Pievani T (eds) Stephen J. Gould: the scientific legacy. Springer, Milan, pp 53–70
Gregory R, DeSalle R (2005) Comparative genomics in prokaryotes. In: Gregory R (ed) The evolution of the genome. Elsevier, Burlington, pp 586–675
Gregory TR (2001) Coincidence, coevolution, or causation? DNA content, cell size, and the C-value enigma. Biol Rev 76:65–101
Gribaldo S, Brochier-Armanet C (2006) The origin and evolution of archaea: a state of the art. Philos Trans R Soc Lond Ser B Biol Sci 361:1007–1022
Groussin M, Boussau B, Szöllősi G, Eme L, Gouy M, Brochier-Armanet C, Daubin V (2016) Gene acquisitions from bacteria at the origins of major archaeal clades are vastly overestimated. Mol Biol Evol 33:305–310
Hahn MW, Wray GA (2002) The g-value paradox. Evol Dev 4:73–75
Hangauer MJ, Vaughn IW, McManus MT (2013) Pervasive transcription of the human genome produces thousands of previously unidentified long Intergenic noncoding RNAs. PLoS Genet 9:e1003569
Harrison PW, Lower RPJ, Kim NKD, Young JPW (2010) Introducing the bacterial 'chromid': not a chromosome, not a plasmid. Trends Microbiol 18:141–148
Hausner G (2003) Fungal mitochondrial genomes, introns and plasmids. In: Arora D, Khachatourians G (eds) Applied mycology and biotechnology. Elsevier, New York, pp 101–131
Henn BM, Cavalli-Sforza LL, Feldman MW (2012) The great human expansion. Proc Natl Acad Sci U S A 109:17758–17764
Hinnebusch J, Tilly K (1993) Linear plasmids and chromosomes in bacteria. Mol Microbiol 10:917–922
Horvath P, Barrangou R (2010) CRISPR/Cas, the immune system of bacteria and archaea. Science 327:167–170
Huerta-Sanchez E, Jin X, Asan BZ, Peter BM, Vinckenbosch N, Liang Y, Yi X, He M, Somel M, Ni P, Wang B, Ou X, Huasang LJ, Cuo ZXP, Li K, Gao G, Yin Y, Wang W, Zhang X, Xu X, Yang H, Li Y, Wang J, Wang J, Nielsen R (2014) Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA. Nature 512:194–197
Husnik F, Nikoh N, Koga R, Ross L, Duncan Rebecca P, Fujie M, Tanaka M, Satoh N, Bachtrog D, Wilson Alex CC, von Dohlen CD, Fukatsu T, McCutcheon John P (2013) Horizontal gene transfer from diverse bacteria to an insect genome enables a tripartite nested mealybug symbiosis. Cell 153:1567–1578
International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. Nature 409:860–921
International Human Genome Sequencing Consortium (2004) Finishing the euchromatic sequence of the human genome. Nature 431:931–945
Irimia M, Roy SW (2014) Origin of spliceosomal introns and alternative splicing. Cold Spring Harb Perspect Biol 6
Iwasaki YW, Siomi MC, Siomi H (2015) PIWI-interacting RNA: its biogenesis and functions. Annu Rev Biochem 84:405–433
Kapranov P, Cheng J, Dike S, Nix DA, Duttagupta R, Willingham AT, Stadler PF, Hertel J, Hackermüller J, Hofacker IL, Bell I, Cheung E, Drenkow J, Dumais E, Patel S, Helt G, Ganesh M, Ghosh S, Piccolboni A, Sementchenko V, Tammana H, Gingeras TR (2007) RNA maps reveal new RNA classes and a possible function for pervasive transcription. Science 316:1484–1488
Kapusta A, Kronenberg Z, Lynch VJ, Zhuo X, Ramsay L, Bourque G, Yandell M, Feschotte C (2013) Transposable elements are major contributors to the origin, diversification, and regulation of vertebrate long noncoding RNAs. PLoS Genet 9:e1003470
Katz LA, Grant JR (2014) Taxon-rich phylogenomic analyses resolve the eukaryotic tree of life and reveal the power of subsampling by sites. Syst Biol 64:406–415
Kazazian HH (2004) Mobile elements: drivers of genome evolution. Science 303:1626–1632
Kellis M, Wold B, Snyder MP, Bernstein BE, Kundaje A, Marinov GK, Ward LD, Birney E, Crawford GE, Dekker J, Dunham I, Elnitski LL, Farnham PJ, Feingold EA, Gerstein M, Giddings MC, Gilbert DM, Gingeras TR, Green ED, Guigo R, Hubbard T, Kent J, Lieb JD, Myers RM, Pazin MJ, Ren B, Stamatoyannopoulos JA, Weng Z, White KP, Hardison RC (2014) Defining functional DNA elements in the human genome. Proc Natl Acad Sci U S A 111:6131–6138
Kiss T (2002) Small nucleolar RNAs: an abundant group of noncoding RNAs with diverse cellular functions. Cell 109:145–148
Kozomara A, Griffiths-Jones S (2011) miRBase: integrating microRNA annotation and deep-sequencing data. Nucleic Acids Res 39:D152–D157

Kramerov DA, Vassetzky NS (2011) Origin and evolution of SINEs in eukaryotic genomes. Heredity 107:487–495
Krause J, Fu Q, Good JM, Viola B, Shunkov MV, Derevianko AP, Paabo S (2010) The complete mitochondrial DNA genome of an unknown hominin from southern Siberia. Nature 464:894–897
Kumar S, Schiffer P, Blaxter M (2012) 959 nematode genomes: a semantic wiki for coordinating sequencing projects. Nucleic Acids Res 40:D1295–D1300
Lambowitz AM, Zimmerly S (2011) Group II introns: mobile ribozymes that invade DNA. Cold Spring Harb Perspect Biol 3
Lane N, Martin WF (2012) The origin of membrane bioenergetics. Cell 151:1406–1416
Lang JM, Darling AE, Eisen JA (2013) Phylogeny of bacterial and archaeal genomes using conserved genes: supertrees and supermatrices. PLoS One 8:e62510
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, Worley KC, Kovar CL, Muzny DM, Gibbs RA, Warren WC, Mardis ER, Weinstock GM, Wilson RK, 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 478:476–482
Luan DD, Korman MH, Jakubczak JL, Eickbush TH (1993) Reverse transcription of R2bm RNA is primed by a nick at the chromosomal target site: a mechanism for non-LTR retrotransposition. Cell 72(4):595–605
Lynch M (2006) Streamlining and simplification of microbial genome architecture. Annu Rev Microbiol 60:327–349
Lynch M (2007) The origins of genome architecture. Sinauer Assoc, Sunderland
Lynch M, Conery JS (2003) The origins of genome complexity. Science 302:1401–1404
Magiorkinis G, Gifford RJ, Katzourakis A, De Ranter J, Belshaw R (2012) Env-less endogenous retroviruses are genomic superspreaders. Proc Natl Acad Sci U S A 109:7385–7390
Maher B (2008) Personal genomes: the case of the missing heritability. Nature 456:18–21
Margulis L (1970) Origin of eukaryotic cells. Yale University Press, New Haven
Marjoram P, Zubair A, Nuzhdin SV (2014) Post-GWAS: where next? More samples, more SNPs or more biology? Heredity 112:79–88
Martin W, Koonin EV (2006) A positive definition of prokaryotes. Nature 442:868–868
Martin W, Muller M (1998) The hydrogen hypothesis for the first eukaryote. Nature 392:37–41
Mattick JS, Makunin IV (2006) Non-coding RNA. Hum Mol Genet 15:R17–R29
McCutcheon JP, Moran NA (2012) Extreme genome reduction in symbiotic bacteria. Nat Rev Microbiol 10:13–26
McInerney JO, O'Connell MJ, Pisani D (2014) The hybrid nature of the Eukaryota and a consilient view of life on earth. Nat Rev Microbiol 12:449–455
McKinney BA, Pajewski NM (2012) Six degrees of epistasis: statistical network models for GWAS. Front Genet 2:109
Meader S, Ponting CP, Lunter G (2010) Massive turnover of functional sequence in human and other mammalian genomes. Genome Res 20:1335–1343
Medini D, Donati C, Tettelin H, Masignani V, Rappuoli R (2005) The microbial pan-genome. Curr Opin Genet Dev 15:589–594
Meyer M, Arsuaga J-L, de Filippo C, Nagel S, Aximu-Petri A, Nickel B, Martínez I, Gracia A, de Castro JMB, Carbonell E, Viola B, Kelso J, Prüfer K, Pääbo S (2016) Nuclear DNA sequences from the middle Pleistocene Sima de los Huesos hominins. Nature 531:504–507
Meyer M, Fu Q, Aximu-Petri A, Glocke I, Nickel B, Arsuaga J-L, Martinez I, Gracia A, de Castro JMB, Carbonell E, Paabo S (2014) A mitochondrial genome sequence of a hominin from Sima de los Huesos. Nature 505:403–406
Meyer M, Kircher M, Gansauge M-T, Li H, Racimo F, Mallick S, Schraiber JG, Jay F, Prüfer K, de Filippo C, Sudmant PH, Alkan C, Fu Q, Do R, Rohland N, Tandon A, Siebauer M, Green RE, Bryc K, Briggs AW, Stenzel U, Dabney J, Shendure J, Kitzman J, Hammer MF, Shunkov MV, Derevianko AP, Patterson N, Andrés AM, Eichler EE, Slatkin M, Reich D, Kelso J, Pääbo S (2012) A high-coverage genome sequence from an archaic Denisovan individual. Science 338:222–226
Moorjani P, Amorim CEG, Arndt PF, Przeworski M (2016) Variation in the molecular clock of primates. Proc Natl Acad Sci U S A 113:10607–10612

Naidoo N, Pawitan Y, Soong R, Cooper D, Ku C-S (2011) Human genetics and genomics a decade after the release of the draft sequence of the human genome. Hum Genomics 5:577–622

Nelson-Sathi S, Sousa FL, Roettger M, Lozada-Chavez N, Thiergart T, Janssen A, Bryant D, Landan G, Schonheit P, Siebers B, McInerney JO, Martin WF (2015) Origins of major archaeal clades correspond to gene acquisitions from bacteria. Nature 517:77–80

Ohno S (1972) So much «junk» DNA in our genome. In: Smith H (ed) Evolution of genetic systems. Gordon and Breach, New York, pp 366–370

Orgel LE, Crick FHC (1980) Selfish DNA: the ultimate parasite. Nature 284:604–607

Pääbo S (2014) The human condition – a molecular approach. Cell 157:216–226

Palazzo AF, Gregory TR (2014) The case for junk DNA. PLoS Genet 10:e1004351

Pang A, MacDonald J, Pinto D, Wei J, Rafiq M, Conrad D, Park H, Hurles M, Lee C, Venter JC, Kirkness E, Levy S, Feuk L, Scherer S (2010) Towards a comprehensive structural variation map of an individual human genome. Genome Biol 11:R52

Pellicer J, Fay MF, Leitch IJ (2010) The largest eukaryotic genome of them all? Bot J Linn Soc 164:10–15

Pennisi E (2007) Human genetic variation. Science 318:1842–1843

Pisani D, Cotton JA, McInerney JO (2007) Supertrees disentangle the chimerical origin of eukaryotic genomes. Mol Biol Evol 24:1752–1760

Prüfer K, Racimo F, Patterson N, Jay F, Sankararaman S, Sawyer S, Heinze A, Renaud G, Sudmant PH, de Filippo C, Li H, Mallick S, Dannemann M, Fu Q, Kircher M, Kuhlwilm M, Lachmann M, Meyer M, Ongyerth M, Siebauer M, Theunert C, Tandon A, Moorjani P, Pickrell J, Mullikin JC, Vohr SH, Green RE, Hellmann I, Johnson PLF, Blanche H, Cann H, Kitzman JO, Shendure J, Eichler EE, Lein ES, Bakken TE, Golovanova LV, Doronichev VB, Shunkov MV, Derevianko AP, Viola B, Slatkin M, Reich D, Kelso J, Paabo S (2014) The complete genome sequence of a Neanderthal from the Altai Mountains. Nature 505:43–49

Quinn JJ, Ilik IA, Qu K, Georgiev P, Chu C, Akhtar A, Chang HY (2014) Revealing long noncoding RNA architecture and functions using domain-specific chromatin isolation by RNA purification. Nat Biotechnol 32:933–940

Ramesh MA, Malik S-B, Logsdon JM Jr (2005) A phylogenomic inventory of meiotic genes: evidence for sex in *Giardia* and an early eukaryotic origin of meiosis. Curr Biol 15:185–191

Rasmussen B, Fletcher IR, Brocks JJ, Kilburn MR (2008) Reassessing the first appearance of eukaryotes and cyanobacteria. Nature 455:1101–1104

Reich D, Green RE, Kircher M, Krause J, Patterson N, Durand EY, Viola B, Briggs AW, Stenzel U, Johnson PLF, Maricic T, Good JM, Marques-Bonet T, Alkan C, Fu Q, Mallick S, Li H, Meyer M, Eichler EE, Stoneking M, Richards M, Talamo S, Shunkov MV, Derevianko AP, Hublin J-J, Kelso J, Slatkin M, Paabo S (2010) Genetic history of an archaic hominin group from Denisova cave in Siberia. Nature 468:1053–1060

Reichow SL, Hamma T, Ferré-D'Amaré AR, Varani G (2007) The structure and function of small nucleolar ribonucleoproteins. Nucleic Acids Res 35:1452–1464

Robinson GE, Hackett KJ, Purcell-Miramontes M, Brown SJ, Evans JD, Goldsmith MR, Lawson D, Okamuro J, Robertson HM, Schneider DJ (2011) Creating a buzz about insect genomes. Science 331:1386

Rochette NC, Brochier-Armanet C, Gouy M (2014) Phylogenomic test of the hypotheses for the evolutionary origin of eukaryotes. Mol Biol Evol 31:832–845

Roy SW, Gilbert W (2006) The evolution of spliceosomal introns: patterns, puzzles and progress. Nat Rev Genet 7:211–221

Sanyal A, Lajoie BR, Jain G, Dekker J (2012) The long-range interaction landscape of gene promoters. Nature 489:109–113

Sempere LF, Cole CN, McPeek MA, Peterson KJ (2006) The phylogenetic distribution of metazoan microRNAs: insights into evolutionary complexity and constraint. J Exp Zool B Mol Del Evol 306B:575–588

Shapiro B, Hofreiter M (2014) A paleogenomic perspective on evolution and gene function: new insights from ancient DNA. Science 343

Shendure J (2014) Life after genetics. Genome Med 6:86

Slatkin M (2008) Linkage disequilibrium – understanding the evolutionary past and mapping the medical future. Nat Rev Genet 9:477–485

Spang A, Saw JH, Jorgensen SL, Zaremba-Niedzwiedzka K, Martijn J, Lind AE, van Eijk R, Schleper C, Guy L, Ettema TJG (2015) Complex archaea that bridge the gap between prokaryotes and eukaryotes. Nature 521:173–179

Stanier RY, van Niel CB (1962) The concept of a bacterium. Arch Mikrobiol 42:17–35

Suh A, Witt CC, Menger J, Sadanandan KR, Podsiadlowski L, Gerth M, Weigert A, McGuire JA, Mudge J, Edwards SV, Rheindt FE (2016) Ancient horizontal transfers of retrotransposons between birds and ancestors of human pathogenic nematodes. Nat Commun 7:11396

Taft RJ, Pheasant M, Mattick JS (2007) The relationship between non-protein-coding DNA and eukaryotic complexity. BioEssays 29:288–299

The 1000 Genomes Project Consortium (2015) A global reference for human genetic variation. Nature 526:68–74

The International HapMap Consortium (2003) The international HapMap project. Nature 426:789–796

The International HapMap Consortium (2005) A haplotype map of the human genome. Nature 437:1299–1320

Thomas CA (1971) The genetic Organization of Chromosomes. Annu Rev Genet 5:237–256

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, Gocayne JD, Amanatides P, Ballew RM, Huson DH, Wortman JR, Zhang Q, Kodira CD, Zheng XH, Chen L, Skupski M, Subramanian G, Thomas PD, Zhang J, Gabor Miklos GL, Nelson C, Broder S, Clark AG, Nadeau J, McKusick VA, Zinder N, Levine AJ, Roberts RJ, Simon M, Slayman C, Hunkapiller M, Bolanos R, Delcher A, Dew I, Fasulo D, Flanigan M, Florea L, Halpern A, Hannenhalli S, Kravitz S, Levy S, Mobarry C, Reinert K, Remington K, Abu-Threideh J, Beasley E, Biddick K, Bonazzi V, Brandon R, Cargill M, Chandramouliswaran I, Charlab R, Chaturvedi K, Deng Z, Francesco VD, Dunn P, Eilbeck K, Evangelista C, Gabrielian AE, Gan W, Ge W, Gong F, Gu Z, Guan P, Heiman TJ, Higgins ME, Ji R-R, Ke Z, Ketchum KA, Lai Z, Lei Y, Li Z, Li J, Liang Y, Lin X, Lu F, Merkulov GV, Milshina N, Moore HM, Naik AK, Narayan VA, Neelam B, Nusskern D, Rusch DB, Salzberg S, Shao W, Shue B, Sun J, Wang ZY, Wang A, Wang X, Wang J, Wei M-H, Wides R, Xiao C, Yan C, Yao A, Ye J, Zhan M, Zhang W, Zhang H, Zhao Q, Zheng L, Zhong F, Zhong W, Zhu SC, Zhao S, Gilbert D, Baumhueter S, Spier G, Carter C, Cravchik A, Woodage T, Ali F, An H, Awe A, Baldwin D, Baden H, Barnstead M, Barrow I, Beeson K, Busam D, Carver A, Center A, Cheng ML, Curry L, Danaher S, Davenport L, Desilets R, Dietz S, Dodson K, Doup L, Ferriera S, Garg N, Gluecksmann A, Hart B, Haynes J, Haynes C, Heiner C, Hladun S, Hostin D, Houck J, Howland T, Ibegwam C, Johnson J, Kalush F, Kline L, Koduru S, Love A, Mann F, May D, McCawley S, McIntosh T, McMullen I, Moy M, Moy L, Murphy B, Nelson K, Pfannkoch C, Pratts E, Puri V, Qureshi H, Reardon M, Rodriguez R, Rogers Y-H, Romblad D, Ruhfel B, Scott R, Sitter C, Smallwood M, Stewart E, Strong R, Suh E, Thomas R, Tint NN, Tse S, Vech C, Wang G, Wetter J, Williams S, Williams M, Windsor S, Winn-Deen E, Wolfe K, Zaveri J, Zaveri K, Abril JF, Guigó R, Campbell MJ, Sjolander KV, Karlak B, Kejariwal A, Mi H, Lazareva B, Hatton T, Narechania A, Diemer K, Muruganujan A, Guo N, Sato S, Bafna V, Istrail S, Lippert R, Schwartz R, Walenz B, Yooseph S, Allen D, Basu A, Baxendale J, Blick L, Caminha M, Carnes-Stine J, Caulk P, Chiang Y-H, Coyne M, Dahlke C, Mays AD, Dombroski M, Donnelly M, Ely D, Esparham S, Fosler C, Gire H, Glanowski S, Glasser K, Glodek A, Gorokhov M, Graham K, Gropman B, Harris M, Heil J, Henderson S, Hoover J, Jennings D, Jordan C, Jordan J, Kasha J, Kagan L, Kraft C, Levitsky A, Lewis M, Liu X, Lopez J, Ma D, Majoros W, McDaniel J, Murphy S, Newman M, Nguyen T, Nguyen N, Nodell M, Pan S, Peck J, Peterson M, Rowe W, Sanders R, Scott J, Simpson M, Smith T, Sprague A, Stockwell T, Turner R, Venter E, Wang M, Wen M, Wu D, Wu M, Xia A, Zandieh A, Zhu X (2001) The sequence of the human genome. Science 291:1304–1351

Vernot B, Akey JM (2014) Resurrecting surviving Neandertal lineages from modern human genomes. Science 343:1017–1021

Visscher PM, Brown MA, McCarthy MI, Yang J (2012) Five years of GWAS discovery. Am J Hum Genet 90:7–24

Wacey D, Kilburn MR, Saunders M, Cliff J, Brasier MD (2011) Microfossils of Sulphur-metabolizing cells in 3.4-billion-year-old rocks of Western Australia. Nat Geosci 4:698–702

Wallin I (1927) Symbioticism and the origin of species. Williams & Wilkins Company, Baltimore

Wang Z, Wu M (2014) Phylogenomic reconstruction indicates mitochondrial ancestor was an energy parasite. PLoS One 9:e110685

Watson J (1990) The human genome project: past, present, and future. Science 248:44–49

White MF, Bell SD (2002) Holding it together: chromatin in the archaea. Trends Genet 18:621–626

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

Williams TA, Foster PG, Cox CJ, Embley TM (2013) An archaeal origin of eukaryotes supports only two primary domains of life. Nature 504:231–236

Woese CR, Fox GE (1977) Phylogenetic structure of the prokaryotic domain: the primary kingdoms. Proc Natl Acad Sci U S A 74:5088–5090

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 U S A 87:4576–4579

Wu D, Hugenholtz P, Mavromatis K, Pukall R, Dalin E, Ivanova NN, Kunin V, Goodwin L, Wu M, Tindall BJ, Hooper SD, Pati A, Lykidis A, Spring S, Anderson IJ, D'haeseleer P, Zemla A, Singer M, Lapidus A, Nolan M, Copeland A, Han C, Chen F, Cheng J-F, Lucas S, Kerfeld C, Lang E, Gronow S, Chain P, Bruce D, Rubin EM, Kyrpides NC, Klenk H-P, Eisen JA (2009) A phylogeny-driven genomic encyclopaedia of bacteria and archaea. Nature 462:1056–1060

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 TJG (2017) Asgard archaea illuminate the origin of eukaryotic cellular complexity. Nature 541:353–358

Zarrei M, MacDonald JR, Merico D, Scherer SW (2015) A copy number variation map of the human genome. Nat Rev Genet 16:172–183

Zhang J (2000) Protein-length distributions for the three domains of life. Trends Genet 16:107–109

Organelle Genomes and Endosymbionts

C. Bleidorn, *Phylogenomics*, DOI 10.1007/978-3-319-54064-1_2

- Mitochondria and chloroplasts are cell organelles of endosymbiotic origin which carry their own genomes.
- Chloroplasts were acquired by their hosts either by primary endosymbiosis (uptake of a cyanobacterium by the last common ancestor of Archaeplastida) or secondary endosymbiotic events (symbiosis of a eukaryote host with chloroplast-bearing algae).
- Organelle genomes differ in size, structure and gene content across taxa.
- Mitochondrial and chloroplast markers are used as standard for DNA barcoding of animals and plants.
- Heritable bacterial endosymbionts are broadly classified as either primary symbionts or secondary symbiont and are commonly found in eukaryotes, especially insects.
- Endosymbiont genomes are highly streamlined and included the smallest reported genomes of all living organisms.

2.1 Mitochondria

2.1.1 Origin and Evolution of Mitochondria

Mitochondria are double-membrane-bound cell organelles which contain their own genome and carry out the replication, transcription and translation of DNA. With the publication of the outstanding book *Origin of Eukaryotic Cells* by Lynn Margulis (1970), the old idea that mitochondria evolved from free-living bacteria via symbiosis got broad attention in the scientific community. The development of cloning and sequencing techniques in the 1970s allowed sequencing of mitochondrial genes and basically confirmed this hypothesis (Gray 2012). Using phylogenomic analyses of different sets of orthologous genes, mitochondria are now firmly placed within Alphaproteobacteria as part of the Rickettsiales (Wang and Wu 2015). Interestingly, this taxon comprises a large variety of bacterial endosymbionts which, similar to mitochondria, also harbour strongly reduced genomes (e.g. the genera *Rickettsia* and *Wolbachia*; ▶ see Sect. 2.3). Whereas the phylogenetic placement of mitochondria seems well settled, the circumstances under which this symbiosis evolved remain under debate (Gray and Archibald 2012). Several eukaryote taxa lack mitochondria, including Microsporidia (fungi), Trichomonadida (Excavata, Fornicata), Diplomonadida (Excavata, Fornicata) and Archamoebae (Amoebozoa) (Cavalier-Smith 1987; Keeling 1998). It has been suggested that these taxa primarily lack mitochondria uniting them as an early branching clade of eukaryotes called Archezoa (Cavalier-Smith 1983). This would mean that the acquisition of mitochondria took place during eukaryote evolution. Interestingly, molecular phylogenetic analyses including the first available ribosomal sequence data initially supported the early branching of amitochondriate taxa, thereby supporting the Archezoa hypothesis (Sogin 1989). However, this idea was later rejected based on several findings. First, all recent amitochondriate taxa have been demonstrated to bear double-membrane-bounded organelles which are referred to as mitosomes or hydrogenosomes (Hjort et al. 2010) and interpreted to be derived from mitochondria. Hydrogenosomes synthesize ATP under anaerobic condition and thereby produce hydrogen, whereas in the case of mitosomes, the function remains unclear. Hydrogenosomes are also found in taxa which never have

been speculated to be primarily amitochondriate, e.g. in some Loricifera (Metazoa, Ecdysozoa) (Danovaro et al. 2010). It is obvious that these types of reduced organelles evolved several times convergently across eukaryotes. With the oxymonad *Monocercomonoides* sp., only a single case of a eukaryotic taxon lacking any form of mitochondrion has been described (Karnkowska et al. 2016). However, also in this case, a secondary loss of mitochondria is clearly supported by a phylogenetic analysis. Second, phylogenetic analyses using other or more genes clearly proofed that the basal branching placement of amitochondriate taxa might be due to systematic errors. Instead, these analyses firmly placed former archezoan taxa as derived eukaryotes. *Microsporidia* are supported as part of the fungi and archaeamobans group deeply within Amoebozoa, and Diplomonadida and Triplomonadida are part of the Excavata (Embley and Martin 2006; Katz and Grant 2014). Finally, PCR-based and genomic analyses found genes which have been transferred from the mitochondrium to the nucleus in amitochondriate taxa (Embley and Martin 2006). Whereas the Archezoa hypothesis has been firmly put to rest and no recent primarily amitochondriate eukaryotes are known, it is still discussed in which order the events leading to the eukaryotic cell have evolved (Fig. 2.1). In a mitochondrial-early scenario, it is assumed that the acquisition of mitochondria basically defines the evolution of eukaryotes, suggesting that the last eukaryote common ancestor already had a mitochondrium. In contrast, mitochondrial-late scenarios assume that early eukaryotes already show some cellular complexity, thereby distinguishing them from archaeans, and that mitochondria were acquired via endosymbiosis later in evolution (Ettema 2016). Genes of eukaryotes are of different ancestry, and the origin of many

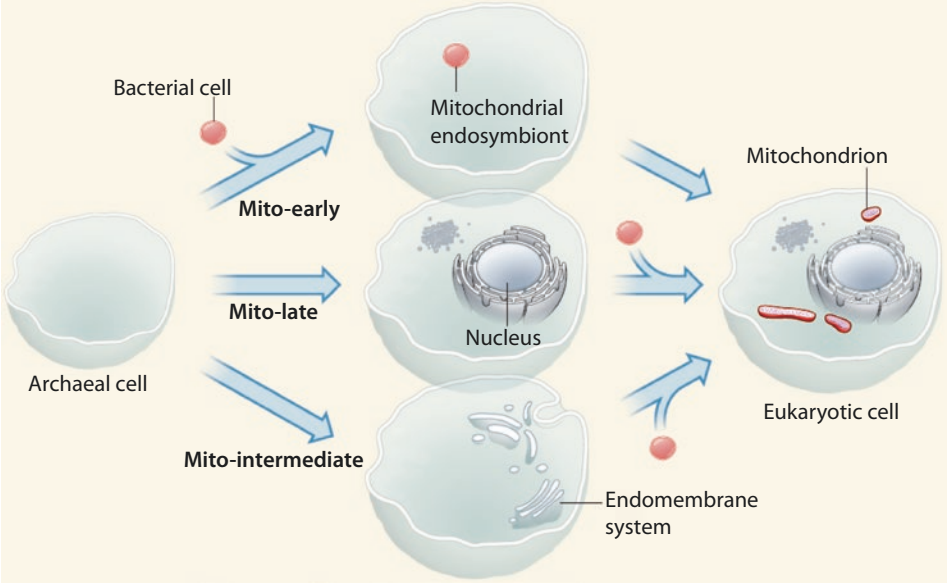


Fig. 2.1 Origin of eukaryotic cells and their mitochondria. Three different scenarios for the acquisition of mitochondria exist. In the mito-early scenario, the acquisition of the mitochondrium via endosymbiosis already took place in the last eukaryote common ancestor. Mito-late and mito-intermediate scenarios assume already complexly organized eukaryotic ancestors which later in evolution acquired the mitochondrion via endosymbiosis (Reprinted by permission from Macmillan Publishers Ltd: Nature (Ettema 2016), Copyright 2016)

can be traced back to archaean or bacterial clades. As mitochondria are likely of alphaproteobacterial origin, genes related to them, either located in the organelle or transferred to the nucleus, can be traced back to these bacteria in a phylogenetic analysis (McInerney et al. 2014). By using comparative studies of the mitochondrial proteome, a conserved core of proteins descended from the ancestral mitochondrion has been identified (Gray 2015). Similarly, core eukaryote nuclear genes of different functional classes can be identified, whose origin also can be traced back in a phylogenomic analysis, often favouring an archaeal origin. With the help of phylogenetic gene family analyses, the relative age of a given group of genes can be estimated. In case of the mitochondria-early hypothesis, it has to be assumed that there are no differences in the age of genes of archaeal and alphaproteobacterial origin. However, a study testing this hypothesis clearly found support that mitochondria-related genes of an alphaproteobacterial origin are significantly younger than eukaryotic genes of other origin (Pittis and Gabaldón 2016). This would support a mitochondria-late scenario, where an already cellularly complex organized eukaryotic host would have acquired the mitochondrium. Whereas most of the genes without mitochondrial origin can be traced back to an archaeal origin, several other genes are of bacterial origin from different clades, underlining the chimeric nature of early eukaryotes. It remains difficult to distinguish if the acquisition of these diverse sets of genes stems from several events of horizontal gene transfer or maybe previous endosymbiotic associations with other bacteria (Pittis and Gabaldón 2016).

Several adaptive hypotheses exist to explain the ancestral function of mitochondria (Lynch 2007). The primary function of recent mitochondrial organelles is the acquisition of energy, and this might also reflect their ancestral role. Other hypotheses describe ancestral functions like oxygen scavenging, photosynthate acquisition or hydrogen acquisition. Under the latter hypothesis, mitochondria are postulated to originate in a hydrogen-dependent autotrophic archaeal host that lived in a fully anaerobic environment. In this relationship the ancestral role of mitochondria was to provide the host with hydrogen produced by fermentation of organic substrates (Martin and Muller 1998).

Given that mitochondria are coevolving with their hosts for more than 1.5 billion years, it comes without surprise that recent mitochondria differ strikingly across taxa. Notably, mitochondrial genomes of extant eukaryotes differ strongly in size, structure and gene content (Burger et al. 2003b; Nosek and Tomáška 2003). Usually the mitochondrial genome is organized as a single circular molecule, as typical for most prokaryote genomes. However, many deviations from this circular organization have been described. Linearly organized mitochondrial genomes are not rare, as, for example, found in ciliates or medusozoan cnidarians (Kayal et al. 2012; Burger et al. 2000). Several different solutions of maintaining the telomeres of linearly organized mitochondrial DNA have been reported, including hairpin structures, inverted and non-inverted repeat sequences or terminal proteins (Nosek and Tomáška 2003). Mitochondrial genomes are not always encoded on a single molecule, but can also be organized on two or more circular (e.g. in several insects and nematode species or in the flowering plant *Amborella trichopoda*) or linear molecules (e.g. the cnidarian *Hydra magnipapillata* with two linear fragments or the opisthokont *Amoebidium parasiticum* bearing many linear fragments) (Burger et al. 2003a; Gibson et al. 2007; Cameron et al. 2011; Voigt et al. 2008; Rice et al. 2013). Mitochondrial genomes show major differences in the size, with the smallest genomes in a range of around 6–7 Kb (e.g. in several apicomplexans) to the biggest known genomes in the size of several Mb

(e.g. 11.3 Mb in the flowering plant *Silene conica*), thereby exceeding the size of some nuclear eukaryote genomes (Hikosaka et al. 2010; Sloan et al. 2012).

The gene content of mitochondrial genomes is greatly reduced compared with Alphaproteobacteria. Large-scale phylogenetic comparisons differ in the number of proteins from 394 to 842 that were (minimally) likely part of the ancestral mitochondrial proteome (Gabaldón and Huynen 2007; Wang and Wu 2014). In contrast, proteins encoded in recent mitochondria range from 3 (as in the apicomplexan *Plasmodium*) to 66 (jakobid Excavata) (Gray 2015). Proteins encoded on mitochondrial genomes are usually involved in the respiratory chain or its corresponding translation system (Lithgow and Schneider 2010). Mitochondrial translation alone requires more than 100 proteins and many other essential housekeeping genes, most of which are encoded in the nucleus and are imported by mitochondria (Dolezal et al. 2006). Altogether, for some species, it is reported that up to 1000 genes are encoded in the nucleus, synthesized in the cytosol and imported to the mitochondria (Lithgow and Schneider 2010). The horizontal transfer of genes from the mitochondrium to the nucleus is further complicated by the presence of differences in the genetic code. The genetic code of mitochondria varies among organisms, and at least 16 deviations from the standard code are reported across eukaryotes, with animals showing the highest diversity (Knight et al. 2001).

2.1.2 Animal Mitochondrial Genomes

Animal mitochondrial genomes usually range in a size between 11 and 20 Kb (Gissi et al. 2008), even though some examples exist with genome sizes of up to 43 Kb as in Placozoa (Dellaporta et al. 2006). Typically, these densely packed genomes encode for 13 protein-coding genes, 22 tRNAs and 2 ribosomal RNAs (Fig. 2.2), which all can be located on either strand (Bernt et al. 2013). One protein-coding gene (*atp8*) has been lost convergently in (among others) many Platyhelminthes, Acoelomorpha and Nematoda. It is not unusual that genes begin with alternative starting codons and stop codons are often incomplete. Moreover, tRNAs are often truncated and undergo RNA editing (Börner et al. 1997) or are missing completely (Gissi et al. 2008). Only few examples of the existence of introns are reported, which are in most cases self-splicing group II introns (Huchon et al. 2015). Animal mitochondria are transmitted maternally, even though some examples of doubly uniparental inheritance (DUI) exist. Transmission via DUI is known for several bivalve molluscs and is characterized by the presence of two distinct gender-associated mitochondrial DNAs, where one is transmitted via eggs (F, female) and the other one transmitted through sperm (M, male) (Passamonti et al. 2011). In this case, females are homoplasmic as they only receive mitochondria from their mother, whereas males are heteroplasmic receiving the organelles from both parents. F and M mitochondrial genomes can differ up to 50% in their nucleotide sequence (Doucet-Beaupré et al. 2010).

Substitution rates of mitochondrial genes are several times faster than those of single-copy nuclear genes in most bilaterian animals (Brown et al. 1979). This made mitochondrial genes ideal markers for population genetic, phylogeographic, phylogenetic and barcoding studies (Avise 2004). However, in non-bilaterian animals like Cnidaria and Porifera, mitochondrial substitutions rates are much lower, making these genes less suitable for barcoding or population genetics (Shearer et al. 2002; Huang et al. 2008).

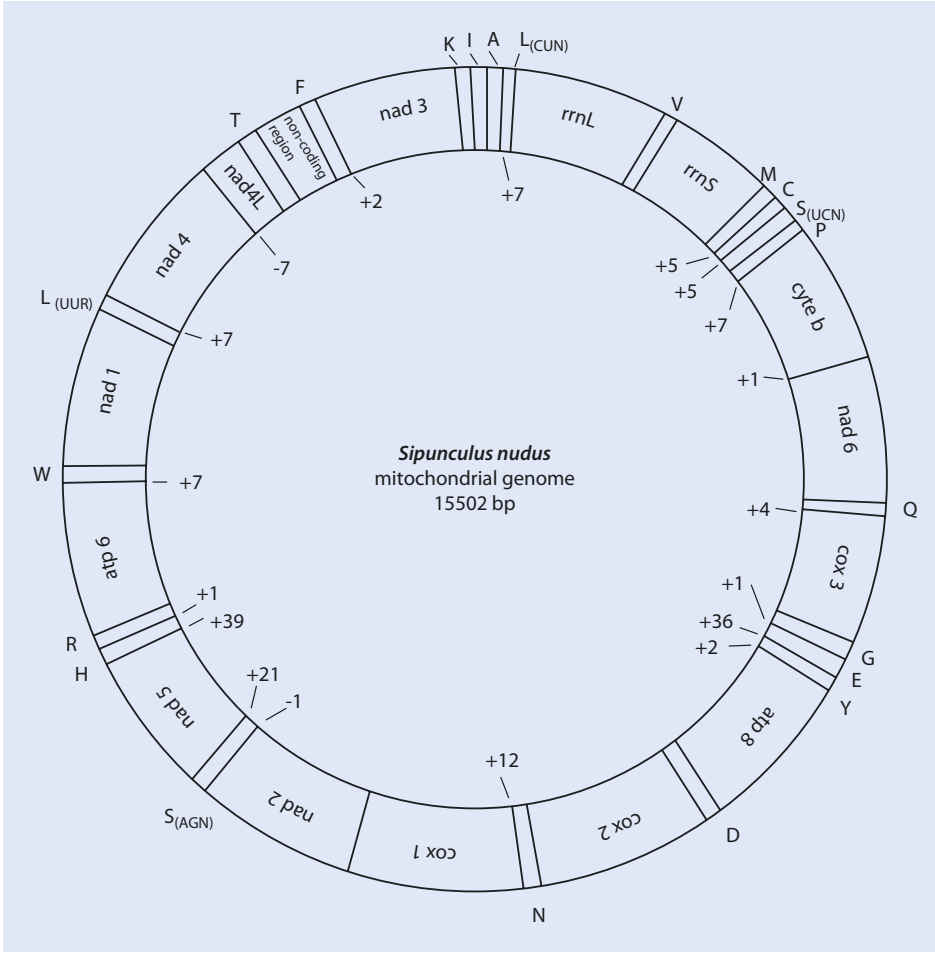


Fig. 2.2 The circular mitochondrial genome of the annelid *Sipunculus nudus* is typical for animal genomes. The densely packed genome encodes for 13 protein-coding genes (*atp1-atp8*, *cox1-cox3*, *cytb*, *nad1-nad6*), 2 ribosomal RNAs (small (*rrns*) and large (*rrnl*) subunit) and 22 tRNAs (specified in one-letter code) (Reprinted from Mwinyi et al. (2009))

2.1.3 Mitochondrial Genomes of Plants and Algae

Whereas animal mitochondrial genomes are rather uniformly organized, plant mitochondrial genomes exhibit a great diversity in size, structure and gene content (Mower et al. 2012; Liu et al. 2012). Mitochondrial genomes have been sequenced for all major clades of plants (rhodophytes, chlorophytes, charophytes, hornworts, liverworts, mosses, ferns, lycophytes, gymnosperms, angiosperms), showing a variety in genome size of a thousand-fold ranging from 13 Kb in chlorophytes up to 11.3 Mb in angiosperms (Mower et al. 2012; Sloan et al. 2012). Remarkable are the mitochondrial genomes of land plants (embryophytes) which are prone to recombination, RNA editing, trans-splicing, insertion of DNA from the chloroplast and nuclear genomes as well as from distant taxa and ongoing gene transfer into the nucleus (Knoop 2012). Land plant mitochondrial genomes are

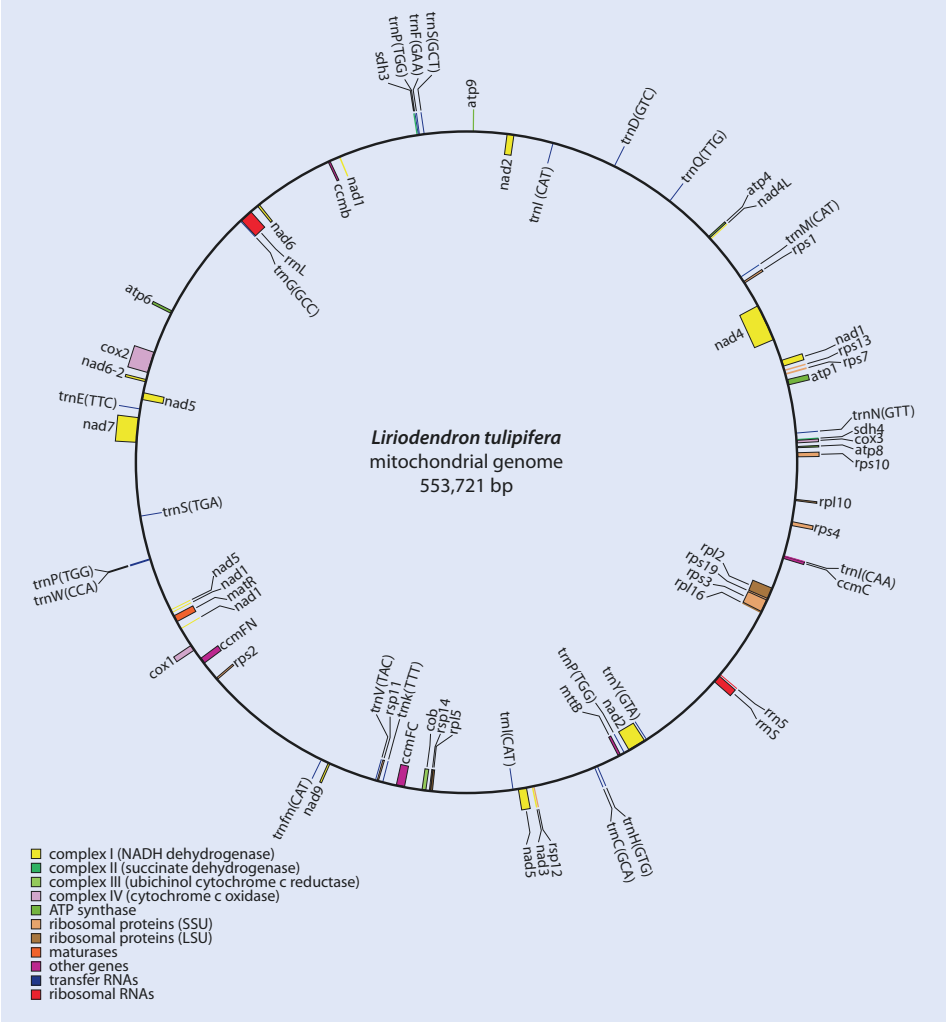


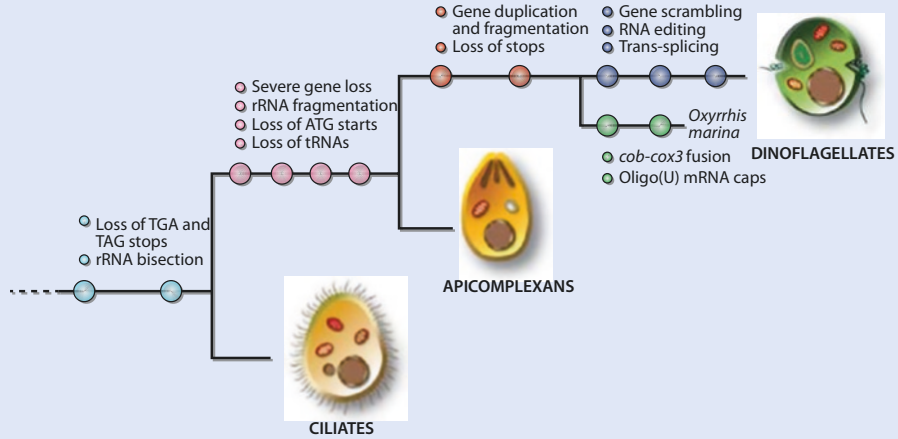
Fig. 2.3 The circular mitochondrial genome of the angiosperm land plant *Liriodendron tulipifera*. As typical for land plants, the genome is expanded in size, harbouring large intergenic regions. Genes inside and outside the circle are transcribed from different strands (The genome has been published by Richardson et al. (2013) and was redrawn using OGDRAW (Lohse et al. 2013))

highly variable in their size, content and structure. The genome size of most land plants exceeds 200 Kb, even though usually less than 20% are encoding for proteins or RNAs (Fig. 2.3). The remaining genome content is dominated by the presence of group I and II introns and large intergenic regions. The number of introns in land plants ranges from 19 to 37, and their position is relatively conserved within clades (Mower et al. 2012). Intergenic regions include repetitive elements, as well as integrations from the nucleus and chloroplast of their own and foreign genomes (Chaw et al. 2008; Rice et al. 2013). However, the origin of most of the excessive intergenic regions of the hugely expanded land plant genomes remains unclear (Sloan et al. 2012).

Plants have a much larger number of genes encoded in their mitochondrial genomes than animals. The number of identified genes in most plants ranges from 42 to 69; however, some chlorophyte green algae only have around 10 genes (Mower et al. 2012; Fan and Lee 2002). As such genes for ribosomal RNAs, tRNAs, ribosomal proteins, a twin-arginine translocase subunit (*tatC*) and *nad* (NADH dehydrogenase), *sdh* (succinate dehydrogenase), *cob* (cytochrome b), *cox* (cytochrome c oxidase), *ccm* (cytochrome c maturation) and *atp* (ATP synthase) subunits are found (Knoop 2012). Interestingly, even though plant mitochondrial genomes show high structural variability, the substitution rates of their encoded genes are rather low in most species (Christensen 2013), making them less suitable as molecular markers in population level or barcoding studies (▶ see Sect. 2.4).

2.1.4 Mitochondrial Genomes of «Other» Eukaryotes

Of outstanding interest from an evolutionary point of view are the mitochondrial genomes of jakobid flagellates, a group of unicellular eukaryotes which are part of the Excavata (Katz and Grant 2014). Jakobid mitochondrial genomes range in their size from 65 to 100 Kb, while showing also a compact organization with a high coding density ranging from 80% to 93% (Burger et al. 2013). These mitochondrial genomes are the most gene rich among eukaryotes, with nearly 100 genes. Unique among eukaryotes is the presence of genes involved in transcription and quality control of translation. Moreover, some highly conserved gene clusters are found which are interpreted as remnants of an operon structure inherited from the bacterial ancestor of mitochondria. In summary, the genome organization of jakobid mitochondria is suggested to most closely resemble the ancestral pattern of all eukaryotes (Lang et al. 1997). In contrast, the most reduced mitochondrial genomes are found among apicomplexans and dinoflagellates (◘ Fig. 2.4), which are sister



◘ **Fig. 2.4** Mitochondrial genome evolution in alveolates. Dinoflagellata and Apicomplexa represent sister taxa, whose current ancestor already possessed a strongly reduced mitochondrial genome, harbouring only three protein-coding genes and fragments of the ribosomal RNAs. Further modifications of the mitochondrial genomes evolved in these lineages (Reprinted by permission from John Wiley and Sons (Waller and Jackson 2009), Copyright 2009)

groups within the taxon Alveolata (SAR clade) (Katz and Grant 2014). The mitochondrial genome of the apicomplexan malaria parasite *Plasmodium falciparum* is encoded on tandemly repeated linear copies of 6 Kb which include only three protein-coding genes (*cox1*, *cox3*, *cob*) and fragments of the small and large subunit of the ribosomal RNA (Feagin et al. 1997). Similarly organized mitochondrial genomes have been revealed for other Apicomplexa, even though the number of ribosomal fragments, the order of genes and the number of copies of tandem repeats (monomeric vs. multiple copies) can vary (Hikosaka et al. 2013). The genome content of dinoflagellates is similar to Apicomplexa (Waller and Jackson 2009). However, some substantial modifications can be found in these genomes, including massive amplification and recombination of the genome. Moreover, trans-splicing is required for generating *cox3* transcripts, and RNA editing of most genes is ubiquitous (Jackson et al. 2012). Dinoflagellates can have surprisingly large genomes given the reduced genome content. The genome of *Symbiodinium minutum* is around 326 Kb, of which 99% are non-coding, even though transcribed (Shoguchi et al. 2015).

2.2 Plastids

2.2.1 Origin and Evolution of Plastids

It is well accepted that plastids originated later than mitochondria in eukaryote evolution, and some eukaryotes bear plastids whereas others not. The first plastids stem from the uptake of a cyanobacterium by the ancestor of Archaeplastida, a clade uniting glaucocystophytes, Rhodophyta and Viridiplantae (green algae and land plants) (Gray 1999). Oxygenic photosynthesis, the conversion of H_2O and CO_2 into energy-rich sugars and O_2, evolved in the lineage of *Cyanobacteria* more than 3.5 billion years ago, which enduringly transformed life on earth due to oxygen enrichment in the atmosphere (Gould et al. 2008; Hohmann-Marriott and Blankenship 2011). Eukaryotes were able to co-opt photosynthesis by integrating their cyanobacterial endosymbiont. Interestingly, whereas it seems that plastids are all closely related, the organisms that contain them are from diverse eukaryotic clades. Besides Archaeplastida, plastids are further found in euglenids, apicomplexans, haptophytes, cryptomonads, heterokonts and dinoflagellates (Keeling 2010). This can be explained by multiple layers of endosymbiotic events, so-called secondary endosymbiosis (Fig. 2.5). In this case a

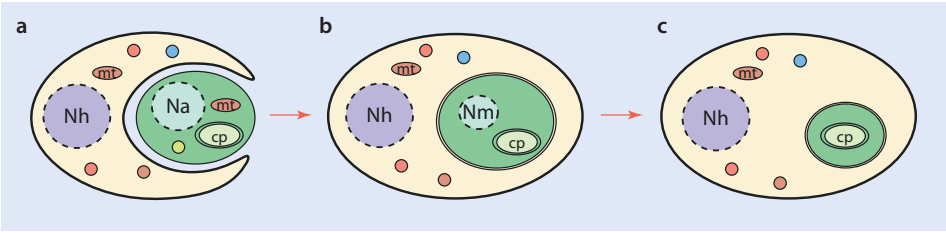


Fig. 2.5 Evolution of secondary endosymbiosis. **a** A *red* or *green* alga with a chloroplast surrounded by a double membrane is engulfed by a eukaryotic host. **b** Eukaryotic host carries a chloroplast with four membranes and the vestigial endosymbiont nucleus (nucleomorph). **c** Eukaryotic host with chloroplast surrounded by four membranes, and the algal nucleus has been completely reduced. *Abbreviations*: *cp* chloroplast, *mt* mitochondrium, *Na* nucleus of the alga, *Nh* nucleus of the host, *Nm* nucleomorph

eukaryote carrying a chloroplast has been phagocytized by another eukaryote leading to a subsequent integration of the new endosymbiont. Such events can be distinguished from primary endosymbiosis by the morphology of the plastids, as they still carry additional cell membranes stemming from the phagocytosis. Whereas primary endosymbionts bear two plastid membranes, secondary endosymbionts have four such membranes, which are sometimes reduced to three, as in euglenids and dinoflagellates (Keeling 2013). Moreover, a few cases of tertiary endosymbiosis events have been documented for some dinoflagellate lineages. Species of the genera *Karenia* and *Karlodinium* lost their plastids and gained new ones from a haptophyte species (Tengs et al. 2000). In the case of secondary endosymbiosis, the algal nucleus is usually completely reduced, while only the chloroplast remains (Fig. 2.5c). However, in some cases a vestigial nucleus of the algal symbiont is retained (Fig. 2.5b). These relicts are called nucleomorphs and can be found in cryptomonads and chlorarachniophytes (Keeling 2010). The actual number of events of secondary and tertiary symbiosis remains still debated, but at least eight distinct evolutionary events are suggested (Cavalier-Smith 2003; Keeling 2013, 2010).

The primary role of the plastid is to conduct photosynthesis, in which case they are called chloroplasts. However, some plastids seem to have lost their photosynthetic ability, e.g. in Apicomplexa (Köhler et al. 1997). As typical for obligate endosymbionts, many unnecessary genes got lost, whereas several other genes were transferred to the host nucleus. Nevertheless, plastids retain a small part of their ancestral genome, which might be due to the fact that hydrophobic proteins are difficult to transport to the organelle or that organelles are need to be in control of expression for genes which are part of the electron transport chain as a redox regulation (Allen 2015; Timmis et al. 2004). Additionally, some cases of transfer of nuclear genes into the plastid genomes are documented (Keeling 2009). Whereas primary plastids are located in the cytoplasm, secondary plastids are found within the endomembrane system. All genes necessary for plastid function which are encoded in the nucleus have a targeting system to arrive at the plastid and to cross its inner and outer envelopes (Strittmatter et al. 2010). Around 40% of the plastid proteome consists of proteins which seem to be derived from the host nuclear genome or various bacterial lineages outside *Cyanobacteria* (Suzuki and Miyagishima 2010).

Similar to mitochondria, plastid genomes are usually organized as circular molecules, with one genome per circle. Additionally, long, polyploid linear molecules and branched molecules undergoing replication seem to be also abundant (Bendich 2004). Plastid genomes of Archaeplastida are usually around 100–200 Kb in size, and the molecule shows a quadripartite structure due to the presence of two large inverted repeats, which divide the molecule into a large and a small single-copy region (Fig. 2.6). Usually around 60–250 genes are encoded on chloroplasts, and they are normally organized as operons. The inverted repeats include the ribosomal RNA genes (16S, 23S and 5S rRNA), as well as some other genes. The number of tRNAs varies between 27 and 31, and a variable number of ribosomal protein genes is usually present. Protein-coding genes are part of photosystems I and II, the cytochrome b6f complex as well as ATP synthase (Green 2011).

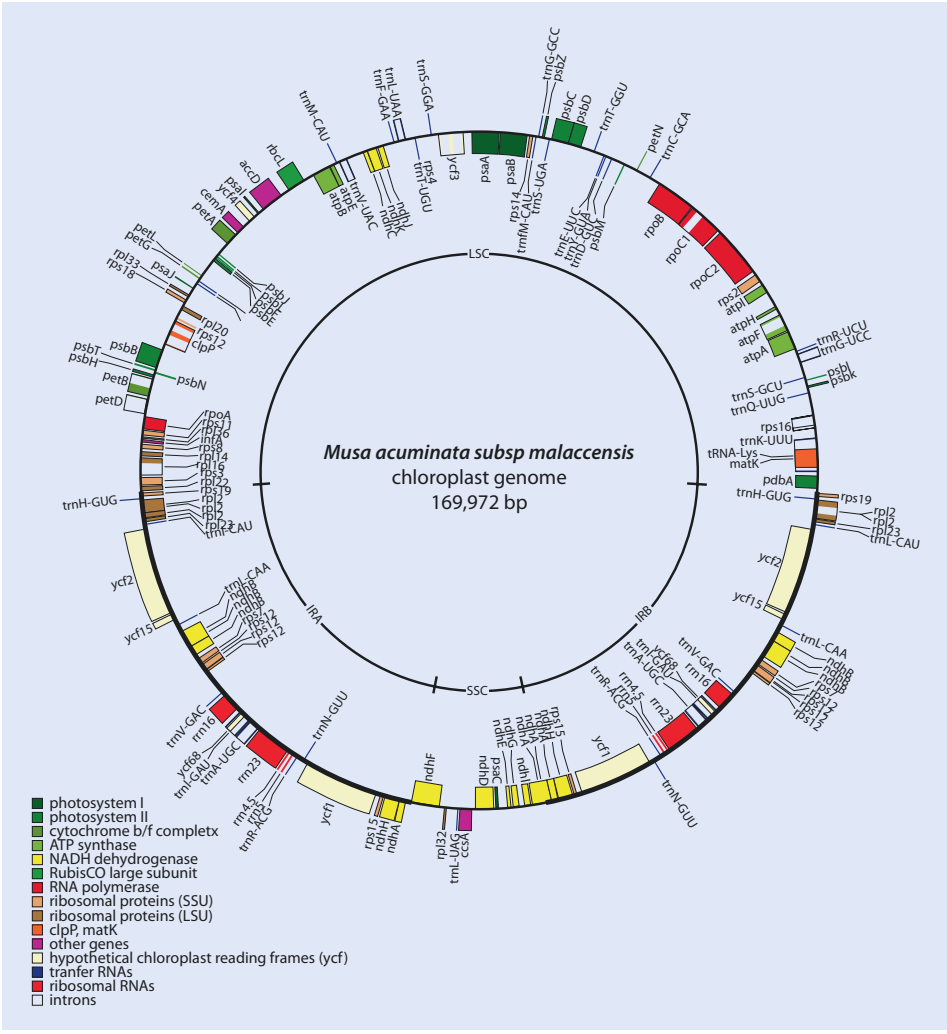


Fig. 2.6 Chloroplast genome of the wild Malaysian banana *Musa acuminata*. Different groups of genes indicated by *colours*, with genes inside and outside the *outer circle* transcribed from different strands. Borders of the large single-copy region (LSC) and small (SSC) single-copy region, as well of inverted repeat regions (IRA, IRB) indicated on the *inner circle* (Chloroplast genome published by Martin et al. (2013) and redrawn with OGDRAW (Lohse et al. 2013))

2.2.2 Plastid Genomes

The composition and size of plastids can vary dramatically across taxa. The largest chloroplast genomes are found in green algae, with the ~500 Kb genome of *Floydiella terrestris* as the actual record holder (Brouard et al. 2010). The large size in these taxa is mainly due

to repetitive regions and not due to the number of retained genes, which is with ~100 similar to that of most land plants. In contrast, the chloroplast genomes of red algae (Rhodophyta) bear with 220–250, the largest number of genes of any sequenced chloroplasts (Janouškovec et al. 2013). It seems that red algae show a much slower rate of plastid-to-nucleus gene transfer. Moreover, red algae plastid genomes also lack inverted repeats. The other extreme in terms of gene number comes from dinoflagellates. The chloroplasts of most photosynthetic dinoflagellates contain the light-harvesting pigment peridinin and are surrounded by three membranes, suggesting a secondary symbiosis, putatively stemming from a red algal host. The organization of these chloroplasts is highly unusual, and only a small number of genes (17) is retained, which are located on different minicircles. Moreover, other chloroplasts types stemming from different symbiosis events are also described in this taxon, which accordingly differ in size and organization (Dorrell and Howe 2015). Half of the described dinoflagellate lineages lost their plastids (Green 2011). Apicomplexa, the sister taxon of dinoflagellates, are parasitic eukaryotes (including important pathogens of humans or livestock as *Plasmodium*, *Toxoplasma*, *Eimeria*) which bear a name-giving organelle called apicoplast, which is also derived from a secondary uptake of a chloroplast of putatively red algal origin. However, unlike chloroplasts these organelles do not show photosynthetic activity. Nevertheless, apicoplasts retain their own genome and expression machinery and are involved in the synthesis of fatty acids, isoprenoids, iron sulphur clusters and haem (Lim and McFadden 2010).

Transcription in chloroplasts can be mediated by two different types of RNA polymerase or by a combination of both. According to the location where the corresponding polymerase is encoded, these types are abbreviated as PEP (plastid-encoded plastid RNA polymerase) or NEP (nuclear-encoded plastid RNA polymerase) (Yagi and Shiina 2014). Chloroplast genes are categorized into three classes according the promoter they bear for their transcription. Class I genes are photosynthesis-related genes and are mainly transcribed by PEP, class II genes comprise mainly of housekeeping genes transcribed by both PEP and NEP, and class III genes (e.g. the gene *accD* and the *rpoB* operon) are transcribed by NEP (Hajdukiewicz et al. 1997). Chloroplasts represent a highly oxidative environment leading to an increased mutation rate. Post-transcriptional repair by RNA editing is widely used to restore affected genes by insertion, deletion or modification of specific nucleotides. Using this mechanism, mainly C to U, but also some U to C, conversions are conducted (Kotera et al. 2005).

Chloroplast genes have been extensively used for phylogenetic and phylogeographic questions. A combination of the genes *rbcL* and *matK* has been proposed as barcode for the identification of plant species (Hollingsworth et al. 2009). Conserved primers for easy amplification of the *rbcL* gene are available, and it is by far the most widely used gene in plant systematics, with over 50,000 published sequences in NCBI GenBank (Li et al. 2015). Several other plastid genome regions or fragments such as *atpF-H*, matK, *psbK-I*, *rbcL*, *ropC1*, *rpoB*, *trnH-psbA* and *trnL-F* have been also widely used in plant molecular systematic studies. Especially the advent of next-generation sequencing techniques enabled an increase in sequencing of complete chloroplast genomes for phylogenomic studies of land plants and green algae (Ruhfel et al. 2014; Lemieux et al. 2014).

2.2.3 Plastids in the Amoeba *Paulinella chromatophora*

The amoeba *Paulinella chromatophora* (Fig. 2.7) also contains two plastids with photosynthetic activities which have been demonstrated to likely originate from an

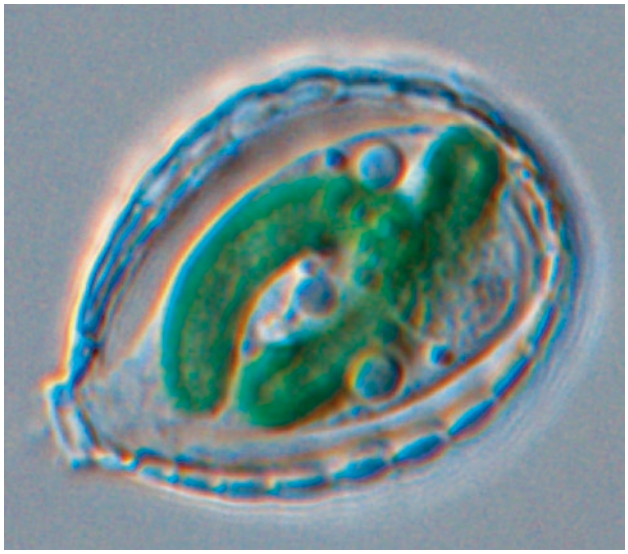


Fig. 2.7 The amoeba *Paulinella chromatophora* contains plastids with photosynthetic activity originating from an independent endosymbiotic uptake (Picture provided by Eva C.M. Nowack)

independent endosymbiotic uptake from a cyanobacterium (Marin et al. 2005; Nowack 2014). Interestingly, the endosymbiotic origin of these organelles is with an assumed age of 60–200 mya much younger than the primary endosymbiotic uptake of plastids by Archaeplastida. Morphological studies and sequencing of rRNA genes suggested that these plastids stem from a member of the cyanobacterial *Synechococcus* clade. The plastids bear two envelope membranes surrounding a thick peptidoglycan wall. It could be shown that the endosymbiont lost 75% of its ancestral genome (Bodył et al. 2012; Nowack et al. 2008). Reduction of the chromatophore genome led to the loss of many important biosynthetic pathways. For compensation, numerous (229) nuclear genes were acquired by horizontal gene transfer of which around 25% came from the endosymbiont (Nowack et al. 2016). The evolution and establishment of such a protein import mechanism qualify the chromatophores of *Paulinella* as cell organelles in a strict sense (Bodył et al. 2012; Keeling and Archibald 2008; Nowack and Grossman 2012). Conversely, some genes seem also to be imported from the host genome into the plastid genome (Mackiewicz and Bodył 2010).

2.3 Heritable Bacterial Endosymbionts

2.3.1 Primary Endosymbionts

Mitochondrial and chloroplast organelles evolved from an ancient symbiosis of bacteria with its archaeal or eukaryote host. Heritable bacterial endosymbionts are widespread across eukaryotic taxa, and complex relationships between host and symbionts have been described. Whereas most described endosymbionts belong to Bacteria, some examples from Archaea are also known (van Hoek et al. 2000). The best investigated examples stem from insects, where bacterial endosymbionts are often inherited maternally, as also typical for mitochondria (Ferrari and Vavre 2011). Endosymbionts can be loosely classified into either primary symbionts (P-symbionts) or secondary symbionts (S-symbionts) (Moran

Fig. 2.8 Tsetse flies (*Glossina* spp.) always carry the P-symbiont *Wigglesworthia*, which resides in a specialized organ called bacteriome. Furthermore, often the presence of two S-symbionts can be observed: *Sodalis* and *Wolbachia* (Reprinted by permission from Elsevier Ltd: Trends in Parasitology (Weiss and Aksoy 2011), Copyright 2011)

et al. 2008). P-symbionts are obligate mutualists which are required for the survival or reproduction of the host. Typically, such endosymbionts reside in specialized organs called bacteriomes. Well-studied examples are endosymbionts of insects with a specialized diet. E.g. tsetse flies (*Glossina* sp.) feed exclusively on blood and rely on microbial symbionts to supply amino acids and vitamins the host is not able to synthesize (Aksoy 2000). These flies harbour the gammaproteobacterium *Wigglesworthia glossinidia* as P-symbiont, which resides in specific epithelial cells (bacteriocytes) forming the bacteriome (Balmand et al. 2013) (Fig. 2.8). *Wigglesworthia* provides its host with vitamins and supports the digestion of the blood meal. Moreover, female tsetse flies cured from its endosymbiont are infertile (Pais et al. 2008). As typical for P-symbionts, *Wigglesworthia* has a streamlined and highly reduced genome of only 700 Kb (Akman et al. 2002).

The smallest reported genomes are found in P-symbionts of sap-feeding insects, often retaining only a minimal gene set (McCutcheon and Moran 2012). With 139 Kb, the smallest genome is reported for the mealybug P-symbiont *Candidatus* Tremblaya princeps (López-Madrigal et al. 2011). The genome of this betaproteobacterium contains only 120 protein-coding genes and misses several essential genes. Interestingly, in the cytoplasm of Tremblaya is another endosymbiont resident, the gammaproteobacterium *Candidatus* Moranella endobia, which supports essential functions of its bacterial host (von Dohlen et al. 2001; Husnik et al. 2013). This highly degenerated endosymbiont genomes led to a complete dependency of their hosts, often blurring the distinction between organelles and endosymbionts (McCutcheon and Keeling 2014). For example, a protein of a gene, which has been horizontally transferred to its host nucleus, has been demonstrated to be transported back to its obligate endosymbiont in aphids (Nakabachi et al. 2014). The evolution of protein targeting systems to redirect the products of horizontally transferred genes back to the symbiont is regarded as one of the major transitions in organelle evolution (Cavalier-Smith and Lee 1985). Aphids usually possess intracellular gammaproteobacteria of the genus *Buchnera*, which are transmitted vertically (to the offspring) via the ovary. This symbiotic relationship is obligate for both partners, as aphids without symbionts have a low fitness or are infertile, and *Buchnera* are unknown outside their aphid hosts (Douglas

1998). The symbiosis between aphids and *Buchnera* is very old, suggested to be established ~200 mya. Consequently, co-diversification between aphid hosts and their *Buchnera* symbionts can be found (Baumann 2005), and phylogenetic analyses of symbiont genes were even helpful to resolve aphid relationships (Novakova et al. 2013).

2.3.2 Secondary Endosymbionts

S-symbionts are bacterial symbionts which can be facultative mutualists and reproductive manipulators or have completely unknown effects on their host (Moran et al. 2008). Usually these symbionts reside in different cell types of their hosts and often invade reproductive organs, but can also be found in fluids of the body cavity. Unlike P-symbionts, there is usually no co-diversification found between S-symbionts and their hosts. Prevalence of S-symbionts in host populations can range from infecting some few up to all individuals, e.g. tsetse flies often carry two S-symbionts: the facultative mutualist *Sodalis glossinidius* (Enterobacteriaceae) and the alphaproteobacterium *Wolbachia pipientis* (Fig. 2.8). Whereas the P-symbiont is found in all tsetse flies, the infection prevalence of the S-symbionts can strongly vary across species and populations ranging from 1.4% up to 93.7% in the case of *Sodalis* (Dennis et al. 2014). *Wolbachia* is found in arthropods and filarial nematodes, and it is estimated that infections occur in ~40% of all terrestrial arthropod species (Zug and Hammerstein 2012). This ubiquity explains why *Wolbachia* is one of the best investigated endosymbionts and relationships with their hosts are investigated in many cases. Within *Wolbachia*, many different supergroups are described, which differ in their host range and their symbiotic relationships (Gerth et al. 2014). By far the most of the known *Wolbachia* strains belong to either supergroups A or B, mostly infecting insects, but also other arthropod species. Interestingly, it has been shown that the origin of these supergroups coincides with the diversification of hyperdiverse insect lineages ~200 mya (Gerth and Bleidorn 2016). *Wolbachia* are well known as reproductive manipulators of their hosts. As vertical transmission is exclusively maternal, several mechanisms to enhance the spread across the host population are described. These include distortion of the host population sex ratio via parthenogenesis, male killing or feminization, as well as induction of cytoplasmic incompatibility (CI). In the case of CI, eggs of uninfected females are incompatible with the sperm of infected males, thereby preventing successful mating (Werren et al. 2008). As typical for S-symbionts, there is usually no co-diversification pattern between *Wolbachia* and their arthropod hosts, suggesting repeated instances of horizontal transmission between unrelated species (Werren et al. 1995).

2.4 DNA Barcoding

DNA-based species identification by a universal DNA barcode of few standard DNA regions became firstly established for animals (Hebert et al. 2003a) and later also standard in plants, fungi and other eukaryotes (Hollingsworth et al. 2009; Schoch et al. 2012; Saunders and McDevit 2012). The idea to use molecular markers for species identification and delimitation was already in use for decades in prokaryotes (Tindall et al. 2010), most commonly utilizing the 16S rRNA gene. For eukaryotes, DNA barcoding has been most successfully developed for animals, where a 658 bp region of the mitochondrial cytochrome oxidase 1 gene (*cox1*) is used as standard marker. The choice of this mitochondrial

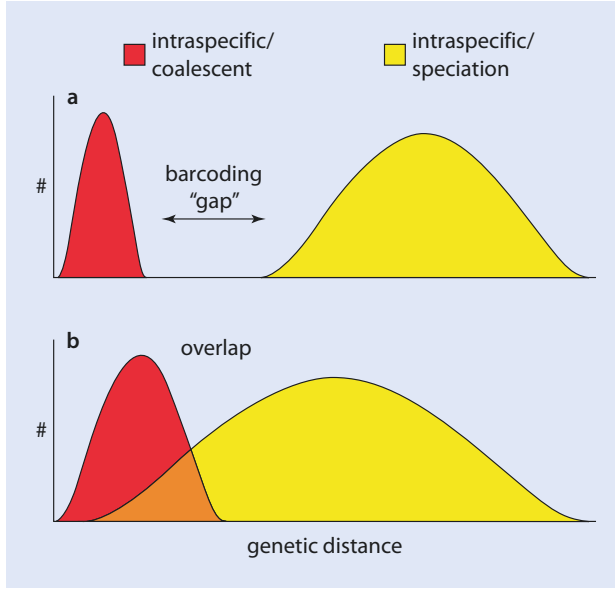


Fig. 2.9 Schematic representation of the DNA barcoding gap. **a** In the ideal case, there is no overlap between intraspecific and interspecific genetic variability, thereby creating a barcoding gap. **b** In many «real-world» examples, an area of overlap of the genetic variability between interspecific and intraspecific comparisons exists (Reprinted from Meyer and Paulay (2005))

marker has several advantages: (I) nearly universal primer pairs for the *cox1* fragment are available (Folmer et al. 1994); (II) mitochondrial DNA is available in a much higher copy number per cell than nuclear DNA, thereby alleviating DNA extraction and amplification; and (III) the existence of a «barcoding gap» is proposed (Fig. 2.9), where interspecific genetic variation clearly exceeds intraspecific variation (Hebert et al. 2003b). Mitochondrial DNA of plants evolves much slower than its animal counterpart, and consequently with the genes *rbcl* and *matk*, two chloroplast markers are currently in use for DNA barcoding of plants (Hollingsworth et al. 2009). In fungi, barcoding relies on the nuclear ITS regions, so-called internal transcribed spacers separating the tandemly repeated ribosomal RNA genes (Schoch et al. 2012). As reference for this approach serves the Barcode of Life Data Systems (BOLD) (Ratnasingham and Hebert 2007), which links specimen information, metadata and genetic sequence data. The Consortium for the Barcode of Life (CBOL) coordinates and promotes the standardization of DNA barcoding. Many country- and taxon-specific initiatives contribute to the growth of the database. In February 2016, around 2.5 million *cox1* sequences from more than 175,000 animal species were accessible in BOLD. DNA barcoding offers several practical applications including protection of endangered species, product authentication, control of invasive and pest species, biodiversity monitoring, diet analyses, linking larval of developmental with adult stages and the discovery of new species.

The reliance on one or few markers also promoted several critiques of the barcoding approach. Especially, focussing solely on organelle markers may be misleading due to reduced effective population size, introgression, maternal inheritance, inconsistent mutation rate, pseudogenization or heteroplasmy (Galtier et al. 2009). The presence of endosymbionts manipulating the host reproduction and thereby altering inheritance patterns of maternally transmitted genes may imply further complications (Gerth et al. 2011). Moreover, the existence of a «barcoding gap» might be an artefact generated through an insufficient sampling across taxa and populations (Wiemers and Fiedler 2007). Nevertheless, DNA barcoding became popular, and especially the advent of

next-generation sequencing techniques allowed metabarcoding studies estimating the diversity of communities previously difficult to handle as, e.g. from soil, permafrost or the deep sea (Valentini et al. 2009). Metabarcoding describes the simultaneous amplification of DNA barcodes from mass collections of organisms or environmental DNA (Yu et al. 2012). Such studies usually discover a huge amount of DNA sequences which do not match with any entry for BOLD, and species-delimitation methods are needed for classification. The most popular methods are based on the generalized mixed Yule coalescent (GMYC) model (Pons et al. 2006) or Poisson tree processes (PTP) (Zhang et al. 2013).

References

Akman L, Yamashita A, Watanabe H, Oshima K, Shiba T, Hattori M, Aksoy S (2002) Genome sequence of the endocellular obligate symbiont of tsetse flies, *Wigglesworthia glossinidia*. Nat Genet 32:402–407

Aksoy S (2000) Tsetse – a haven for microorganisms. Parasitol Today 16:114–118

Allen J (2015) Why chloroplasts and mitochondria retain their own genomes and genetic systems: colocation for redox regulation of gene expression. Proc Natl Acad Sci U S A 112:10231–10238

Avise JC (2004) Molecular markers, natural history, and evolution. Sinauer Associates, Inc, Sunderland

Balmand S, Lohs C, Aksoy S, Heddi A (2013) Tissue distribution and transmission routes for the tsetse fly endosymbionts. J Invertebr Pathol 112(Suppl 1):S116–S122

Baumann P (2005) Biology of bacteriocyte-associated endosymbionts of plant sap-sucking insects. Annu Rev Microbiol 59:155–189

Bendich AJ (2004) Circular chloroplast chromosomes: the grand illusion. Plant Cell 16:1661–1666

Bernt M, Braband A, Schierwater B, Stadler PF (2013) Genetic aspects of mitochondrial genome evolution. Mol Phylogenet Evol 69:328–338

Bodył A, Mackiewicz P, Gagat P (2012) Organelle evolution: *Paulinella* breaks a paradigm. Curr Biol 22:R304–R306

Börner GV, Yokobori S-I, Mörl M, Dörner M, Pääbo S (1997) RNA editing in metazoan mitochondria: staying fit without sex. FEBS Lett 409:320–324

Brouard J-S, Otis C, Lemieux C, Turmel M (2010) The exceptionally large chloroplast genome of the green alga *Floydiella terrestris* illuminates the evolutionary history of the Chlorophyceae. Genome Biol Evol 2:240–256

Brown WM, George M, Wilson AC (1979) Rapid evolution of animal mitochondrial DNA. Proc Natl Acad Sci U S A 76:1967–1971

Burger G, Forget L, Zhu Y, Gray MW, Lang BF (2003a) Unique mitochondrial genome architecture in unicellular relatives of animals. Proc Natl Acad Sci U S A 100:892–897

Burger G, Gray MW, Forget L, Lang BF (2013) Strikingly bacteria-like and gene-rich mitochondrial genomes throughout jakobid protists. Genome Biol Evol 5:418–438

Burger G, Gray MW, Franz Lang B (2003b) Mitochondrial genomes: anything goes. Trends Genet 19:709–716

Burger G, Zhu Y, Littlejohn TG, Greenwood SJ, Schnare MN, Lang BF, Gray MW (2000) Complete sequence of the mitochondrial genome of *Tetrahymena pyriformis* and comparison with *Paramecium aurelia* mitochondrial DNA. J Mol Biol 297:365–380

Cameron SL, Yoshizawa K, Mizukoshi A, Whiting MF, Johnson KP (2011) Mitochondrial genome deletions and minicircles are common in lice (Insecta: Phthiraptera). BMC Genomics 12:394

Cavalier-Smith T (1983) A 6-kingdom classification and a unified phylogeny. In: Schenk HEA, Schwemmler WS (eds) Endocytobiology. II. Intracellular Space as Oligogenetic Ecosystem. Walter de Gruyter, Berlin, pp 1027–1034

Cavalier-Smith T (1987) Eukaryotes with no mitochondria. Nature 326:332–333

Cavalier-Smith T (2003) Genomic reduction and evolution of novel genetic membranes and protein-targeting machinery in eukaryote-eukaryote chimaeras (meta-algae). Philos Trans R Soc Lond Ser B Biol Sci 358:109–134

Cavalier-Smith T, Lee JJ (1985) Protozoa as hosts for endosymbioses and the conversion of symbionts into organelles. J Protozool 32:376–379

Chaw S-M, Chun-Chieh Shih A, Wang D, Wu Y-W, Liu S-M, Chou T-Y (2008) The mitochondrial genome of the gymnosperm *Cycas taitungensis* contains a novel family of short interspersed elements, Bpu sequences, and abundant RNA editing sites. Mol Biol Evol 25:603–615

Christensen AC (2013) Plant mitochondrial genome evolution can be explained by DNA repair mechanisms. Genome Biol Evol 5:1079–1086

Danovaro R, Dell'Anno A, Pusceddu A, Gambi C, Heiner I, Møbjerg Kristensen R (2010) The first metazoa living in permanently anoxic conditions. BMC Biol 8:1–10

Dellaporta SL, Xu A, Sagasser S, Jakob W, Moreno MA, Buss LW, Schierwater B (2006) Mitochondrial genome of *Trichoplax adhaerens* supports Placozoa as the basal lower metazoan phylum. Proc Natl Acad Sci U S A 103:8751–8756

Dennis JW, Durkin SM, Horsley Downie JE, Hamill LC, Anderson NE, MacLeod ET (2014) *Sodalis glossinidius* prevalence and trypanosome presence in tsetse from Luambe National Park, Zambia. Parasit Vectors 7:1–11

Dolezal P, Likic V, Tachezy J, Lithgow T (2006) Evolution of the molecular machines for protein import into mitochondria. Science 313:314–318

Dorrell RG, Howe CJ (2015) Integration of plastids with their hosts: lessons learned from dinoflagellates. Proc Natl Acad Sci USA 112(33):10247–10254

Doucet-Beaupré H, Breton S, Chapman EG, Blier PU, Bogan AE, Stewart DT, Hoeh WR (2010) Mitochondrial phylogenomics of the Bivalvia (Mollusca): searching for the origin and mitogenomic correlates of doubly uniparental inheritance of mtDNA. BMC Evol Biol 10:50

Douglas AE (1998) Nutritional interactions in insect-microbial symbioses: aphids and their symbiotic bacteria *Buchnera*. Annu Rev Entomol 43:17–37

Embley TM, Martin W (2006) Eukaryotic evolution, changes and challenges. Nature 440:623–630

Ettema TJG (2016) Evolution: mitochondria in the second act. Nature 531:39–40

Fan J, Lee RW (2002) Mitochondrial genome of the colorless green alga *Polytomella parva*: two linear DNA Molecules with homologous inverted repeat termini. Mol Biol Evol 19:999–1007

Feagin JE, Mericle BL, Werner E, Morris M (1997) Identification of additional rRNA fragments encoded by the *Plasmodium falciparum* 6 kb element. Nucleic Acids Res 25:438–446

Ferrari J, Vavre F (2011) Bacterial symbionts in insects or the story of communities affecting communities. Philos Trans R Soc Lond Ser B Biol Sci 366:1389–1400

Folmer O, Black M, Hoeh W, Lutz R, Vrijenhoek R (1994) DNA primers for amplification of mitochondrial cytochrome c oxidase subunit I from diverse metazoan invertebrates. Mol Mar Biol Biotechnol 3:294–299

Gabaldón T, Huynen MA (2007) From endosymbiont to host-controlled organelle: the hijacking of mitochondrial protein synthesis and metabolism. PLoS Comput Biol 3:e219

Galtier N, Nabholz B, Glemin S, Hurst GDD (2009) Mitochondrial DNA as a marker of molecular diversity: a reappraisal. Mol Ecol 18:4541–4550

Gerth M, Bleidorn C (2016) Comparative genomics provides a timeframe for *Wolbachia* evolution and exposes a recent biotin synthesis operon transfer. Nat Microbiol 2:16241

Gerth M, Gansauge MT, Weigert A, Bleidorn C (2014) Phylogenomic analyses uncover origin and spread of the *Wolbachia* pandemic. Nat Commun 5:5117

Gerth M, Geißler A, Bleidorn C (2011) *Wolbachia* infections in bees (Anthophila) and possible implications for DNA barcoding. Syst Biodivers 9:319–327

Gibson T, Blok VC, Phillips MS, Hong G, Kumarasinghe D, Riley IT, Dowton M (2007) The mitochondrial subgenomes of the nematode *Globodera pallida* are mosaics: evidence of recombination in an animal mitochondrial genome. J Mol Evol 64:463–471

Gissi C, Iannelli F, Pesole G (2008) Evolution of the mitochondrial genome of Metazoa as exemplified by comparison of congeneric species. Heredity 101:301–320

Gould SB, Waller RF, McFadden GI (2008) Plastid evolution. Annu Rev Plant Biol 59:491–517

Gray MW (1999) Evolution of organellar genomes. Curr Opin Genet Dev 9:678–687

Gray MW (2012) Mitochondrial evolution. Cold Spring Harb Perspect Biol 4:a011403

Gray MW (2015) Mosaic nature of the mitochondrial proteome: implications for the origin and evolution of mitochondria. Proc Natl Acad Sci U S A 112:10133–10138

Gray MW, Archibald JM (2012) Origins of mitochondria and plastids. In: Bock R, Knoop V (eds) Genomics of chloroplasts and mitochondria, Advances in photosynthesis and respiration, vol 35. Springer Science + Business Media B.V, Dordrecht, pp 1–30

Green BR (2011) Chloroplast genomes of photosynthetic eukaryotes. Plant J 66:34–44

Hajdukiewicz PTJ, Allison LA, Maliga P (1997) The two RNA polymerases encoded by the nuclear and the plastid compartments transcribe distinct groups of genes in tobacco plastids. EMBO J 16:4041–4048
Hebert PDN, Cywinska A, Ball SL, deWaard JR (2003a) Biological identifications through DNA barcodes. Proc R Soc Lond B Biol Sci 270:313–321
Hebert PDN, Ratnasingham S, de Waard JR (2003b) Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species. Proc R Soc Lond B Biol Sci 270:S96–S99
Hikosaka K, Kita K, Tanabe K (2013) Diversity of mitochondrial genome structure in the phylum Apicomplexa. Mol Biochem Parasitol 188:26–33
Hikosaka K, Watanabe Y-I, Tsuji N, Kita K, Kishine H, Arisue N, Palacpac NMQ, Kawazu S-I, Sawai H, Horii T, Igarashi I, Tanabe K (2010) Divergence of the mitochondrial genome structure in the apicomplexan parasites, *Babesia* and *Theileria*. Mol Biol Evol 27:1107–1116
Hjort K, Goldberg AV, Tsaousis AD, Hirt RP, Embley TM (2010) Diversity and reductive evolution of mitochondria among microbial eukaryotes. Philos Trans R Soc Lond Ser B Biol Sci 365:713–727
Hohmann-Marriott MF, Blankenship RE (2011) Evolution of photosynthesis. Annu Rev Plant Biol 62:515–548
Hollingsworth PM, Forrest LL, Spouge JL, Hajibabaei M, Ratnasingham S, van der Bank M, Chase MW, Cowan RS, Erickson DL, Fazekas AJ, Graham SW, James KE, Kim K-J, Kress WJ, Schneider H, van AlphenStahl J, Barrett SCH, van den Berg C, Bogarin D, Burgess KS, Cameron KM, Carine M, Chacón J, Clark A, Clarkson JJ, Conrad F, Devey DS, Ford CS, Hedderson TAJ, Hollingsworth ML, Husband BC, Kelly LJ, Kesanakurti PR, Kim JS, Kim Y-D, Lahaye R, Lee H-L, Long DG, Madriñán S, Maurin O, Meusnier I, Newmaster SG, Park C-W, Percy DM, Petersen G, Richardson JE, Salazar GA, Savolainen V, Seberg O, Wilkinson MJ, Yi D-K, Little DP (2009) A DNA barcode for land plants. Proc Natl Acad Sci U S A 106:12794–12797
Huang D, Meier R, Todd PA, Chou LM (2008) Slow mitochondrial COI sequence evolution at the base of the metazoan tree and its implications for DNA barcoding. J Mol Evol 66:167–174
Huchon D, Szitenberg A, Shefer S, Ilan M, Feldstein T (2015) Mitochondrial group I and group II introns in the sponge orders Agelasida and Axinellida. BMC Evol Biol 15:278
Husnik F, Nikoh N, Koga R, Ross L, Duncan Rebecca P, Fujie M, Tanaka M, Satoh N, Bachtrog D, Wilson Alex CC, von Dohlen CD, Fukatsu T, McCutcheon John P (2013) Horizontal gene transfer from diverse bacteria to an insect genome enables a tripartite nested mealybug symbiosis. Cell 153:1567–1578
Jackson CJ, Gornik SG, Waller RF (2012) The mitochondrial genome and transcriptome of the basal dinoflagellate *Hematodinium* sp.: character evolution within the highly derived mitochondrial genomes of dinoflagellates. Genome Biol Evol 4:59–72
Janouškovec J, Liu S-L, Martone PT, Carré W, Leblanc C, Collén J, Keeling PJ (2013) Evolution of red algal plastid genomes: ancient architectures, introns, horizontal gene transfer, and taxonomic utility of plastid markers. PLoS One 8:e59001
Karnkowska A, Vacek V, Zubáčová Z, Treitli Sebastian C, Petrželková R, Eme L, Novák L, Žárský V, Barlow Lael D, Herman Emily K, Soukal P, Hroudová M, Doležal P, Stairs Courtney W, Roger Andrew J, Eliáš M, Dacks Joel B, Vlček Č, Hampl V (2016) A eukaryote without a mitochondrial organelle. Curr Biol 26:1274–1284
Katz LA, Grant JR (2014) Taxon-rich phylogenomic analyses resolve the eukaryotic tree of life and reveal the power of subsampling by sites. Syst Biol 64:406–415
Kayal E, Bentlage B, Collins AG, Kayal M, Pirro S, Lavrov DV (2012) Evolution of linear mitochondrial genomes in medusozoan cnidarians. Genome Biol Evol 4:1–12
Keeling PJ (1998) A kingdom's progress: Archezoa and the origin of eukaryotes. BioEssays 20:87–95
Keeling PJ (2009) Role of horizontal gene transfer in the evolution of photosynthetic eukaryotes and their plastids. In: Gogarten MB, Gogarten JP, Olendzenski LC (eds) Horizontal gene transfer, Methods in molecular biology, vol 532. Humana Press, New York, pp 501–515
Keeling PJ (2010) The endosymbiotic origin, diversification and fate of plastids. Philos Trans R Soc Lond Ser B Biol Sci 365:729–748
Keeling PJ (2013) The number, speed, and impact of plastid endosymbioses in eukaryotic evolution. Annu Rev Plant Biol 64:583–607
Keeling PJ, Archibald JM (2008) Organelle evolution: what's in a name? Curr Biol 18:R345–R347
Knight RD, Freeland SJ, Landweber LF (2001) Rewiring the keyboard: evolvability of the genetic code. Nat Rev Genet 2:49–58
Knoop V (2012) Seed plant mitochondrial genomes: complexity evolving. In: Bock R, Knoop V (eds) Genomics of chloroplasts and mitochondria, Advances in photosynthesis and respiration, vol 35. Springer Science + Business Media B.V, Dordrecht, pp 175–200

Köhler S, Delwiche CF, Denny PW, Tilney LG, Webster P, Wilson RJM, Palmer JD, Roos DS (1997) A plastid of probable green algal origin in apicomplexan parasites. Science 275:1485–1489
Kotera E, Tasaka M, Shikanai T (2005) A pentatricopeptide repeat protein is essential for RNA editing in chloroplasts. Nature 433:326–330
Lang BF, Burger G, O'Kelly CJ, Cedergren R, Golding GB, Lemieux C, Sankoff D, Turmel M, Gray MW (1997) An ancestral mitochondrial DNA resembling a eubacterial genome in miniature. Nature 387:493–497
Lemieux C, Otis C, Turmel M (2014) Chloroplast phylogenomic analysis resolves deep-level relationships within the green algal class Trebouxiophyceae. BMC Evol Biol 14:211
Li X, Yang Y, Henry RJ, Rossetto M, Wang Y, Chen S (2015) Plant DNA barcoding: from gene to genome. Biol Rev 90:157–166
Lim L, McFadden GI (2010) The evolution, metabolism and functions of the apicoplast. Philos Trans R Soc Lond Ser B Biol Sci 365:749–763
Lithgow T, Schneider A (2010) Evolution of macromolecular import pathways in mitochondria, hydrogenosomes and mitosomes. Philos Trans R Soc Lond Ser B Biol Sci 365:799–817
Liu Y, Wang B, Cui P, Li L, Xue J-Y, Yu J, Qiu Y-L (2012) The mitochondrial genome of the lycophyte *Huperzia squarrosa*: the most archaic form in vascular plants. PLoS One 7:e35168
Lohse M, Drechsel O, Kahlau S, Bock R (2013) OrganellarGenomeDRAW – a suite of tools for generating physical maps of plastid and mitochondrial genomes and visualizing expression data sets. Nucleic Acids Res 41:W575–W581
López-Madrigal S, Latorre A, Porcar M, Moya A, Gil R (2011) Complete genome sequence of «*Candidatus* Tremblaya princeps» strain PCVAL, an intriguing translational machine below the living-cell status. J Bacteriol 193:5587–5588
Lynch M (2007) The origins of genome architecture. Sinauer Assoc, Sunderland
Mackiewicz P, Bodył A (2010) A hypothesis for import of the nuclear encoded *PsaE* Protein of *Paulinella chromatophora* (Cercozoa, Rhizaria) into its cyanobacterial endosymbionts/plastids via the endomembrane system. J Phycol 46:847–859
Margulis L (1970) Origin of eukaryotic cells. Yale University Press, New Haven
Marin B, Nowack EC, Melkonian M (2005) A plastid in the making: evidence for a second primary endosymbiosis. Protist 156:425–432
Martin G, Baurens F-C, Cardi C, Aury J-M, D'Hont A (2013) The complete chloroplast genome of banana (*Musa acuminata*, Zingiberales): insight into plastid monocotyledon evolution. PLoS One 8:e67350
Martin W, Muller M (1998) The hydrogen hypothesis for the first eukaryote. Nature 392:37–41
McCutcheon JP, Keeling PJ (2014) Endosymbiosis: protein targeting further erodes the organelle/symbiont distinction. Curr Biol 24:R654–R655
McCutcheon JP, Moran NA (2012) Extreme genome reduction in symbiotic bacteria. Nat Rev Microbiol 10:13–26
McInerney JO, O'Connell MJ, Pisani D (2014) The hybrid nature of the Eukaryota and a consilient view of life on earth. Nat Rev Microbiol 12:449–455
Meyer CP, Paulay G (2005) DNA barcoding: error rates based on comprehensive sampling. PLoS Biol 3:e422
Moran NA, McCutcheon JP, Nakabachi A (2008) Genomics and evolution of heritable bacterial symbionts. Annu Rev Genet 42:165–190
Mower JP, Sloan DB, Alverson AJ (2012) Plant mitochondrial genome diversity: the genomics revolution. In: Wendel FJ, Greilhuber J, Dolezel J, Leitch JI (eds) Plant genome diversity, Plant genomes, their residents, and their evolutionary dynamics, vol 1. Springer, Vienna, pp 123–144
Mwinyi A, Meyer A, Bleidorn C, Lieb B, Bartolomaeus T, Podsiadlowski L (2009) Mitochondrial genome sequence and gene order of *Sipunculus nudus* give additional support for an inclusion of Sipuncula into Annelida. BMC Genomics 10:27
Nakabachi A, Ishida K, Hongoh Y, Ohkuma M, Miyagishima S-Y (2014) Aphid gene of bacterial origin encodes a protein transported to an obligate endosymbiont. Curr Biol 24:R640–R641
Nosek J, Tomáška Ľ (2003) Mitochondrial genome diversity: evolution of the molecular architecture and replication strategy. Curr Genet 44:73–84
Novakova E, Hypša V, Klein J, Foottit R, Von Dohlen CD, Moran NA (2013) Reconstructing the phylogeny of aphids (Hemiptera: Aphididae) using DNA of the obligate symbiont *Buchnera aphidicola*. Mol Phylogenet Evol 68:42–54
Nowack ECM (2014) *Paulinella chromatophora* – rethinking the transition from endosymbiont to organelle. Acta Soc Bot Pol 83:387–397

Nowack ECM, Grossman AR (2012) Trafficking of protein into the recently established photosynthetic organelles of *Paulinella chromatophora*. Proc Natl Acad Sci U S A 109:5340–5345
Nowack ECM, Melkonian M, Glöckner G (2008) Chromatophore genome sequence of *Paulinella* sheds light on acquisition of photosynthesis by eukaryotes. Curr Biol 18:410–418
Nowack ECM, Price DC, Bhattacharya D, Singer A, Melkonian M, Grossman AR (2016) Gene transfers from diverse bacteria compensate for reductive genome evolution in the chromatophore of *Paulinella chromatophora*. Proc Natl Acad Sci U S A 113:12214–12219
Pais R, Lohs C, Wu Y, Wang J, Aksoy S (2008) The obligate mutualist *Wigglesworthia glossinidia* Influences reproduction, digestion, and immunity processes of Its host, the Tsetse fly. Appl Environ Microbiol 74:5965–5974
Passamonti M, Ricci A, Milani L, Ghiselli F (2011) Mitochondrial genomes and doubly uniparental inheritance: new insights from *Musculista senhousia* sex-linked mitochondrial DNAs (Bivalvia Mytilidae). BMC Genomics 12:442
Pittis AA, Gabaldón T (2016) Late acquisition of mitochondria by a host with chimaeric prokaryotic ancestry. Nature 531:101–104
Pons J, Barraclough TG, Gomez-Zurita J, Cardoso A, Duran DP, Hazell S, Kamoun S, Sumlin WD, Vogler AP (2006) Sequence-based species delimitation for the DNA taxonomy of undescribed insects. Syst Biol 55:595–609
Ratnasingham S, Hebert PDN (2007) BOLD: the barcode of life data system (http://www.barcodinglife.org). Mol Ecol Notes 7:355–364
Rice DW, Alverson AJ, Richardson AO, Young GJ, Sanchez-Puerta MV, Munzinger J, Barry K, Boore JL, Zhang Y, dePamphilis CW, Knox EB, Palmer JD (2013) Horizontal transfer of entire genomes via mitochondrial fusion in the angiosperm *Amborella*. Science 342:1468–1473
Richardson AO, Rice DW, Young GJ, Alverson AJ, Palmer JD (2013) The «fossilized» mitochondrial genome of *Liriodendron tulipifera*: ancestral gene content and order, ancestral editing sites, and extraordinarily low mutation rate. BMC Biol 11:29
Ruhfel B, Gitzendanner M, Soltis P, Soltis D, Burleigh J (2014) From algae to angiosperms-inferring the phylogeny of green plants (Viridiplantae) from 360 plastid genomes. BMC Evol Biol 14:23
Saunders GW, McDevit DC (2012) Methods for DNA barcoding photosynthetic protists emphasizing the macroalgae and diatoms. In: Kress JW, Erickson LD (eds) DNA barcodes: methods and protocols. Humana Press, Totowa, pp 207–222
Schoch CL, Seifert KA, Huhndorf S, Robert V, Spouge JL, Levesque CA, Chen W, Consortium FB (2012) Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. Proc Natl Acad Sci U S A 109:6241–6246
Shearer TL, van Oppen MJH, Romano SL, Wörheide G (2002) Slow mitochondrial DNA sequence evolution in the Anthozoa (Cnidaria). Mol Ecol 11:2475–2487
Shoguchi E, Shinzato C, Hisata K, Satoh N, Mungpakdee S (2015) The large mitochondrial genome of *Symbiodinium minutum* reveals conserved noncoding sequences between dinoflagellates and apicomplexans. Genome Biol Evol 7:2237–2244
Sloan DB, Alverson AJ, Chuckalovcak JP, Wu M, McCauley DE, Palmer JD, Taylor DR (2012) Rapid evolution of enormous, multichromosomal genomes in flowering plant mitochondria with exceptionally high mutation rates. PLoS Biol 10:e1001241
Sogin ML (1989) Evolution of eukaryotic microorganisms and their small subunit ribosomal RNAs. Am Zool 29:487–499
Strittmatter P, Soll J, Bölter B (2010) The chloroplast protein import machinery: a review. In: Economou A (ed) Protein secretion, Methods in molecular biology, vol 619. Humana Press, New York, pp 307–321
Suzuki K, Miyagishima S-Y (2010) Eukaryotic and eubacterial contributions to the establishment of plastid proteome estimated by large-scale phylogenetic analyses. Mol Biol Evol 27:581–590
Tengs T, Dahlberg OJ, Shalchian-Tabrizi K, Klaveness D, Rudi K, Delwiche CF, Jakobsen KS (2000) Phylogenetic analyses Indicate that the 19′Hexanoyloxy-fucoxanthin-containing dinoflagellates have tertiary plastids of haptophyte origin. Mol Biol Evol 17:718–729
Timmis JN, Ayliffe MA, Huang CY, Martin W (2004) Endosymbiotic gene transfer: organelle genomes forge eukaryotic chromosomes. Nat Rev Genet 5:123–135
Tindall BJ, Rosselló-Móra R, Busse H-J, Ludwig W, Kämpfer P (2010) Notes on the characterization of prokaryote strains for taxonomic purposes. Int J Syst Evol Microbiol 60:249–266
Valentini A, Pompanon F, Taberlet P (2009) DNA barcoding for ecologists. Trends Ecol Evol 24:110–117

van Hoek AHAM, van Alen TA, Sprakel VSI, Leunissen JAM, Brigge T, Vogels GD, Hackstein JHP (2000) Multiple acquisition of methanogenic archaeal symbionts by anaerobic ciliates. Mol Biol Evol 17:251–258

Voigt O, Erpenbeck D, Wörheide G (2008) A fragmented metazoan organellar genome: the two mitochondrial chromosomes of *Hydra magnipapillata*. BMC Genomics 9:350

von Dohlen CD, Kohler S, Alsop ST, McManus WR (2001) Mealybug beta-proteobacterial endosymbionts contain gamma-proteobacterial symbionts. Nature 412:433–436

Waller RF, Jackson CJ (2009) Dinoflagellate mitochondrial genomes: stretching the rules of molecular biology. BioEssays 31:237–245

Wang Z, Wu M (2014) Phylogenomic reconstruction indicates mitochondrial ancestor was an energy parasite. PLoS One 9:e110685

Wang Z, Wu M (2015) An integrated phylogenomic approach toward pinpointing the origin of mitochondria. Sci Rep 5:7949

Weiss B, Aksoy S (2011) Microbiome influences on insect host vector competence. Trends Parasitol 27:514–522

Werren JH, Baldo L, Clark ME (2008) *Wolbachia*: master manipulators of invertebrate biology. Nat Rev Microbiol 6:741–751

Werren JH, Zhang W, Guo LR (1995) Evolution and phylogeny of *Wolbachia*: reproductive parasites of arthropods. Proc R Soc Lond B Biol Sci 261:55–63

Wiemers M, Fiedler K (2007) Does the DNA barcoding gap exist? – a case study in blue butterflies (Lepidoptera: Lycaenidae). Front Zool 4:8

Yagi Y, Shiina T (2014) Recent advances in the study of chloroplast gene expression and its evolution. Front Plant Sci 5:61

Yu DW, Ji Y, Emerson BC, Wang X, Ye C, Yang C, Ding Z (2012) Biodiversity soup: metabarcoding of arthropods for rapid biodiversity assessment and biomonitoring. Methods Ecol Evol 3:613–623

Zhang J, Kapli P, Pavlidis P, Stamatakis A (2013) A general species delimitation method with applications to phylogenetic placements. Bioinformatics 29:2869–2876

Zug R, Hammerstein P (2012) Still a host of hosts for *Wolbachia*: analysis of recent data suggests that 40% of terrestrial arthropod species are infected. PLoS One 7:e38544

Sequencing Techniques

C. Bleidorn, *Phylogenomics*, DOI 10.1007/978-3-319-54064-1_3

- Sanger sequencing is based on the chain termination method which relies on separating DNA by size and the incorporation of labelled modified nucleotides.
- 454 pyrosequencing measures the amount of light produced by the incorporation of nucleotides in a cascade of enzymatic reactions under the presence of a luciferase.
- Reversible terminator sequencing is a sequencing-by-synthesis approach where the incorporation of modified nucleotides is detected stepwise.
- Ion semiconductor sequencing analyses changes of hydrogen ion concentration during the incorporation of nucleotides into the DNA strand.
- Single-molecule real-time (SMRT) sequencing monitors without interruption the incorporation of differently fluorescent-tagged nucleotides by the polymerase activity.
- Nanopore sequencing detects the identity of nucleotides within the DNA strand while it is passing through a nanopore.
- The availability of next-generation sequencing (NGS) platforms transformed the field of genomics and led to a dramatic decrease in sequencing costs.

3.1 Sanger Sequencing

Two DNA sequencing techniques were developed in the mid-1970s. Allan Maxam and Walter Gilbert proposed a chemical cleavage method, which was initially widely used around molecular laboratories (Maxam and Gilbert 1977). Around the same time, Frederick Sanger and colleagues developed a chain termination method (Sanger et al. 1977). After the chemistry needed for this method became commercially available, Sanger sequencing got the standard sequencing technique for all applications. For nearly three decades, Sanger sequencing was synonym to DNA sequencing. DNA sequencing revolutionized many fields of biological and medical sciences, and fittingly Frederick Sanger and Walter Gilbert were awarded with a Nobel Prize in Chemistry in 1980, which they shared with Paul Berg (Brenner 2014).

Sanger's chain termination method is basically based on two principles: (I) DNA can be separated by size and (II) DNA polymerases is able to incorporate modified nucleotides. As DNA is negatively charged, gel electrophoresis allows the separation of DNA strands by size as larger DNA strands migrate slower to a positive electrode. By using polyacrylamide gels, it is even possible to detect minute differences of single base pairs between two different DNA strands. When separating DNA into single strands, it is possible to replicate one strand by the use of DNA polymerase II which will add one nucleotide after another complementary to the template DNA strand. Requirements are a DNA primer (oligonucleotide) that fits to the template and the availability of nucleotides (dNTPs) for incorporation. Nucleotides bear a 3′ OH group and a 5′ phosphate group which will be connected by the polymerase while removing two of the phosphates of the 5′ end. However, if the 3′ OH group is missing, it is impossible to connect another nucleotide, and the elongation of the template strand is terminated. Sanger used exactly such modified nucleotides, which have an H at the 3′ end instead of the OH group (dideoxynucleotides, ddNTPs) for his sequencing reaction. The sequencing reaction mix is composed of DNA polymerase, a primer for the target region, and a mix of dNTPs and ddNTPs. The ratio of dNTPs/ddNTPs is usually around 100 to 1, so that termination could be obtained at least once for every position of the

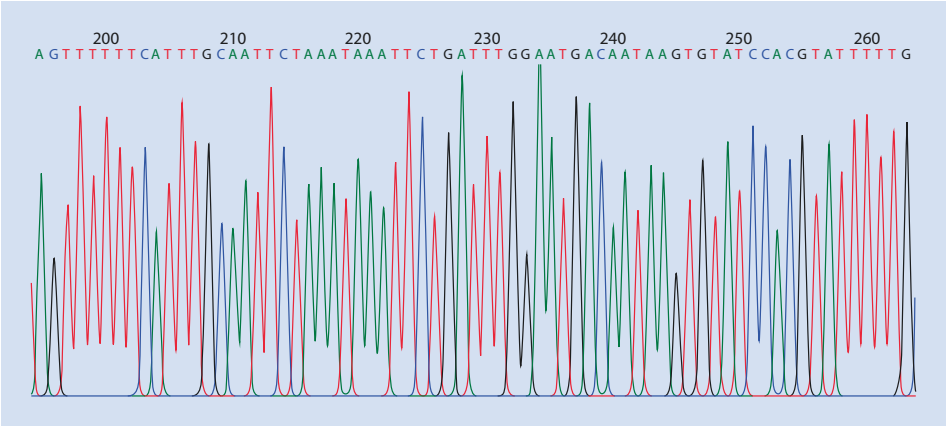


Fig. 3.1 Chromatogram of a Sanger sequencing run. All four bases are labelled differently and are separated according to their size

DNA template during the amplification process. Whenever a ddNTP is incorporated, the elongation of the template DNA stops, thereby generating DNA molecules of different sizes. Using electrophoresis, these differently sized DNA molecules can be separated, and the identity of the incorporated ddNTP will allow reconstructing the nucleotide sequence (Fig. 3.1). Initially, ddNTPs were labelled radioactively, and reactions were performed for each of the four bases separately. Huge gels had to be inspected by the eye to reconstruct the sequence order, a reason why DNA output from sequencers is still known as reads. Later on, unique fluorescent labels for all four different bases (adenosine, thymine, guanine and cytosine) were introduced which could be detected by a laser while migrating through the gel. This innovation strongly decreased analysis time, as all reactions could run at the same time and increased the accuracy of base detection (Prober et al. 1987). Computer-based analyses allow that the detected fluorescence will be automatically associated with the corresponding nucleotide to generate a chromatogram for every sequence read (Fig. 3.1).

Current machines for high-throughput analyses like the ABI 3730xl are equipped with 96 capillaries. The time for a single sequencing run will take, according to the envisaged quality, 2–3 h, with up to 1000 bp sequence reads of high quality. The output for a single run will be around 100 Kb and during 24 h sequencing of 1 Mb might be realistic (Liu et al. 2012). After trimming the ends of sequencing reads, an accuracy of 99.999% can be achieved, and sequence errors are mainly due to errors in preceding amplification steps (Kircher and Kelso 2010). As such, Sanger sequencing remains a good option for high-quality sequence reads. The first human genome sequence has been sequenced with this technique.

3.2 454 Pyrosequencing

The first next-generation sequencing (NGS) technique that was commercially available has been 454 sequencing. This technique presented solutions for each of the bottlenecks typically found in large-scale classical Sanger sequencing: library preparation, template preparation and sequencing itself (Rothberg and Leamon 2008). The principle for 454 sequencing goes back to a real-time sequencing approach called pyrosequencing developed by Ronaghi et al. (1996), and a highly parallelized high-throughput automatization was presented by Margulies

et al. (2005). The great advantage of this method compared to Sanger sequencing is that sequences are read out while being synthesized, therefore omitting electrophoresis steps for size separation of DNA fragments. In this approach, nucleotides are released one after another and washed over the template DNA strand. A cascade of enzymatic reactions leads to the emission of detectable light signal which is in its strength proportional to the number of nucleotides incorporated during this step. For example, if C's are washed over the template DNA, the polymerase incorporates as much consecutive C's as are present in the target sequence. When being incorporated by the polymerase, a pyrophosphate is released for each nucleotide, which is subsequently converted to ATP by ATP sulphurylase (Fig. 3.2). Present

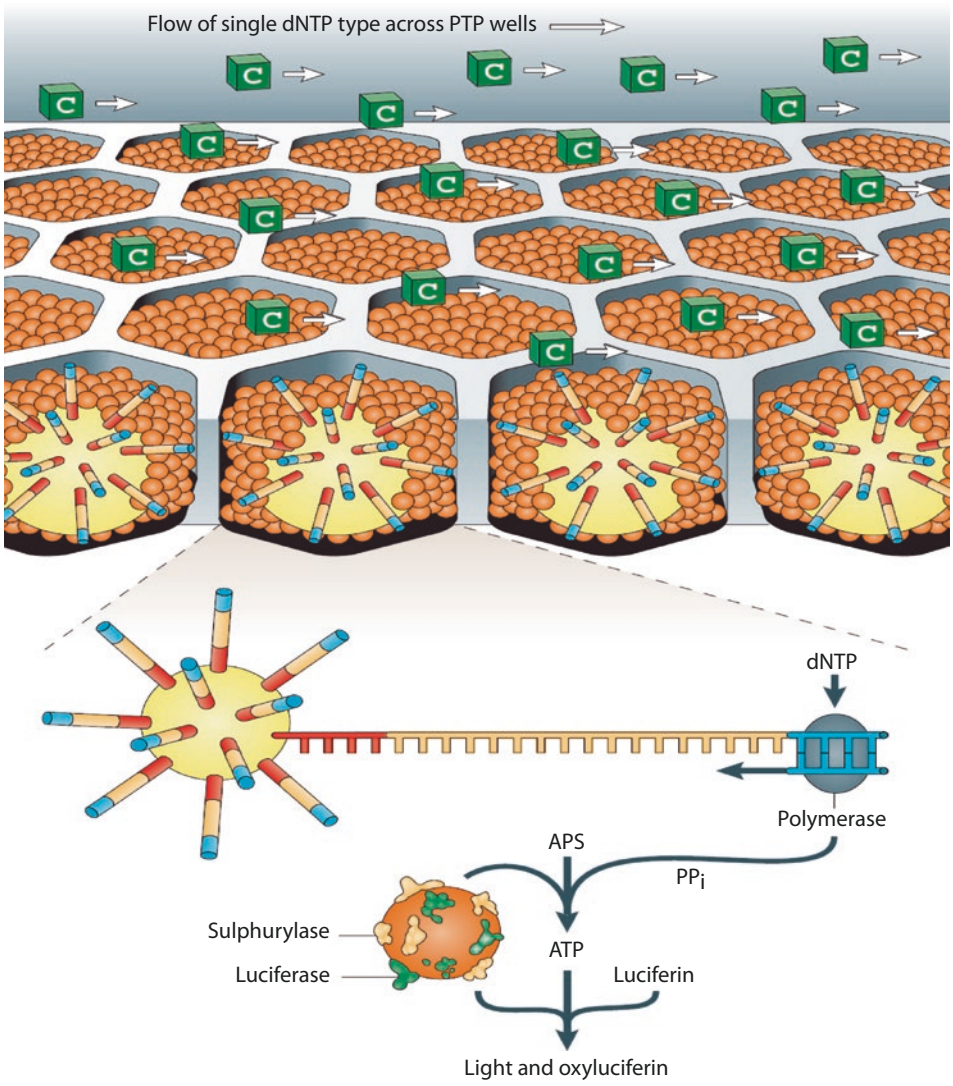


Fig. 3.2 Principles of 454 pyrosequencing in picotiter well plates. The dNTP cytosine is flushed over the wells containing beads carrying DNA fragments and reaction mix. Incorporation of a dNTP by the polymerase leads to the release of a pyrophosphate, which is converted to ATP by ATP sulphurylase. Present luciferases use this energy to oxidize luciferin, which leads to the generation of light (Reprinted by permission from Macmillan Publishers Ltd: *Nature Review Genetics* (Metzker 2010), copyright 2009)

luciferases can use this energy to oxidize luciferin, which leads to the generation of light (Ronaghi 2001). After detection of the signal, superfluous nucleotides are removed in a washing step, and the next nucleotide is provided in a subsequent flow cycle.

For the preparation of sequencing libraries, DNA is linked with special adaptors by sequential ligation and subsequently separated into single strands. One of these adaptors allows the binding to beads by hybridization. During a process called emulsion PCR, library fragments are loaded onto beads and amplified. To do this, a water mixture containing the capture beads, sequencing libraries and PCR reagents are mixed with synthetic oil in a plastic vessel. Vigorous shaking leads to the formation of droplets around the beads which by chance will usually contain a single DNA library fragment. As each droplet contains also all PCR reagents, an amplification of the library fragment will produce millions of copies which bind to the capture bead. After finishing the PCR, beads will be cleaned from the oil, and those which do not carry DNA are removed. A key to high-throughput sequencing with this method is the use of picotiter well plates for sequencing (Fig. 3.2). These plates bear approximately 1.6 million wells with a volume of 75 picoliters (Margulies et al. 2005). These wells are designed to exactly fit a single bead which carries the DNA fragments to be sequenced. Each well is filled with a capture bead and smaller beads carrying enzymes like ATP sulphurylase and luciferases. As described above, free nucleotides are washed over the well plate, and light emission is detected by a high-resolution camera. The number of filled wells determines the number of sequences to be generated in a single run.

Sequencers of Roche's 454 GS FLX series can produce around 1,000,000 sequences with an average read length of 700 bp and reads up to 1000 bp. A single run will take around 24 h, with an output of 700 Mb altogether. Four hundred fifty-four sequences are of high quality with an accuracy of ~99.75% after removing reads with N's (Huse et al. 2007). Sequencing errors are mainly indels (insertion and deletions), often found in stretches of homopolymers (Gilles et al. 2011). Due to the considerably low output in comparison to other methods, 454 sequencing lost its relevance for genome projects. However, especially for high-throughput metabarcoding studies based on amplicon sequencing, this technique is still frequently used (Yu et al. 2012; Petrosino et al. 2009). However, Roche decided to stop supporting the 454 platform in 2016.

3.3 Reversible Terminator Sequencing (Illumina)

The by far most widely used NGS platform is Illumina sequencing. Also known under its former company name as Solexa sequencing, this technique is based on cyclic reversible terminator technology. The sequencing reaction takes place on a flow cell, where literally billions of sequences can be processed during a single run. Based on a sequencing-by-synthesis approach, all four nucleotides are added simultaneously to the flow cell, together with a polymerase. Similar to Sanger sequencing, the PCR reaction is stopped after incorporating a modified base. Every incorporated nucleotide is chemically blocked at its 3′ OH group and carries a removable fluorophore which can be identified by laser. Most Illumina sequencers (GA II, HiSeq, MiSeq) use a system with four colours, one for each base. Based on four different images coming from four different colour channels, bases are called for each sequence. The Illumina NextSeq system uses only two channels for base detection. Here, red refers to a C and green to T, mixed signals (red and green) are interpreted as A and the missing of a dye refers to a G. As only two images are needed for base

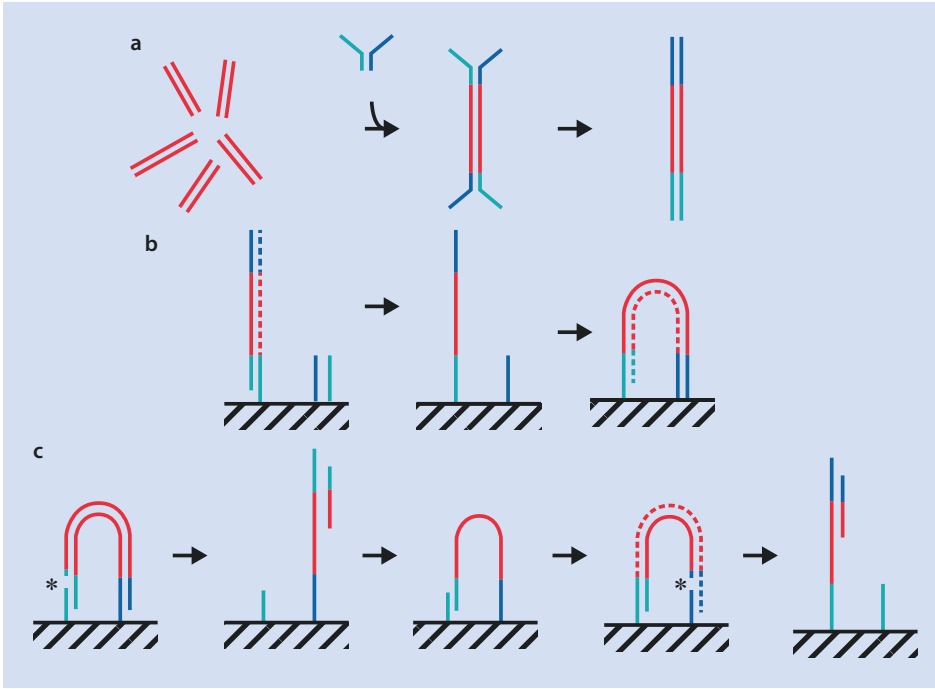


Fig. 3.3 Illumina sequencing library preparation and bridge amplification. **a** For library preparation, two different adaptors are ligated to the end of sheared DNA fragments. **b** The sequencing library is pumped into the flow cell and can bind to short oligonucleotides on its surface, which are complementary to adaptor sequences. Single-stranded molecules are copied starting from the hybridized anchor region. Newly synthesized double-stranded molecules are denatured, and the original template DNA strand is washed away. Single-stranded strands start to bend over to hybridize at their end with adjacent free anchor oligonucleotides, thereby building a bridge. Multiple PCR cycles are used to generate clusters of clonal sequences. **c** To generate single-stranded templates for sequencing by synthesis, DNA fragments are linearized by cleavage within one adaptor sequence and subsequently denatured. For paired-end sequencing, the DNA template forms a bridge, and the second strand of the DNA fragment is synthesized. This time, cleavage will take place in the opposite adaptor region to provide a template for sequencing (Reprinted by permission from Macmillan Publishers Ltd: *Nature* (Bentley et al. 2008), copyright 2008)

calling, the detection becomes faster. After detection, the blocking and the fluorophore are removed, and the process of sequencing is continued by incorporating the next nucleotide (Bentley et al. 2008). In contrast to Sanger sequencing, all incorporated nucleotides terminate the elongation process and carry fluorophores.

For library preparation, two different adaptors (P5 and P7) are ligated to the ends of all DNA molecules (Fig. 3.3a). In an additional step, indices can be added by an indexing PCR creating unique libraries, which allows pooling during sequencing. In contrast to other barcoding approaches, the indexes are placed within the adaptors and not at the ends of the molecules to be sequenced (Meyer and Kircher 2010). The resulting library is amplified with longer primers which further extend the adaptor sequences. In the next steps, the double-stranded library will be separated into single strands which are pumped into the flow cell. Two different types of short oligonucleotides which are complementary to the ends of the library adaptors are distributed as anchors across the flow cell. These anchors can hybridize to the end of adaptors and by adding nucleotides and a polymerase.

Single-stranded molecules are copied starting from the hybridized anchor region. The newly synthesized double-stranded molecule is denatured, and the original template DNA strand is washed away. As a result, all newly synthesized strands are covalently attached to the flow cell. In a step called bridge amplification, the single-stranded strands start to bend over to hybridize at their end with adjacent free anchor oligonucleotides (Fig. 3.3b). Again, the hybridized primer is extended by polymerase, which leads to the formation of a double-stranded bridge. This bridge is denatured, resulting in two copies of covalently bound single-stranded DNA templates. The bridge amplification step is repeated several times until cluster of some thousand copies are generated. In the end, the bridges are again denatured, and all reversed strands are cleaved and washed away. After blocking of the free 3′ ends, the sequencing as described above begins. While sequencing, the number of cycles will determine the number of nucleotides to be sequenced. With current machines and chemistry, usually 96–250 cycles are used to produce sequences in according length. It is also possible to use paired-end (PE) sequencing, which means that both ends of a single molecule from the sequencing library will be determined. In this case, the blocking of the free sequence ends has to be removed, and a new bridge is formed due to bending of the sequence ends to corresponding adjacent anchors (Fig. 3.3c). The single strand is replicated using a polymerase, and afterwards double strands are separated again. This time, all original forward strands are cleaved and washed away, and after blocking free ends, sequencing begins again.

Illumina sequencing can currently generate the biggest output of sequence data, and several different platforms are available. The Illumina HiSeq models are able to generate around 600 million per lane of a flow cell, which comprises eight lanes altogether, resulting into nearly 5 billion sequences for a full run. According to the number of cycles used, this would produce 750 Gb sequence output for 150 cycles. Including copying the data, such a run would last around 10 days. Faster «rapid» runs using flow cells with only two lanes, which also produce «only» around 66% of the output per lane, are available, and these can be finished in 24 h. In January 2017, Illumina announced the release of the NovaSeq models, which will replace the HiSeq series in the near future. Using NovaSeq, an output of up to 3 Tb in a 40 h run is envisaged. Two Illumina machines are available as desktop machines with the NextSeq 500 and the MiSeq. The NextSeq 500 platform runs a single lane with an output to up to 400 million sequences. The Illumina MiSeq is the cheapest model, and according to the model, the output of a single lane flow cell is between 5 and 25 million sequences. The accuracy of Illumina sequencing is with ~99.25% slightly worse than for Sanger or 454 (Quail et al. 2012). However, the error profile differs from 454 as substitutions are found much more frequently instead of indels, which is usually easier to handle in downstream analyses like mapping or assembly.

Illumina sequencing is currently standard for transcriptome sequencing and the resequencing of genomes. The read length of earlier chemistry and machines was limited to between 36 bp and 76 bp. However, with the availability of longer reads, this technique is now also frequently used in metabarcoding and metagenomic studies as well.

3.4 Ion Semiconductor Sequencing (Ion Torrent)

The Personal Genome Machine (PGM) released by Ion Torrent in 2010 was a totally new approach based on ion semiconductor sequencing (Rothberg et al. 2011). After Ion Torrent was purchased by Life Technologies, a platform with even higher throughput was

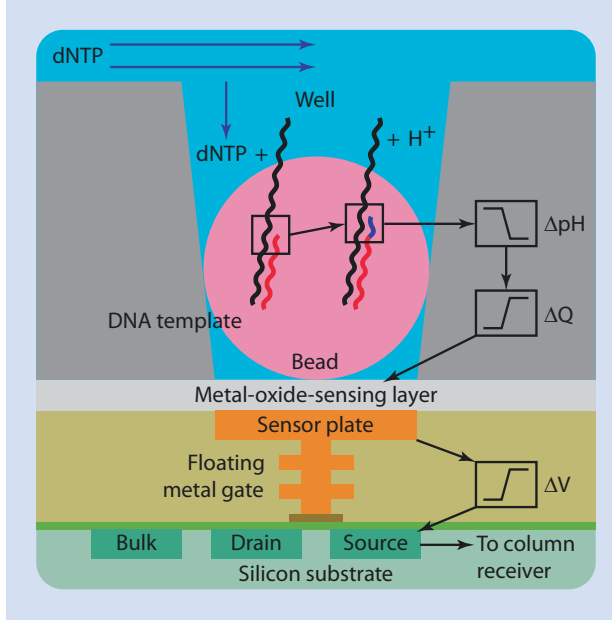


Fig. 3.4 Principle of ion semiconductor sequencing. A well containing a bead with a DNA fragment is shown. Incorporation of a nucleotide releases a proton (H^+), which changes the pH in the well. This release changes the potential in an underlying metal-oxide sensing layer, which is received by a transistor (Reprinted by permission from Macmillan Publishers Ltd: *Nature* (Rothberg et al. 2011), copyright 2011)

released: the Ion Proton. In difference to Sanger, Illumina and 454, this technique does not rely on analysing optical signals. Instead, changes of hydrogen ion concentration are analysed. Anytime when a nucleotide is incorporated into a DNA strand by polymerase activity, a hydrogen (or proton) is released (Fig. 3.4). The release of this proton can be measured in real time by ion-sensitive field-effect transistors (ISFETs) (Sakurai and Husimi 1992). By using available methodology and software from modern imaging devices (laptops, digital cameras), an array has been built for the large-scale use of ISFETs. This technique is called complementary metal-oxide semiconductor (CMOS) process (Rothberg et al. 2011). Every sensor in this array directly monitors the hydrogen ion release during sequencing. Each chip of the sequencer contains between 1 and 660 million sensors, which are composed of a well with an acrylamide bead with a DNA template containing also the dNTPs. The chip size can be chosen according to the required number of reads for the sequencing project. As in 454 sequencing, the wells are flooded in cycles with one sort dNTPs at a time. Below the well lies a metal-oxide sensing layer, which itself is on top of a sensor plate and floating metal «gate» for the transmission of electronic information about the pH changes to the semiconductor (Fig. 3.4). The detected changes in pH allow inferring if and how many bases have been incorporated to a sequence read. Relying on a purely electronic detection system without any optical components allows a considerably cheap instrument cost compared to other NGS platforms.

The library preparation is similar to the 454 technique. Adaptors are ligated to DNA molecules, which are loaded onto magnetic beads. Molecules are amplified using emulsion PCRs. Wells are suited to fit one bead, which are loaded on to the chips by a centrifugation step.

Two different sequence platforms are available (PGM and Ion Proton), and these can be run with differently sized chips. Chips for the PGM bear less sensors (PGM 314: 1.2 million; PGM 316: 6.1 million; PGM 318: 11 million), with an expected output of 500,000 to 5.5 million sequences. With a read length of up to 400 bp, a single run using the largest chip would

generate an output of ~2 Gb. The runtime is according to the chip size with 2 to 7 h relatively fast compared with other techniques. The chips of the Ion Proton system are larger sized (PI, 165 million sensors; PII, 660 million). Using the largest size ~330 million reads with up to 200 bp can be generated in a run which lasts between 2 and 4 h. This would equal an output of 66.6 Gb. With ~98%, the accuracy is lower than for 454 and Illumina platforms (Quail et al. 2012; Merriman et al. 2012). Similar to the 454 technique, indels are the prevailing error type. However, the biggest advantage for this technique is speed. The library preparation should last less than 6 h, and sequencing runs are finished in a few hours. Using this approach, it is possible to get bacterial genomes completed from extracted DNA to assembly in less than 3 days. This rapid methodology has been proven useful while monitoring and characterizing bacterial genomes during an *E. coli* outbreak in Germany in spring 2011 (Mellmann et al. 2011).

3.5 Single-Molecule Real-Time (SMRT) Sequencing (PacBio)

The so far discussed sequencing techniques produce rather sort sequence reads, mostly below 1000 bps. Machines based on a new sequencing method targeting single molecules are available since 2011 from the company Pacific Biosciences (PacBio), which are able to produce considerably longer reads. Here, polymerase activity is monitored without interruption while incorporating four differently fluorescent-tagged dNTPs (Eid et al. 2009). These phospholinked nucleotides carry a fluorescent label on the phosphate group of the nucleotide and are cleaved away after incorporation. By using real-time imaging, incorporated nucleotides are detected while they are synthesized along a single DNA template molecule. The detection takes place in a zero-mode waveguide (ZMW) microwell which is a nanophotonic structure surrounded by aluminium. Each ZMW measures only 70 nm in diameter and 100 nm in depth, leaving an observation volume of 20×10^{-21} litres. A single molecule of Φ29 DNA polymerase is attached to the surface of the ZMW, and its activity can be measured (Fig. 3.5). The small volume of the ZMWs reduces the amount of background noise due to the presence of fluorescent-labelled nucleotides. While detecting the level of fluorescence intensity in a single ZMW, a more or less stable background level is measured. An association of a phospho-linked nucleotide with the template DNA in the polymerase active site triggers a pulse of fluorescence intensity for the corresponding dye. This light emission lasts some milliseconds, which is recorded by the detector of the ZMW. The fluorescence label is cleaved by the DNA polymerase leaving a phosphodiester bond which allows the elongation of the DNA template. The cleaved dye diffuses, leading to a drop of the recorded emission intensity back to the background level. The next nucleotide can be incorporated and the measurement repeats (Fig. 3.5). The synthesis rate is around two to four bases per second. In contrast to all other methods described so far, SMRT sequencing does not interrupt the process of DNA synthesis. Interestingly, it has been shown that the emission spectra contain more information besides the nucleotide identity. The duration of an emission and the interval between successive emissions also reveal information about nucleotide modifications. Using this data, a genome-wide mapping of methylation patterns becomes possible (Flusberg et al. 2010).

High-quality DNA is needed in a high quantity for the library preparation, as no additional amplification step is included. Genomic DNA is sheared to the desired average DNA length, ends are repaired and hairpin adaptors are ligated to these ends. Hairpin adaptors represent single-stranded loops to which the sequencing primer can bind. The construct of the double-stranded template DNA flanked by two hairpin loops is called SMRTbell (Travers et al. 2010). Using polymerase with strand displacement activity, a primer binding to the

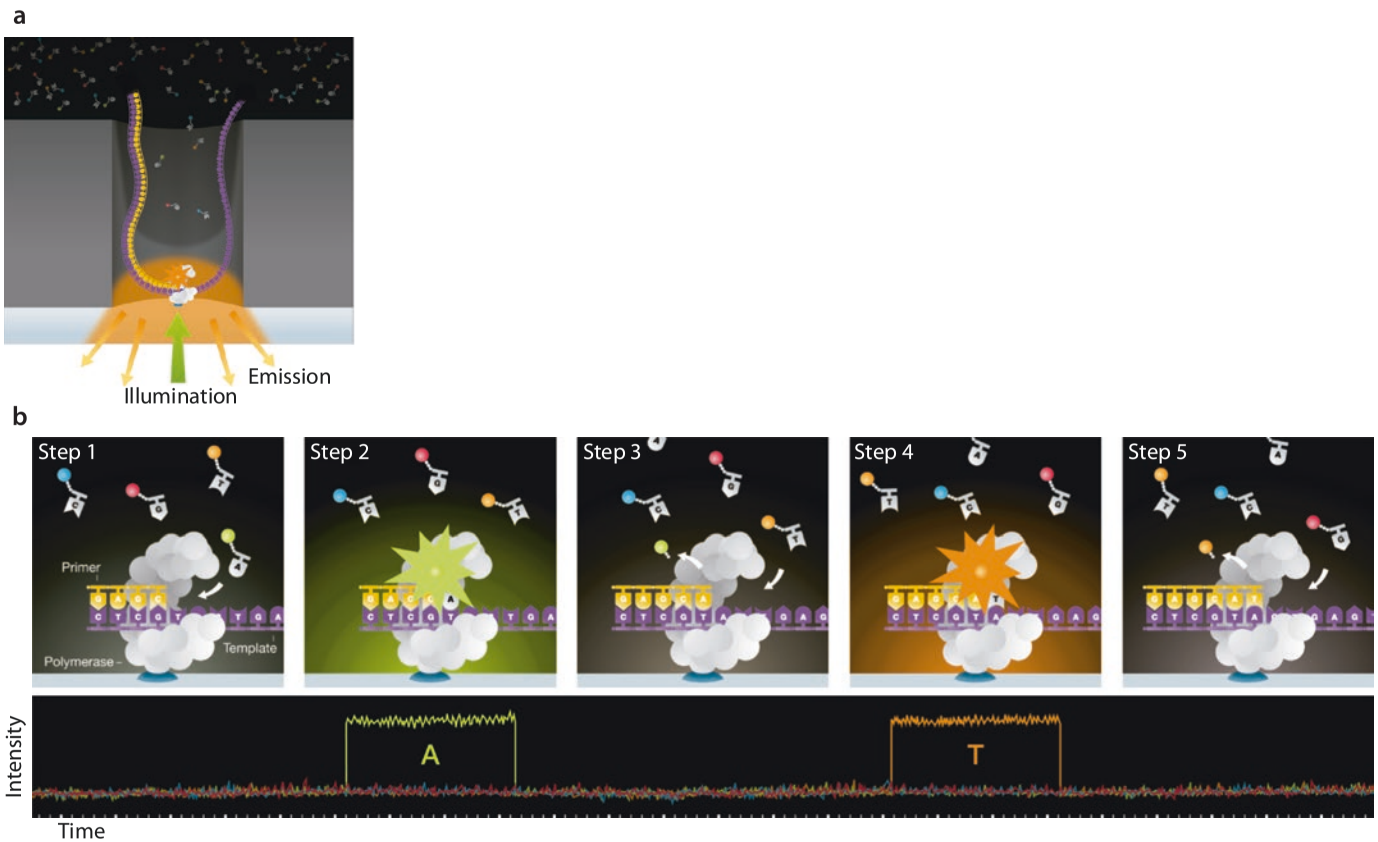


Fig. 3.5 Principle of single-molecule real-time (SMRT) sequencing. **a** A single molecule of Φ29 DNA polymerase is attached to the surface of a zero-mode waveguide (ZMW) microwell, and its activity while copying a DNA molecule is measured. **b** Emission spectra are detected, while fluorescent-labelled nucleotides are incorporated by the polymerase (Reprinted by permission of Pacific Biosciences)

hairpin adaptor can be extended displacing one DNA strand, while the other is used as a template. Sequencing has been even facilitated without any library preparation (Coupland et al. 2012). Whereas there seems to be no negative effect on the read length, the output was considerably lower as for standard library preparation. As no ligated adaptors are present, known primer regions or random hexamer primers can be used for sequencing.

SMRT sequencing has the advantage that the sequencing process is with 4 h rather fast. The read length is long, averaging around 15 Kb, and reads may exceed lengths of 50 Kb and more (Lee et al. 2014). Moreover, single molecules are sequenced, and modifications are detected as additional information. However, a caveat of the technique is the high error rate. An accuracy of ~80–85% is given for single pass reads (Hackl et al. 2014). Even though a large fraction of the sequencing errors seem to stem from deletions and insertions, no significant sequencing bias has been found (Ross et al. 2013). The output is with 2.8 Gb sequencing data (PacBio RSII) per day rather low compared with other NGS techniques. Nevertheless, the availability of long reads from this technique dramatically increases the quality of genome and transcriptome assemblies (Koren and Phillippy 2015; Tilgner et al. 2014). Currently, long reads via SMRT sequencing represent the gold standard for the de novo assembly of genomes, as a more complete picture of gene content, structural variation and repeat biology can be achieved (Gordon et al. 2016).

3.6 Nanopore Sequencing

The principle of DNA (and RNA) sequencing using nanopores was firstly proven back in the mid-1990s (Kasianowicz et al. 1996). For the first sequencing experiments, a staphylococcal nanopore α-hemolysin protein pore was incorporated into a phospholipid bilayer separated by two reservoirs with a salt solution. By applying an electric current using electrodes placed on the opposite sides of the bilayer, negatively charged DNA molecules are forced passing through a small nanopore channel with a diameter of a few nm. Nucleotides passing the pore characteristically decrease the amplitude of the ionic current and can be detected. Using this system, even methylated cytosines can be distinguished from the four standard DNA bases (Clarke et al. 2009; Branton et al. 2008).

The company Oxford Nanopore Technologies (ONT) constructed a series of sequencing devices based on this technique, which are available (MinION) or currently entering the market (PromethION, SmidgeION). The biggest problem of the technique is the immense speed by which the DNA strand is processed through the nanopore. This leads to a decrease of resolution when detecting nucleotides in the channel of the pore. Currently ONT is developing two different systems for DNA sequencing: strand exonuclease and strand sequencing (Clarke et al. 2009), of which only for the latter sequencing data is available while writing this chapter.

In strand sequencing, double-stranded DNA is ratcheted through the nanopore by a «motor protein», a process by which it becomes single stranded. For library preparation, sheared DNA is end-repaired, and a hairpin adaptor is ligated to one end of the molecule, while the motor protein is ligated to the other (Goodwin et al. 2015). During sequencing, one strand is passing the pore, followed by the hairpin adaptor and the other strand. If both strands are sequenced, consensus sequences of the two complementary strands are produced, which are termed 2D reads. Due to the speed of this process, multiple bases are present in the pore at a time, and a collection of overlapping *k*-mers (usually 5-mers, so consecutive five nucleotide fragments) are recorded as signal. Base calling, which is

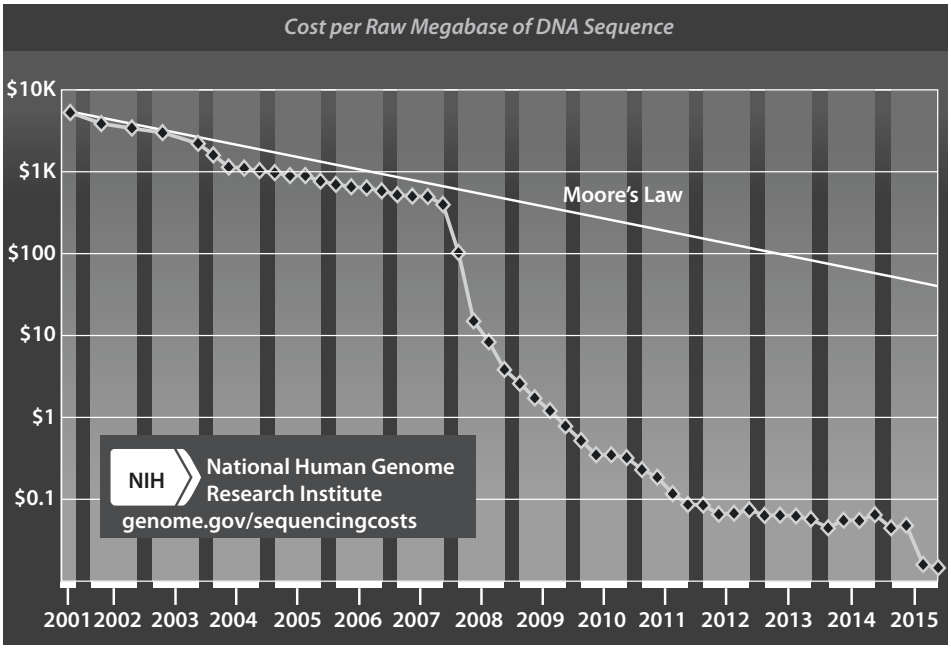
Fig. 3.6 The MinION sequencing device from Oxford Nanopore Technologies (Picture reprinted with permission of Oxford Nanopore Technologies)

conducted by using the software MinKNOW as a cloud application, needs to distinguish 4^5 (1024) possible ionic current states for all possible 5-mers. Not surprisingly, a high error rate is reported for all reads produced by this technique so far, averaging around 12% (Ip et al. 2015). However, several strategies for error correction are available (Loman et al. 2015; Goodwin et al. 2015), and also development of new sequencing chemistries already improved the quality of reads dramatically. The development of this technique progresses so fast that numbers mentioned in this chapter will likely be already outdated when printed. ONT developed several sequencing platforms using this technique; however, while writing this chapter in winter 2016, only data for the MinION nanopore was available. An early release of this device to selected laboratories was announced in 2013 (MinION access program) and facilitated early 2014. Since 2015, the MinION miniature DNA sequencer in the size of an MP3 player (Fig. 3.6) is commercially available. A starter pack including the sequencing device, two flow cells and a library preparation kit can be purchased for $1000.

The MinION is equipped with 512 channels with four nanopores each, each of them detecting 50 to 250 bp per second depending on run mode and chemistry. With the R7 chemistry, an output ranging from 90 to 490 mbp per 48 h is reported, with average read lengths around 6 kbp and maximum read lengths of up to 150 kbp (Ashton et al. 2015; Quick et al. 2014; Goodwin et al. 2015). Initial numbers for the currently distributed R9 chemistry are higher (Istace et al. 2016). Using MinION long reads, it was possible to assemble a complete *Escherichia coli* genome and, in a hybrid assembly with Illumina short reads, the yeast genome (Goodwin et al. 2015; Loman et al. 2015). With the genome of the nematode *Caenorhabditis elegans*, the first animal genome has been sequenced using MinION nanopore long reads only (Tyson et al. 2017). A system with higher output (PromethION) is currently delivered to selected laboratories by ONT. Moreover, further developments exploring other biological or focussing on synthetic nanopores are under investigation (Feng et al. 2015; Wang et al. 2014). If the still rather high error rate can be decreased in future updates, nanopore sequencing might challenge PacBio's status as the gold standard for whole-genome sequencing. Due to the speed and the quite simple library preparation, real-time monitoring in metagenomic frameworks and sequencing in the field are possible. For example, Ebola virus surveillance in the field using nanopore sequencing during an outbreak in Western Africa has been demonstrated (Quick et al. 2016). Moreover, methods like «real-time selective sequencing» will truly help to exploit the power of real-time sequencing (Loose et al. 2016). Currently, we just see the potential of this technique unravelling.

3.7 Comparison of Sequencing Platforms

A broad array of different sequencing techniques became commercially distributed in the last decade, revolutionizing the field of evolutionary genomics. Besides the techniques discussed here, some other platforms are available (e.g. SOLiD, Helicos) (Bowers et al. 2009; Valouev et al. 2008), which are less used in phylogenomic studies and have or will have the greatest potential of applications in clinical studies screening nucleotide polymorphisms, ChIP-seq (chromatin immunoprecipitation DNA sequencing) or resequencing genomes. In 2015, the Beijing Genomics Institute released its sequencing platform called BGISEQ-500, which comes close to the output of Illumina's HiSeq platforms (Goodwin et al. 2016). Several new approaches for DNA sequencing are under development, which are still years away from being commercially available, e.g. transmission electron microscopy DNA sequencing (Bell et al. 2012). The impact of the new sequencing techniques and how it transformed the field of genomics can be most easily seen in the dramatic decrease sequencing costs. Starting with the year 2000, the price per raw megabase of DNA sequencing decreased the first 7 years of this century in line with Moore's law (Moore 1965). This basically means that the number of sequence data to be generated by a fixed price should double exponentially approximately every 2 years (Mardis 2008). Starting in 2007, with the arrival of newly available sequence platforms (454, Illumina), the costs per raw megabase dramatically decreased (◘ Fig. 3.7), basically allowing small laboratories the access to genome and transcriptome sequencing. The development of these new techniques made the $1000 human genome became reality (► see Infobox 3.1).



◘ **Fig. 3.7** Decrease in sequencing costs per raw megabase of DNA sequence over the last 15 years (Reprinted from: Wetterstrand KA. DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP) Available at: ► www.genome.gov/sequencingcostsdata. Accessed January 2017)

Infobox 3.1

The $1000 Genome

Deciphering the human genome took an international collaboration more than a decade, and the costs were estimated to be around 3 billion US dollars when announced in 2001. It became clear that sequencing costs have to be reduced dramatically if subsequent studies targeting haplotype diversity across humans and the use of genomic analysis for routine medical applications should be realized. Discussions about this topic in an expert round delivered the catch phrase that the «$1000 genome» should be targeted. When announced in the early 2000s, this claim seemed utopic and was firstly revised to target the «$100,000 genome», which still would be a more than 100-fold decrease in sequencing costs. It is important to keep in mind that a 100-fold decrease in costs would allow a 100-fold increase in data for the same price – adding much needed statistical power for many desired studies targeting the genetic background of diseases. To achieve these goals, the US-based National Human Genome Research Institute (NHGRI) launched programs actively funding sequencing technology developments. Several sequencing centres and university start-up companies greatly benefited from this rich source of funding. With 454 pyrosequencing, the first of the many NGS techniques to follow became available while directly being supported by these programs. Using this technique, the complete genome sequence of James Watson – the scientist who was directly involved in the discovery of the DNA double helix – was sequenced in less than 2 months for well under 1 million US dollar (Wolinsky 2007; Wheeler et al. 2008). The advent and development of Illumina sequencing again strongly decreased the costs. Already in 2013, costs around $5000 for a 30× coverage human genome with Illumina short reads were estimated by the NHGRI. The launch of Illumina's HiSeq ×10 system, which is an array of ten sequencers with massive output, finally achieved the goal to sequence human genomes for less than $1000 in early 2014. In January 2017, Illumina announced that with the NovaSeq sequencing platform, it might be possible in the near future to sequence a human genome for $100. However, there is no real consensus how to calculate these costs, as to the pure costs for sequencing, additional costs for personal, electricity and analysis should be added. Whereas genome sequencing became extremely cheap, costs for analysing all these data remain high as highly trained scientists have to do this step. Or as Elaine Mardis phrased it famously, «The $1000 genome, the $100,000 analysis?» (Mardis 2010).

Sanger sequencing, which is still the gold standard in terms of read quality, is now also known as the first generation of sequencing. Second-generation sequencing platforms are 454, Illumina and Ion Torrent, all of them massively parallelized for high-throughput data generation, but restricted to short-read lengths. The newly launched nanopore sequencers and PacBio's SMRT sequencing are the third generation of sequencers, which have less output than second-generation machines, but are capable of single-molecule sequencing which in parallel also allows the detection of epigenetic modifications. While the read lengths of these machines are much higher than for all other available sequencing techniques, they are still error prone. However, the development of refined sequencing chemistry and technical updates of sequencing machines stipulate hope for higher-quality data in the near future.

With the availability of this number of different sequencing platforms varying in costs, quality and output (Fig. 3.8), it becomes more difficult to strategically decide which technique should be used in planning phylogenomic studies and, if possible, which sequencers should be purchased by laboratories working in the field of evolution. By far, the highest output is generated by Illumina's HiSeq platforms. The acquisition of a HiSeq is expensive, and these machines can be usually only fully exploited by sequencing centres or very large laboratories. The same is true for PacBio systems. Illumina's MiSeq and Ion Torrent's PGM are affordable for smaller labs. However, the price per base cost for these machines is usually much higher than Illumina's HiSeq (Fig. 3.8). Nevertheless, these machines are well suited for targeted sequencing strategies or sequencing of complete

Platform	Roche 454 FLX plus	Illumina MiSeq	Illukina Next Seq 500	Illumina HiSeq 2500 RR	Illumina HiSeq 4000	Illumina HiSeq X
Reads: (M)	1,25	25	400	600	5000	6000
Read length: (paired-end*)	700	300*	150*	100*	150*	150*
Run time: (d)	0,9	2	1,2	1,125	3,5	3
Yield: (Gb)	0,7	15	120	120	1500	1800
Rate: (Gb/d)	0,75	7,5	100	106,6	400	600
Reagents: ($K)	6,2	1	4,41	6,145	29,9	12,75
per-Gb: ($)	8K	93	36,75	51,2	20	7
hg-30x: ($)	--	11160	4410	6144	2400	840
Machine: ($)	500K	99K	250K	740K	900K	1M
Platform	Ion Torrent 318 HiQ 520	Ion Proton P1	PacBio RS P6-C4	PacBio Sequel	MinIoN MK1	PromethION
Reads: (M)	3-5	165	5,5	38,5	0,05	--
Read length: (paired-end*)	200 400	200	15K	12K	10K	10K
Run time: (d)	0,37	--	4,3	4,3	2	--
Yield: (Gb)	1.2-2	10	12	84	2,75	3100
Rate: (Gb/d)	5,5	--	2,8	19,5	1,375	--
Reagents:($K)	0,6	1	2,4	11,2	0,5	2,5
per-Gb: ($)	--	100	200	80	180	20
hg-30x: ($)	--	12000	24000	9600	21800	2400
Machine: ($)	50K	243K	695K	350K	1K	75K

Fig. 3.8 Output (in bp), runtime (in days) and costs (in $) of different sequencing platforms discussed in this chapter. The cost of resequencing a human genome with 30× coverage is indicated (hg-30×) (Reprinted with permission of Albert Vilella, ► http://twitter.com/albertvilella)

prokaryote genomes. Genome project dealing with eukaryote genomes should envisage a hybrid strategy combining a high coverage of short-read data from second-generation sequencers (e.g. Illumina) and low coverage of long reads from the third generation (e.g. PacBio, Nanopore). Assemblies solely based on long read might result in the highest quality, but could still be too expensive for most projects.

References

Ashton PM, Nair S, Dallman T, Rubino S, Rabsch W, Mwaigwisya S, Wain J, O'Grady J (2015) MinION nanopore sequencing identifies the position and structure of a bacterial antibiotic resistance island. Nat Biotechnol 33:296–300

Bell DC, Thomas WK, Murtagh KM, Dionne CA, Graham AC, Anderson JE, Glover WR (2012) DNA base identification by electron microscopy. Microsc Microanal 18:1049–1053

Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, Hall KP, Evers DJ, Barnes CL, Bignell HR, Boutell JM, Bryant J, Carter RJ, Keira Cheetham R, Cox AJ, Ellis DJ, Flatbush MR, Gormley NA, Humphray SJ, Irving LJ, Karbelashvili MS, Kirk SM, Li H, Liu X, Maisinger KS, Murray LJ, Obradovic B, Ost T, Parkinson ML, Pratt MR, Rasolonjatovo IMJ, Reed MT, Rigatti R, Rodighiero C, Ross MT, Sabot A, Sankar SV, Scally A, Schroth GP, Smith ME, Smith VP, Spiridou A, Torrance PE, Tzonev SS, Vermaas EH, Walter K, Wu X, Zhang L, Alam MD, Anastasi C, Aniebo IC, Bailey DMD, Bancarz IR, Banerjee S, Barbour SG, Baybayan PA, Benoit VA, Benson KF, Bevis C, Black PJ, Boodhun A, Brennan JS, Bridgham JA, Brown RC, Brown AA, Buermann DH, Bundu AA, Burrows JC, Carter NP, Castillo N, Chiara E, Catenazzi M, Chang S, Neil Cooley R, Crake NR, Dada OO, Diakoumakos KD, Dominguez-Fernandez B, Earnshaw DJ,

Egbujor UC, Elmore DW, Etchin SS, Ewan MR, Fedurco M, Fraser LJ, Fuentes Fajardo KV, Scott Furey W, George D, Gietzen KJ, Goddard CP, Golda GS, Granieri PA, Green DE, Gustafson DL, Hansen NF, Harnish K, Haudenschild CD, Heyer NI, Hims MM, Ho JT, Horgan AM, Hoschler K, Hurwitz S, Ivanov DV, Johnson MQ, James T, Huw Jones TA, Kang G-D, Kerelska TH, Kersey AD, Khrebtukova I, Kindwall AP, Kingsbury Z, Kokko-Gonzales PI, Kumar A, Laurent MA, Lawley CT, Lee SE, Lee X, Liao AK, Loch JA, Lok M, Luo S, Mammen RM, Martin JW, PG MC, McNitt P, Mehta P, Moon KW, Mullens JW, Newington T, Ning Z, Ling Ng B, Novo SM, O'Neill MJ, Osborne MA, Osnowski A, Ostadan O, Paraschos LL, Pickering L, Pike AC, Pike AC, Chris Pinkard D, Pliskin DP, Podhasky J, Quijano VJ, Raczy C, Rae VH, Rawlings SR, Chiva Rodriguez A, Roe PM, Rogers J, Rogert Bacigalupo MC, Romanov N, Romieu A, Roth RK, Rourke NJ, Ruediger ST, Rusman E, Sanches-Kuiper RM, Schenker MR, Seoane JM, Shaw RJ, Shiver MK, Short SW, Sizto NL, Sluis JP, Smith MA, Ernest Sohna Sohna J, Spence EJ, Stevens K, Sutton N, Szajkowski L, Tregidgo CL, Turcatti G, Vandevondele S, Verhovsky Y, Virk SM, Wakelin S, Walcott GC, Wang J, Worsley GJ, Yan J, Yau L, Zuerlein M, Rogers J, Mullikin JC, Hurles ME, NJ MC, West JS, Oaks FL, Lundberg PL, Klenerman D, Durbin R, Smith AJ (2008) Accurate whole human genome sequencing using reversible terminator chemistry. Nature 456:53–59

Bowers J, Mitchell J, Beer E, Buzby PR, Causey M, Efcavitch JW, Jarosz M, Krzymanska-Olejnik E, Kung L, Lipson D, Lowman GM, Marappan S, McInerney P, Platt A, Roy A, Siddiqi SM, Steinmann K, Thompson JF (2009) Virtual terminator nucleotides for next-generation DNA sequencing. Nat Methods 6:593–595

Branton D, Deamer DW, Marziali A, Bayley H, Benner SA, Butler T, Di Ventra M, Garaj S, Hibbs A, Huang X, Jovanovich SB, Krstic PS, Lindsay S, Ling XS, Mastrangelo CH, Meller A, Oliver JS, Pershin YV, Ramsey JM, Riehn R, Soni GV, Tabard-Cossa V, Wanunu M, Wiggin M, Schloss JA (2008) The potential and challenges of nanopore sequencing. Nat Biotechnol 26:1146–1153

Brenner S (2014) Frederick Sanger (1918–2013). Sci 343:262

Clarke J, Wu H-C, Jayasinghe L, Patel A, Reid S, Bayley H (2009) Continuous base identification for single-molecule nanopore DNA sequencing. Nat Nanotechnol 4:265–270

Coupland P, Chandra T, Quail M, Reik W, Swerdlow H (2012) Direct sequencing of small genomes on the Pacific Biosciences RS without library preparation. BioTechniques 53:365–372

Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, Peluso P, Rank D, Baybayan P, Bettman B, Bibillo A, Bjornson K, Chaudhuri B, Christians F, Cicero R, Clark S, Dalal R, deWinter A, Dixon J, Foquet M, Gaertner A, Hardenbol P, Heiner C, Hester K, Holden D, Kearns G, Kong X, Kuse R, Lacroix Y, Lin S, Lundquist P, Ma C, Marks P, Maxham M, Murphy D, Park I, Pham T, Phillips M, Roy J, Sebra R, Shen G, Sorenson J, Tomaney A, Travers K, Trulson M, Vieceli J, Wegener J, Wu D, Yang A, Zaccarin D, Zhao P, Zhong F, Korlach J, Turner S (2009) Real-time DNA sequencing from single polymerase molecules. Sci 323:133–138

Feng Y, Zhang Y, Ying C, Wang D, Du C (2015) Nanopore-based fourth-generation DNA sequencing technology. Genomics Proteomics Bioinformatics 13:4–16

Flusberg BA, Webster DR, Lee JH, Travers KJ, Olivares EC, Clark TA, Korlach J, Turner SW (2010) Direct detection of DNA methylation during single-molecule, real-time sequencing. Nat Methods 7:461–465

Gilles A, Meglecz E, Pech N, Ferreira S, Malausa T, Martin J-F (2011) Accuracy and quality assessment of 454 GS-FLX titanium pyrosequencing. BMC Genomics 12:245

Goodwin S, Gurtowski J, Ethe-Sayers S, Deshpande P, Schatz MC, McCombie WR (2015) Oxford nanopore sequencing, hybrid error correction, and de novo assembly of a eukaryotic genome. Genome Res 25:1750–1756

Goodwin S, McPherson JD, McCombie WR (2016) Coming of age: ten years of next-generation sequencing technologies. Nat Rev Genet 17:333–351

Gordon D, Huddleston J, Chaisson MJP, Hill CM, Kronenberg ZN, Munson KM, Malig M, Raja A, Fiddes I, Hillier LW, Dunn C, Baker C, Armstrong J, Diekhans M, Paten B, Shendure J, Wilson RK, Haussler D, Chin C-S, Eichler EE (2016) Long-read sequence assembly of the gorilla genome. Science 352:aae0344

Hackl T, Hedrich R, Schultz J, Förster F (2014) Proovread: large-scale high-accuracy PacBio correction through iterative short read consensus. Bioinformatics 30:3004–3011

Huse S, Huber J, Morrison H, Sogin M, Welch D (2007) Accuracy and quality of massively parallel DNA pyrosequencing. Genome Biol 8:R143

Ip C, Loose M, Tyson J, de Cesare M, Brown B, Jain M, Leggett R, Eccles D, Zalunin V, Urban J, Piazza P, Bowden R, Paten B, Mwaigwisya S, Batty E, Simpson J, Snutch T, Birney E, Buck D, Goodwin S, Jansen H, O'Grady J, Olsen H, null n (2015) MinION Analysis and Reference Consortium: Phase 1 data release and analysis [version 1; referees: 2 approved]. F1000Res 4:1075

Istace B, Friedrich A, d'Agata L, Faye S, Payen E, Beluche O, Caradec C, Davidas S, Cruaud C, Liti G, Lemainque A, Engelen S, Wincker P, Schacherer J, Aury J-M (2016) de novo assembly and population genomic survey of natural yeast isolates with the Oxford Nanopore MinION sequencer. bioRxiv. doi.org/10.1101/066613

Kasianowicz JJ, Brandin E, Branton D, Deamer DW (1996) Characterization of individual polynucleotide molecules using a membrane channel. Proc Natl Acad Sci U S A 93:13770–13773

Kircher M, Kelso J (2010) High-throughput DNA sequencing – concepts and limitations. BioEssays 32:524–536

Koren S, Phillippy AM (2015) One chromosome, one contig: complete microbial genomes from long-read sequencing and assembly. Curr Opin Microbiol 23:110–120

Lee H, Gurtowski J, Yoo S, Marcus S, McCombie WR, Schatz M (2014) Error correction and assembly complexity of single molecule sequencing reads. bioRxiv:006395

Liu L, Li Y, Li S, Hu N, He Y, Pong R, Lin D, Lu L, Law M (2012) Comparison of next-generation sequencing systems. J Biomed Biotechnol 2012:251364

Loman NJ, Quick J, Simpson JT (2015) A complete bacterial genome assembled de novo using only nanopore sequencing data. Nat Methods 12:733–735

Loose M, Malla S, Stout M (2016) Real-time selective sequencing using nanopore technology. Nat Methods 13:751–754

Mardis ER (2008) The impact of next-generation sequencing technology on genetics. Trends Genet 24:133–141

Mardis E (2010) The $1,000 genome, the $100,000 analysis? Genome Med 2:84

Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen Y-J, Chen Z, Dewell SB, Du L, Fierro JM, Gomes XV, Godwin BC, He W, Helgesen S, Ho CH, Irzyk GP, Jando SC, Alenquer MLI, Jarvie TP, Jirage KB, Kim J-B, Knight JR, Lanza JR, Leamon JH, Lefkowitz SM, Lei M, Li J, Lohman KL, Lu H, Makhijani VB, McDade KE, McKenna MP, Myers EW, Nickerson E, Nobile JR, Plant R, Puc BP, Ronan MT, Roth GT, Sarkis GJ, Simons JF, Simpson JW, Srinivasan M, Tartaro KR, Tomasz A, Vogt KA, Volkmer GA, Wang SH, Wang Y, Weiner MP, Yu P, Begley RF, Rothberg JM (2005) Genome sequencing in microfabricated high-density picolitre reactors. Nature 437:376–380

Maxam AM, Gilbert W (1977) A new method for sequencing DNA. Proc Natl Acad Sci U S A 74:560–564

Mellmann A, Harmsen D, Cummings CA, Zentz EB, Leopold SR, Rico A, Prior K, Szczepanowski R, Ji Y, Zhang W, McLaughlin SF, Henkhaus JK, Leopold B, Bielaszewska M, Prager R, Brzoska PM, Moore RL, Guenther S, Rothberg JM, Karch H (2011) Prospective genomic characterization of the German enterohemorrhagic *Escherichia coli* O104:H4 outbreak by rapid next generation sequencing technology. PLoS One 6:e22751

Merriman B, D Team IT, Rothberg JM (2012) Progress in ion torrent semiconductor chip based sequencing. Electrophoresis 33:3397–3417

Metzker ML (2010) Sequencing technologies - the next generation. Nat Rev Genet 11:31–46

Meyer M, Kircher M (2010) Illumina sequencing library preparation for highly multiplexed target capture and sequencing. Cold Spring Harb Protoc 2010:pdb.prot5448

Moore G (1965) Cramming more components onto integrated circuits. Electrodiagn Ther 38:114–117

Petrosino JF, Highlander S, Luna RA, Gibbs RA, Versalovic J (2009) Metagenomic pyrosequencing and microbial identification. Clin Chem 55:856–866

Prober J, Trainor G, Dam R, Hobbs F, Robertson C, Zagursky R, Cocuzza A, Jensen M, Baumeister K (1987) A system for rapid DNA sequencing with fluorescent chain-terminating dideoxynucleotides. Science 238:336–341

Quail M, Smith M, Coupland P, Otto T, Harris S, Connor T, Bertoni A, Swerdlow H, Gu Y (2012) A tale of three next generation sequencing platforms: comparison of ion torrent, Pacific biosciences and Illumina MiSeq sequencers. BMC Genomics 13:341

Quick J, Quinlan A, Loman N (2014) A reference bacterial genome dataset generated on the MinION portable single-molecule nanopore sequencer. Gigascience 3:22

Quick J, Loman NJ, Duraffour S, Simpson JT, Severi E, Cowley L, Bore JA, Koundouno R, Dudas G, Mikhail A, Ouédraogo N, Afrough B, Bah A, Baum JHJ, Becker-Ziaja B, Boettcher JP, Cabeza-Cabrerizo M, Camino-Sánchez Á, Carter LL, Doerrbecker J, Enkirch T, Dorival IG, Hetzelt N, Hinzmann J, Holm T, Kafetzopoulou LE, Koropogui M, Kosgey A, Kuisma E, Logue CH, Mazzarelli A, Meisel S, Mertens M, Michel J, Ngabo D, Nitzsche K, Pallasch E, Patrono LV, Portmann J, Repits JG, Rickett NY, Sachse A, Singethan K, Vitoriano I, Yemanaberhan RL, Zekeng EG, Racine T, Bello A, Sall AA, Faye O, Faye O, Magassouba NF, Williams CV,

Amburgey V, Winona L, Davis E, Gerlach J, Washington F, Monteil V, Jourdain M, Bererd M, Camara A, Somlare H, Camara A, Gerard M, Bado G, Baillet B, Delaune D, Nebie KY, Diarra A, Savane Y, Pallawo RB, Gutierrez GJ, Milhano N, Roger I, Williams CJ, Yattara F, Lewandowski K, Taylor J, Rachwal P, Turner DJ, Pollakis G, Hiscox JA, Matthews DA, MKO S, Johnston AM, Wilson D, Hutley E, Smit E, Di Caro A, Wölfel R, Stoecker K, Fleischmann E, Gabriel M, Weller SA, Koivogui L, Diallo B, Keïta S, Rambaut A, Formenty P, Günther S, Carroll MW (2016) Real-time, portable genome sequencing for Ebola surveillance. Nature 530:228–232

Ronaghi M (2001) Pyrosequencing sheds light on DNA sequencing. Genome Res 11:3–11

Ronaghi M, Karamohamed S, Pettersson B, Uhlén M, Nyrén P (1996) Real-time DNA sequencing using detection of pyrophosphate release. Anal Biochem 242:84–89

Ross M, Russ C, Costello M, Hollinger A, Lennon N, Hegarty R, Nusbaum C, Jaffe D (2013) Characterizing and measuring bias in sequence data. Genome Biol 14:R51

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

Rothberg JM, Hinz W, Rearick TM, Schultz J, Mileski W, Davey M, Leamon JH, Johnson K, Milgrew MJ, Edwards M, Hoon J, Simons JF, Marran D, Myers JW, Davidson JF, Branting A, Nobile JR, Puc BP, Light D, Clark TA, Huber M, Branciforte JT, Stoner IB, Cawley SE, Lyons M, Fu Y, Homer N, Sedova M, Miao X, Reed B, Sabina J, Feierstein E, Schorn M, Alanjary M, Dimalanta E, Dressman D, Kasinskas R, Sokolsky T, Fidanza JA, Namsaraev E, McKernan KJ, Williams A, Roth GT, Bustillo J (2011) An integrated semiconductor device enabling non-optical genome sequencing. Nature 475:348–352

Sakurai T, Husimi Y (1992) Real-time monitoring of DNA polymerase reactions by a micro ISFET pH sensor. Anal Chem 64:1996–1997

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

Tilgner H, Grubert F, Sharon D, Snyder MP (2014) Defining a personal, allele-specific, and single-molecule long-read transcriptome. Proc Natl Acad Sci U S A 111:9869–9874

Travers KJ, Chin C-S, Rank DR, Eid JS, Turner SW (2010) A flexible and efficient template format for circular consensus sequencing and SNP detection. Nucleic Acids Res 38:e159

Tyson JR, O'Neil NJ, Jain M, Olsen HE, Hieter P, Snutch TP (2017) Whole genome sequencing and assembly of a *Caenorhabditis elegans* genome with complex genomic rearrangements using the MinION sequencing device. bioRxiv. doi.org/10.1101/099143

Valouev A, Ichikawa J, Tonthat T, Stuart J, Ranade S, Peckham H, Zeng K, Malek JA, Costa G, McKernan K, Sidow A, Fire A, Johnson SM (2008) A high-resolution, nucleosome position map of *C. elegans* reveals a lack of universal sequence-dictated positioning. Genome Res 18:1051–1063

Wang Y, Yang Q, Wang Z (2014) The evolution of nanopore sequencing. Front Genet 5:449

Wheeler DA, Srinivasan M, Egholm M, Shen Y, Chen L, McGuire A, He W, Chen Y-J, Makhijani V, Roth GT, Gomes X, Tartaro K, Niazi F, Turcotte CL, Irzyk GP, Lupski JR, Chinault C, X-z S, Liu Y, Yuan Y, Nazareth L, Qin X, Muzny DM, Margulies M, Weinstock GM, Gibbs RA, Rothberg JM (2008) The complete genome of an individual by massively parallel DNA sequencing. Nature 452:872–876

Wolinsky H (2007) The thousand-dollar genome. EMBO Rep 8:900–903

Yu DW, Ji Y, Emerson BC, Wang X, Ye C, Yang C, Ding Z (2012) Biodiversity soup: metabarcoding of arthropods for rapid biodiversity assessment and biomonitoring. Methods Ecol Evol 3:613–623

Sequencing Strategies

C. Bleidorn, *Phylogenomics*, DOI 10.1007/978-3-319-54064-1_4

- Shotgun strategies help sequencing whole genomes in small fragments which are assembled into longer contigs afterwards.
- RADseq strategies provide a reduced but consistent set of sequences of the genome which are especially used for population genetics.
- Hybrid enrichment describes the specific enhancement of preselected sequences.
- RNA-Seq analyses characterize the sequence content and according expression level of transcriptomes.
- Technical developments paved the way to sequence genomes and transcriptome of single cells.

4.1 Shotgun Sequencing

The length of prokaryote and eukaryote genomes exceeds by far the length of sequence reads produced by available technologies. Moreover, in the case of eukaryotes, the genomic information is distributed across a number of chromosomes. Therefore, different strategies have been developed for complete genome sequencing. Many of these methods have been explored in the course of the human genome project, e.g. transposon-based methods to integrate random insertions into cloned DNA or multiplex PCR strategies (Green 2001; Church and Kieffer-Higgins 1988). However, the most common method is shotgun sequencing, which was developed in the early 1980s (Anderson 1981; Gardner et al. 1981). For shotgun sequencing, a large stretch of DNA is fragmented into smaller pieces. In the next step, random pieces of the fragmented DNA are sequenced to generate redundant amounts of sequence data. Finally, individual sequence reads are assembled to reconstruct the sequence of the analysed genome (Green 2001). Two different strategies using shotgun sequencing have been used in genome-sequencing projects (Fig. 4.1): (I) hierarchical shotgun sequencing and (II) whole-genome shotgun sequencing.

For hierarchical shotgun sequencing (Fig. 4.1a), large fragments of DNA are cloned using bacterial artificial chromosomes (BACs). BACs are cloning vectors derived from *Escherichia coli* plasmids and have the advantage that the insertion of relatively large DNA fragments (>100–300 Kb) is possible (Shizuya et al. 1992). Alternatively, other cloning systems have been used, but less frequent than BACs. In a second step, a physical map of the cloned DNA is established. Various physical mapping approaches have been developed, including BAC restriction-based fingerprinting (Marra et al. 1997), iterative hybridization (Mozo et al. 1999), and the use of BAC-end sequences for connecting BAC clones by sequence identity (Mahairas et al. 1999). Restriction-based fingerprinting methods digest BAC clones by using a set of restriction enzymes (e.g. two enzymes in case of double digest), thereby generating a set of different sized fragments which can be visualized using gel electrophoresis. For each BAC, a unique pattern of bands on a gel is derived, and the presence and absence of fragment sizes can be scored. Finally, all BACs are ordered in relative position according to their similarity regarding shared fragment sizes (Soderlund et al. 1997). Based on this information, a minimal set of overlapping BACs which is in total completely covering a selected genomic region (minimal tiling path) is selected. For individual sequencing of BACs, their inserted DNA is purified and physically shared to generate smaller fragments for sequencing. For Sanger sequencing, broken ends of the sheared fragments are enzymatically repaired, and all fragments are size fractionized using gel

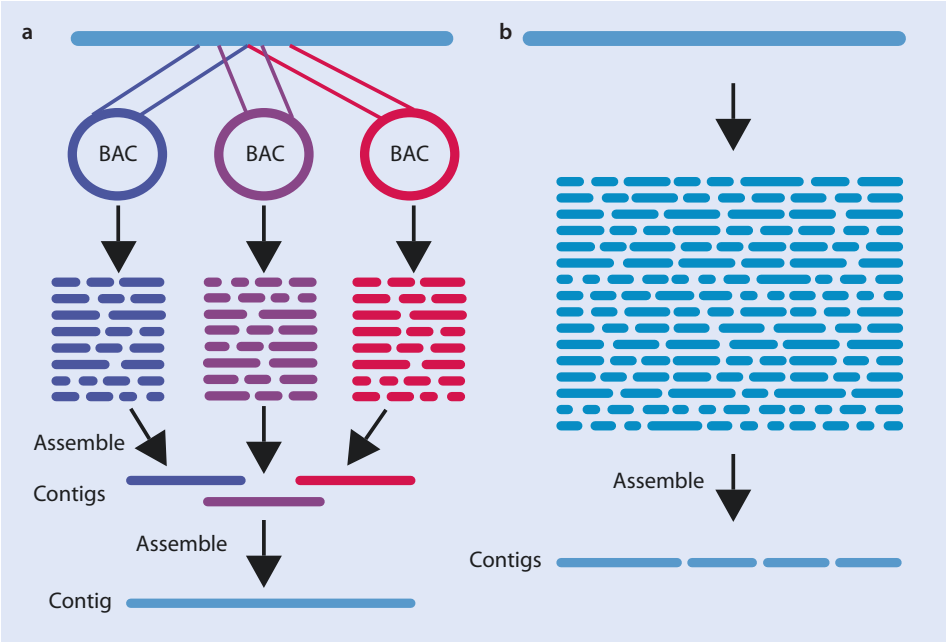


Fig. 4.1 Overview of shotgun-sequencing methods. **a** For hierarchical shotgun sequencing, large fragments of the original chromosome are cloned into BAC clones. BAC clones with overlapping fragments are chosen according to physical mapping information and fragmented into small fragments. BAC clone fragments are sequenced and assembled for each clone separately. Assembled contigs will be overlapped according mapping information to the final contig. **b** For whole-genome shotgun sequencing, chromosomes will be directly fragmented, without mapping information. Fragments will be sequenced and reads will be assembled into contigs

electrophoresis. Medium-sized (2–3 Kb) fragments are selected and cloned into sequencing vectors, which can be finally sequenced using conserved primer sites in the vector. A random collection of sequences of ~10x coverage is generated for each BAC, which can be used for BAC contig assembly. Contigs for all BACs of the minimal tiling path are overlapped according the information from the physical mapping to generate the final sequence, which should represent the sequenced genomic region. The first available larger eukaryote genomes, e.g. *Arabidopsis thaliana* (The Arabidopsis Genome Initiative 2000) and *Caenorhabditis elegans* (The *C. elegans* Sequencing Consortium 1998), have been sequenced with this approach. The International Human Genome Sequencing Consortium (2001) used hierarchical shotgun sequencing for the human genome.

Whole-genome shotgun (wgs) sequencing (Fig. 4.1b) directly involves sequencing of sheared genomic DNA, thereby leaving out the time-consuming step of establishing a physical map (Green 2001). In case of using Sanger sequencing, sheared DNA is end repaired, subcloned into sequencing vectors and sequenced in a high coverage. Assembly of this kind of sequence data usually leads to less continuous contigs, as the topological information from physical mapping is missing. Initially, this approach was mainly used for (small) bacterial genomes. Weber and Myers (1997) used simulations to demonstrate the practicability of wgs for sequencing large eukaryote genomes. Most famously, this was validated in practice by Craig Venter and colleagues by sequencing and assembling the human genome using wgs data (Venter et al. 2001).

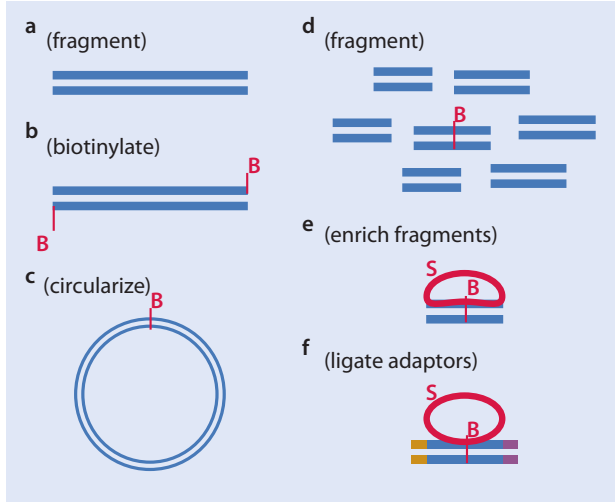


Fig. 4.2 Construction of mate-pair libraries (Illumina). **a** DNA is sheared into fragments. **b** DNA fragments are end repaired and biotinylated. **c** Biotinylated fragments are circularized. **d** Circularized DNA molecules are sheared into ~500 bp fragments. **e** Fragments containing biotinylated ends are selected using streptavidin-covered magnetic beads; remaining fragments are washed away. **f** Adaptors for sequencing are ligated to selected fragments

Next-generation sequencing (NGS) techniques dramatically increased the output of sequencing reads, and wgs approaches became a standard. However, the most powerful methods in terms of sequence reads output (Illumina, Ion Torrent) are also the methods producing the shortest reads (100–250 bp). Especially the assembly of eukaryote genomes, which are often rich in repetitive sequences, became a major challenge. One strategy to provide extra information for assembling wgs data is the use of mate-pair sequencing. Mate pairs describe the sequenced ends of DNA fragments separated by a specific size. For example, if the ends of a 3 Kb fragments are sequenced, the topological information that these sequences should be separated by roughly this size can be used to improve assemblies. Mate pair libraries have been developed for all major short-read sequencing techniques (Illumina, 454, Ion Torrent), and even though details may vary, the principle remains the same. Most frequently, mate-pair sequencing is conducted with Illumina, and therefore details are explained for this method.

In the first step, genomic DNA is sheared into fragments of the desired size (Fig. 4.2a). Typical sizes for mate pair libraries range from 2 to 5 Kb, even though larger libraries (5 to 25 Kb) are also feasible (van Heesch et al. 2013). DNA fragments are end repaired and the 3′-ends are labelled with biotin (Fig. 4.2b). The B-vitamin biotin is widely used in molecular biology and can be covalently attached to proteins or nucleic acids. Biotin binds with high specificity and very fast to streptavidin. Magnetic beads covered with this protein can be used to specifically enrich biotinylated molecules. The size of prepared fragments can be selected using agarose gel electrophoresis, and size information is essential for subsequent computational analysis. Biotinylated fragments are circularized by intramolecular ligation (Fig. 4.2c), and remaining linear molecules are enzymatically removed. The circularized DNA molecules are sheared again into a size of ~500 bp (Fig. 4.2d). The fragments containing the biotinylated ends are selected using streptavidin-covered magnetic beads (Fig. 4.2e), and remaining fragments are washed away. The selected fragments contain the 3′-ends of the original DNA fragments. Finally, sequencing adaptors are attached to the selected fragments to prepare the sequencing library (Fig. 4.2f). Sequencing of these fragments generates read pairs which align towards the ends of the original size-selected fragment and are outward facing from each other. The gap between these reads is approximately of the size of the original fragment, and this information is valuable for contig assembly and scaffolding of genomes (Chaisson et al. 2009).

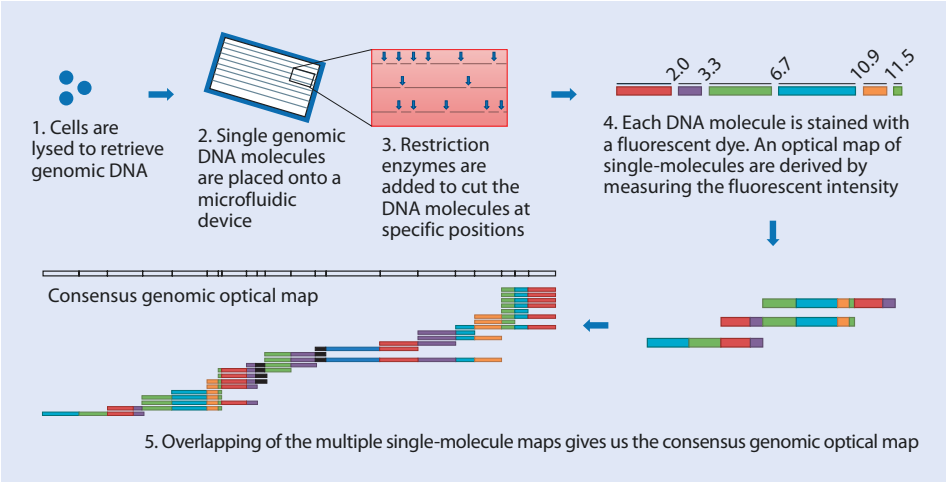


Fig. 4.3 A workflow for optical mapping (By Fong Chun Chan and Kendric Wang (Own work) [CC BY 3.0 (▶ http://creativecommons.org/licenses/by/3.0)], via Wikimedia Commons)

Mapping strategies have been developed to improve and validate wgs assemblies, e.g. optical mapping (Schwarz et al. 2014). This method is similar to the restriction-based fingerprinting approach described above. For optical mapping, large DNA molecules are immobilized on a surface and digested with one or more restriction enzymes (Fig. 4.3). The digested DNA molecules are stained with a fluorescent dye. The length between adjoining cut sites is estimated by measuring the fluorescence intensity. Mapping data of each single DNA molecule is used to produce a consensus genomic optical map, which includes an ordered series of DNA fragment sizes (Mendelowitz and Pop 2014). Recently, a high-throughput method of optical mapping using nanochannels has been proposed (Lam et al. 2012). With this approach, DNA fragments are nicked by an enzyme at specific sequence sites and subsequently fluorescently labelled. With the help of an electric field, molecules are driven through a nanoscale channel, where the DNA is stretched. In this channel, distances between fluorescent labels can be measured using a microscope. A unique optical pattern resembling a barcode is created by the distance measure of the labels (Michaeli and Ebenstein 2012).

A mapping strategy which became recently popular has been commercialized by Dovetail Genomics and is based on a Hi-C approach (Lieberman-Aiden et al. 2009). The idea behind Hi-C is that, after fixation of chromatin structure, DNA segments which are in close proximity in the nucleus are more likely to be ligated together. This is reflected by the finding that the number of intra-chromosomal ligation pairs decreases while the genomic distance between them increases. With the so-called cHiCago protocol, Hi-C mapping is used for the localization of chromatin interactions to infer the relative order and orientation of contigs (Putnam et al. 2016). Using this protocol, chromatin is reconstituted in vitro and fixed with formaldehyde. The fixed chromatin is then cut with a restriction enzyme, thereby generating free sticky ends, which are filled with biotinylated and thiolated nucleotides. In the next step, free blunt ends are ligated, and chromatin crosslinks to generate ligation mate pairs, which are fusions of fragments which are distantly located in the genome. After library preparation, these fragments can be sequenced with NGS methods. The mapping of these fragments helps to dramatically improve

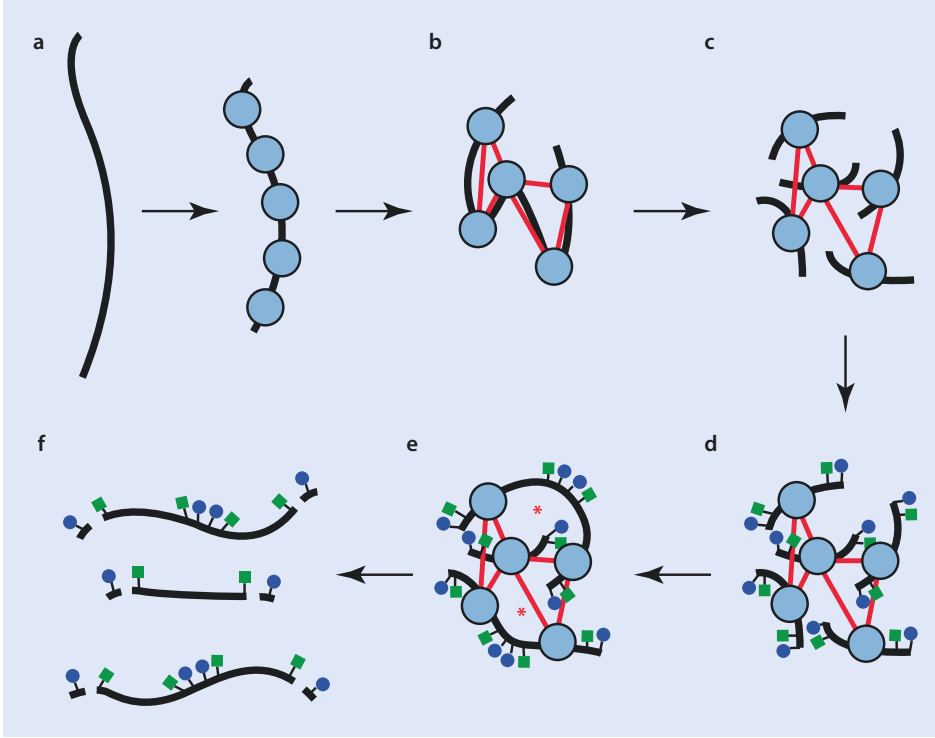


Fig. 4.4 Diagram of the cHiCago library preparation protocol as used by Dovetail Genomics. **a** Chromatin (nucleosomes in blue) is reconstituted in vitro upon naked DNA (black strand). **b** Fixation of chromatin by formaldehyde. *Red lines* indicate crosslinks. **c** Cutting of fixed chromatin using restriction enzymes. **d** Filling of sticky ends with biotinylated (*blue circles*) and thiolated (*green squares*) nucleotides. **e** Ligation of free blunt ends (*red asterisks*). **f** Fragments for library preparation are yielded by reversion of crosslinks and removal of proteins. Terminal biotinylated nucleotides are removed (Reprinted from Putnam et al. (2016))

genome assemblies based on various NGS techniques (e.g. Illumina, PacBio). For example, by using the cHiCago protocol, the scaffold N50 of the Illumina-based genome assembly of the American alligator could be increased from 508 Kb to 10 Mb (Putnam et al. 2016) (Fig. 4.4).

A different way to improve wgs assemblies is by using long sequencing reads. This can be directly done by sequencing with third-generation techniques such as single-molecule real-time sequencing or nanopore sequencing. Alternatively, long reads can also be generated synthetically for Illumina short-read sequencing. Illumina itself distributes a technique called TruSeq, which was formerly known under the name Moleculo. With this approach, ~10 Kb DNA fragments are amplified and barcoded before sequencing, and long reads can be created afterwards based on this information. The company 10X Genomics released an instrument called Chromium which used a similar but more powerful approach for the generation of synthetic long reads. Up to 100 Kb long DNA fragments are amplified and barcoded with an emulsion PCR step. Subsequently, these fragments are sequenced in a very low coverage, and sequenced barcodes localize clouds of short reads which are used to scaffold de novo assemblies (Lee et al. 2016). The advantage of both these methods is their considerably lower price compared to true long-read

sequencing. However, synthetically generated «long» reads are prone to biases of the Illumina technology, e.g. less or no coverage in regions with high GC content. Also, tandem repeats are still difficult to tackle with this approach.

4.2 RADseq

Due to the advent of NGS techniques, genome sequencing became feasible and affordable even for non-model organisms and also smaller laboratories. However, for many studies, it is sufficient to analyse a snapshot of the genome, but for a high number of individuals. A set of related methods used to sequence a reduced, but consistent representation of the genome is known as restriction site-associated DNA sequencing (RADseq). Applications of RADseq include discovery of genetic markers for phylogenetics and population genetics (Cruaud et al. 2014; Davey et al. 2011), mapping of quantitative trait loci (QTLs) (Houston et al. 2012), linking mapping (Gonen et al. 2014) or local genome assembly (Etter et al. 2011). The name RADseq was introduced for one specific approach of reduced representation sequencing (Baird et al. 2008), but is now used to describe several similar methods (Andrews et al. 2016). Besides the original RADseq approach, this family includes methods like ddRAD (Peterson et al. 2012), ezRAD (Toonen et al. 2013), 2bRAD (Wang et al. 2012), and the widely used genotyping by sequencing (GBS) (Elshire et al. 2011).

The original RADseq protocol starts with the digestion of genomic DNA with one restriction enzyme (Fig. 4.5a). Restriction enzymes are able to cleave DNA in either random (type I) or specific positions (type II). The first restriction enzyme cutting specific sequence motive (*HindII*) was isolated from the bacterium *Haemophilus influenzae* (Smith and Welcox 1970). Since that time, several thousand restriction enzymes (targeting different sequence motives) have been described, and hundreds are commercially available. A list of available restriction enzymes and their properties are collected in the database REBASE (Roberts et al. 2015). The choice of the restriction enzyme greatly influences in how many pieces the genome is cut. By a rule of thumb, the longer the recognized sequence motive, the less fragments are generated. For example, a six-base pair motive as recognized by the *EcoR1* enzyme (Fig. 4.5a) will cut every 4,000 bp, whereas an eight-base pair motive would only cut every 65,500 bp (Andrews et al. 2016). These numbers are rough estimates and are greatly influenced by the base composition of the investigated genome. Restriction enzymes can either cut symmetrically, thereby generating blunt ends, or asymmetrically. By using an asymmetrical cutting enzyme, all fragments will bear so-called sticky ends, which describe the overhang created by cutting with the restriction enzyme. An adaptor can be ligated to these sticky ends, which includes a known primer site for PCR amplification (Fig. 4.5b). If adaptors bearing unique barcode sequences are used, multiple libraries can be mixed at this point (multiplexing). This barcode will be read during sequencing and allows the separation of multiplexed samples. The complete DNA library will be sheared, followed by reparation of sequence ends. Using blunt-end ligation, a second adaptor is ligated to all fragments (Fig. 4.5c). This second adaptor is Y-shaped, containing an only partially overlapping sequence. The resulting DNA library will be amplified using a primer pair (e.g. P1 and P2) (Fig. 4.5d). One sequencing primer site is nested in the first adaptor (P1). The second primer site is identical to one of the nonoverlapping sequence parts of the y-shaped adaptors (P2). The y-shaped adaptor is completed when fragments containing the first adaptor are bound by P1 and copied. Primer P2 only binds to the Y-shaped adaptor after completion. Thereby, specificity of the

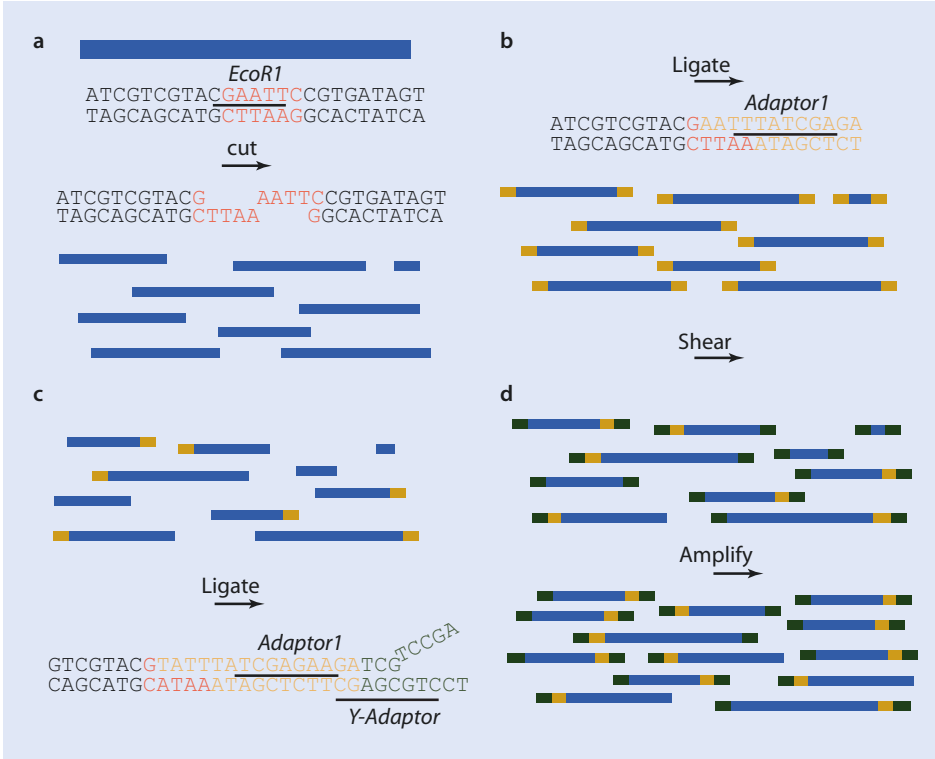


Fig. 4.5 Workflow of the original RADseq protocol. **a** Genomic DNA is cut with a chosen restriction enzyme (in this example *EcoR1*) for fragmentation. **b** Using the overhang as created by the restriction enzyme, an adaptor is ligated to sequence fragments. The complete pool of DNA is sheared mechanically. **c** Y-shaped adaptors are ligated to the sheared pool of DNA fragments. **d** Using priming sites in both adaptors, the DNA library is amplified. Only fragments containing both adaptors can be successfully amplified

amplification is enhanced, as only fragments containing both adaptors are amplified (Fig. 4.5d). The enhanced library can be sequenced using NGS. With this method, thousands of single nucleotide polymorphic (SNP) loci can be generated (Davey et al. 2011).

Several variants of the original RADseq protocol have been developed (see above), which differ in details of restriction enzyme digestion, size selection or adaptor ligation (Andrews et al. 2016). Commonly used alternative protocols are ddRAD and GBS. In the case of double digest RADseq (ddRAD), two different restriction enzymes are utilized to digest the genomic DNA (Peterson et al. 2012). Adaptors are ligated to each cut site, and size selection is facilitated by choosing those fragments, which are flanked by restriction enzyme recognition sites that are neither too close or too distant (Fig. 4.6). Using this method, all reads of a given locus share the same fragment size, as no shearing step is involved. Moreover, size selection further decreases the number of analysed loci, which in turn increases the coverage in terms of sequence reads. In contrast, in the case of RADseq (see above), each sequenced fragment has a cut site at one end and a randomly sheared end at the other. Thereby a range of fragment sizes is produced for each locus (Andrews et al. 2016).

GBS is basically a simplified protocol of the RADseq approaches described above. DNA is digested with one restriction enzyme, and a pair of adaptors is ligated to each fragment. One adaptor contains a barcode unique for each library (e.g. for single individuals); the

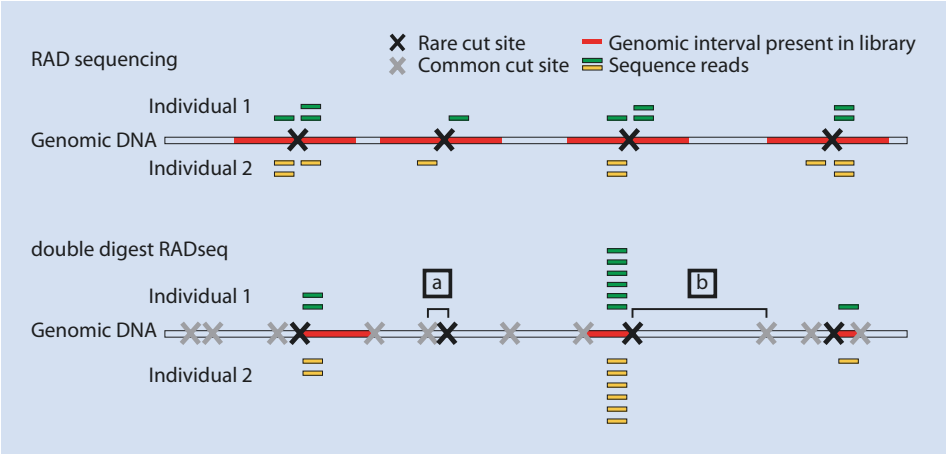


Fig. 4.6 Comparison of analysed loci by RADseq **a** and ddRADseq **b**. In the case of ddRADseq, **b** size selection excludes regions flanked by either [*a*] very close or [*b*] very distant restriction enzyme recognition sites (Figure from Peterson et al. (2012))

other adaptor is a common adopter used in all libraries (Elshire et al. 2011). Subsequently, all libraries are pooled and a PCR is performed with primer sites nesting in the ligated adaptors. The pooled and amplified library can be sequenced using NGS. Modifications of this simple protocol, using two restriction enzymes and y-shaped adaptors, have been published (Poland et al. 2012). GBS approaches have been especially widely used for SNP discovery in large plant genomes (Deschamps et al. 2012), but also population genomic analyses (Friis et al. 2016).

The number of loci identified by RADseq methods is influenced by the frequency of cut sites of the chosen restriction enzymes, size selection (if applied), genome size of the target organism and chosen RADseq method. If a reference genome is available, in silico analyses can be performed to optimize RADseq experiments (Lepais and Weir 2014). Such analyses are used to predict the number of retrieved loci given the choice of restriction enzyme or based on alternative methods. Even though in many cases there are no reference genomes available, genome-wide surveys of frequencies of restriction enzyme recognition sequences show a high variability across eukaryotic taxonomic groups (Herrera et al. 2015). The frequency of this cleavage sites seems to be similar among closely related species, which helps to choose enzymes for RADseq experiments with organisms lacking a reference genome. Moreover, as RADseq methods differ in costs and hands-on time in the lab, these factors further influence the numbers of samples which can be analysed. Pooling samples without using individually barcoded adaptors are a cost-efficient alternative, but may prohibit some downstream population genetic analyses (Futschik and Schlötterer 2010; Andrews et al. 2014).

Advantages and disadvantages of different RADseq methods have been discussed in detail (Puritz et al. 2014; Andrews et al. 2014; Andrews et al. 2016). Several biases due to methodological artefacts may influence the analysis of RADseq data in general. A common problem is the introduction of PCR duplicates. These duplicates do not represent independent samples from the analysed genomic DNA pool. As independence of samples is an underlying assumption of most population genetic analyses, this may result in skewing allele frequencies, genotyping errors or false-positive alleles (Andrews et al. 2014). Putative PCR duplicates can be identified when using RADseq methods that include a

random-shearing step, as in the original RADseq protocol (see above). By analysing paired-end sequence reads, PCR duplicates can be identified as fragments that are identical across forward and reverse reads (Davey et al. 2011). Additional sources of bias introduced during PCR are preferential amplification of loci based on GC content and fragment size, which may impact the variance of sequence read coverage across loci (Puritz et al. 2014). Critical for all RADseq methods are problems due to non-random sampling leading to systematic underestimation of polymorphisms (Arnold et al. 2013; Huang and Knowles 2014). Non-random sampling results from polymorphic recognition sequences of the used restriction enzymes, resulting in missing data for some chromosomes/individuals (allelic dropout).

4.3 Hybrid Enrichment

Hybrid enrichment methods are used for the specific capture and enrichment of selected sequences (Lemmon and Lemmon 2013). In short, capture probes (DNA or RNA) that are complementary to targeted regions in the genome are hybridized to a DNA library, and target DNA is enriched by washing away nontargeted DNA prior to high-throughput sequencing. This method has been used to enrich selected single-copy orthologous loci for phylogenetic analyses, as in anchored hybrid enrichment (AHE) (Lemmon et al. 2012) or enrichment of ultraconserved elements (UCE) (Faircloth et al. 2012). Moreover, it is widely used for the enrichment of exonic DNA (Li et al. 2013) or organelle DNA (Briggs et al. 2009). Prior to the enrichment, long oligonucleotides (usually ~60–120 bp) which cover the target regions have to be designed and synthesized. For this purpose, genomic or transcriptomic resources of the target species or closely related species are used as a reference. In the case of AHE, and when targeting UCEs, it has been shown that capture probes could even be successfully designed for vertebrates across multiple evolutionary timescales, in some cases spanning divergence times of ~500 million years (Lemmon et al. 2012; Faircloth et al. 2012). Capture probes can be designed for several hundred to thousands of loci in parallel, which may involve several thousand oligonucleotides. Most target enrichment applications follow a solution-based enrichment protocol (sometimes with modifications) as developed by Gnirke et al. (2009) (Fig. 4.7). Designed oligonucleotides are synthesized on a microarray (Lipshutz et al. 1999), cleaved and eluted. After initial PCR, a T7 promoter sequence is added to the double-stranded DNA. This promoter can be used to transcribe DNA to RNA with the help of T7 RNA polymerase. This polymerase is promoter specific in only transcribing double-stranded DNA downstream of a T7 promoter sequence (Studier and Moffatt 1986). The transcription takes place under the presence of biotin-UTPs, thereby generating biotinylated single-stranded RNA capture baits (Fig. 4.7a). Meanwhile, genomic DNA of the target organism is sheared, end repaired, adaptor ligated (grey) and PCR amplified (Fig. 4.7b). Capture of targets will take place in solution. For this purpose, strands of genomic DNA are separated and hybridized with the prepared biotinylated RNA baits (Fig. 4.7c). After hybridization, target DNA (and unbound probes) can be captured using magnetic streptavidin-coated beads (Fig. 4.7c). Unbound DNA is washed away, whereas captured and thereby

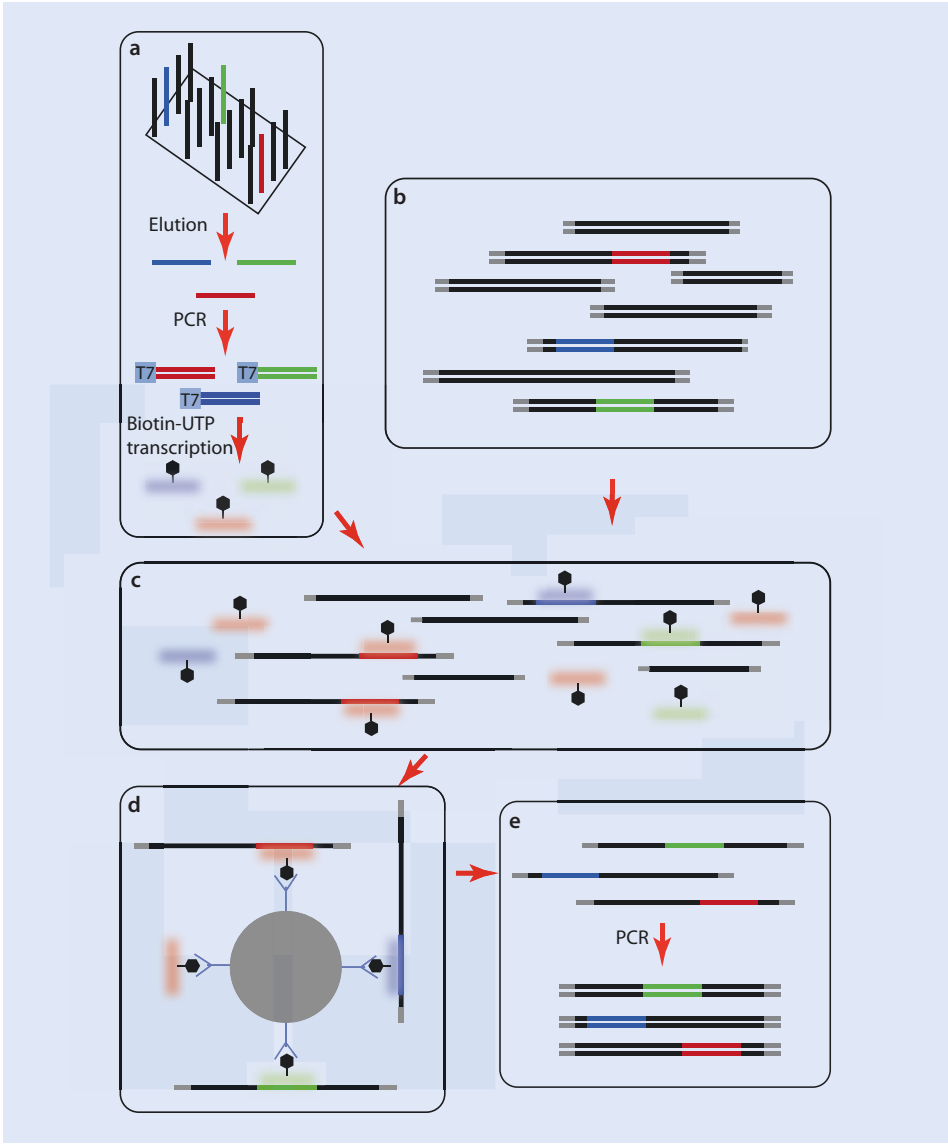


Fig. 4.7 Principle of solution hybrid selection. *Colours* represent differently targeted DNA regions. *Black diamonds* represent biotin label. **a** Long oligonucleotides are synthesized on a microarray, cleaved and eluted. After initial PCR, a T7 promoter is added to double-stranded DNA. In the presence of biotin-UTP, biotinylated single-stranded RNA baits are generated (*milky lines with black diamonds*). **b** Genomic DNA of the target organism is sheared, end repaired, adaptor ligated (*grey*) and PCR amplified. **c** Strands of genomic DNA are separated and hybridized in solution with biotinylated RNA baits. **d** Free biotinylated RNA baits and those hybridizing to target DNA are captured using streptavidin-coated magnetic beads. **e** Captured DNA fragments are eluted and amplified by PCR

enriched target DNA is eluted, PCR amplified and ready to be sequenced using NGS platforms (Fig. 4.7d).

Especially two approaches became widely used for phylogenomic studies. Anchored hybrid enrichment as introduced by Lemmon et al. (2012) identifies conserved DNA regions flanked by less conserved regions for probe design. Usually alignments of genomically well-characterized model species are exploited to design oligonucleotides. AHE has been mostly used for phylogenetic analyses of different groups of vertebrates (Prum et al. 2015; Eytan et al. 2015; Ruane et al. 2015). Faircloth et al. (2012) targeted UCEs, which have been initially described as perfectly conserved segments of mammalian genomes which are not functionally transcribed (Dermitzakis et al. 2005). Such regions have been also described in other animals, but also plants and fungi (Siepel et al. 2005; Zheng and Zhang 2008). Using UCEs has the advantage that a set of loci can be characterized in highly divergent reference genomes and later applied to a diverse set of taxa, without the need of always designing new probes (Jones and Good 2016). As UCEs are often flanked by variable regions, this method also works across shallow evolutionary timescales as, for example, demonstrated in the phylogenetic analysis of a cichlid radiation (McGee et al. 2016).

Hybridization enrichment strategies have been also successfully used when working with ancient DNA. Often only a very low level of endogenous DNA is preserved in ancient specimens (1–2%), while the majority represents environmental DNA (Carpenter et al. 2013). Moreover, the DNA is normally highly degenerated, and only short and also damaged fragments are present. Consequently, wgs approaches might be not effective and too costly when dealing with ancient DNA. Fu et al. (2013) developed capture probes targeting the complete mitochondrial genome and representative portions from the nuclear genome in ancient humans. It was furthermore possible to sequence complete mitochondrial genomes from the oldest so far investigated ancient humans (> ~300,000 years ago) (Meyer et al. 2014). This method has been also demonstrated to work with highly degraded and ultrashort DNA in non-permafrost-preserved cave bears from the Middle Pleistocene (Dabney et al. 2013). Target capture of mitochondrial genomes in permafrost-preserved horse fossils even allowed the analyses of specimens which dated 560,000 to 780,000 years ago (Orlando et al. 2013).

Alternatively to in solution hybridization methods, capture can take place directly on a microarray (Albert et al. 2007). DNA microarrays have been initially used to study gene expression pattern (Schena et al. 1995), an application which is now more and more supplanted by RNA-Seq (see below). DNA microarrays are a collection of DNA sequences which are attached to a surface (e.g. glass). Specific PCR products or designed oligonucleotides can be printed at specified sites on glass slides using high-precision arraying robots (Schulze and Downward 2001). Complementary DNA can be directly hybridized to DNA microarrays and thereby captured. If this DNA is fluorescently labelled, the intensity of bound DNA can be measured, e.g. to infer the relative expression of mRNA. In the case of hybridization enrichment, genomic DNA is sheared, adaptor ligated, amplified and hybridized with the array (Albert et al. 2007). Non-hybridized DNA is washed away, while the captured (and thereby enriched) DNA fragments are eluted and prepared for subsequent NGS library preparation. Liu et al. (2016) demonstrated the successful enrichment of mitochondrial genomes of insects using such a microarray capture approach.

4.4 Expressed Sequence Tags and RNA-Seq

The transcriptome comprises the complete set of transcripts, as well as their quantity, of a cell or population of cells. Several technologies are available to sequence and quantify the transcriptome, including hybridization-based approaches using microarrays (see above) or direct sequencing (Wang et al. 2009). Using Sanger-based techniques, sequencing of expressed sequence tags (ESTs) was established in the 1990s to characterize transcriptomes (Adams et al. 1991), even though the lack of sequencing power usually did not allow the quantification of gene expression. By harnessing the power of NGS techniques, RNA-Seq became the method of choice to sequence transcriptomes and to determine gene expression levels. In general, for both methods RNA is reverse transcribed to a library of cDNA fragments. The RNA can be total, selected for transcripts carrying a poly-A-tail or depleted in ribosomal RNA. Similarly, specific libraries targeting small RNAs (e.g. tRNAs, microRNAs) can be constructed. For EST sequencing, cDNA is cloned into an appropriate vector, which is sequenced from both ends. Alternatively, directional cloning of cDNA is possible, so that only 5′-ends of the sequences are sequenced, thereby avoiding poly-A-tail sequences. Sequencing takes place with the Sanger technique and usually an amount of a few hundred or thousands transcript ends is manageable. This method played an important role in gene discovery (Schuler 1997) and also paved the way for the first broadscale phylogenomic studies in animals (Dunn et al. 2008). With dbEST, an entire database hosted by NCBI GenBank is dedicated to EST sequences (Boguski et al. 1993).

Transcriptome sequencing by RNA-Seq exploits available NGS high-throughput technologies (Wang et al. 2009). As for EST sequencing, RNA is firstly converted to a cDNA library. The cDNA fragments will then be prepared for NGS methods by attaching adaptors to both ends. The library is finally sequenced in a high-throughput manner to obtain a high coverage of short sequence reads. RNA-Seq can be used for transcriptome assembly, as well as expression profiling at the same time. Especially for non-model organisms, RNA-Seq became the method of choice for de novo transcriptome assembly, gene discovery and gene expression comparisons (Ekblom and Galindo 2011; McCormack et al. 2013; Todd et al. 2016). By using RNA-Seq, hundreds to thousands of putatively orthologous genes can be discovered, and thereby transcriptome-based phylogenomic analyses became state of the art to understand animal evolution (Telford et al. 2015; Dunn et al. 2014). Moreover, RNA-Seq is a powerful tool for gene expression analyses. The expression level of genes is measured by the number of sequenced fragments that map back to each transcript. For RNA-Seq, abundance levels are given as mapped reads per kilobase (RPKM) (Mortazavi et al. 2008). Compared to microarray studies, the RNA-Seq approach offers several advantages (Table 4.1), e.g. identification of gene isoforms and allele-specific expression, nucleotide polymorphisms and post-transcriptional base modifications (Malone and Oliver 2011; Rapaport et al. 2013). Importantly, this approach also enabled comparative gene expression studies for organisms where reference genomes or transcriptomes are missing (Todd et al. 2016). Consequently, RNA-Seq became a powerful approach to study differential gene expression, which aims to investigate qualitative and quantitative differences of genes expressed in different cell types (Gilbert 2013).

As powerful and straightforward the counting of mapped reads appears, several pitfalls have to be avoided when working with RNA-Seq data (Tarazona et al. 2011;

Table 4.1 Comparison of different methods investigating gene expression (partly adopted from Wang et al. (2009))

	Microarray	ESTs	RNA-Seq
Principle	Hybridization	Sanger	NGS (e.g. Illumina)
Resolution	Several to 100 bp	Single base pair	Single base pair
Throughput	High	Low	High
Prior genomic resources	Required	Not required	Not required
Isoform distinction	No	Yes	Yes
Allelic expression	No	Yes	Yes

Vijay et al. 2013). The expression signal of any given transcript is obviously limited by the sequencing depth and is thereby also dependent on the level of expression of other transcripts (Rapaport et al. 2013). Additionally, there is a transcript length bias, as more reads map to long transcripts compared to short transcripts of similar expression (Oshlack and Wakefield 2009). Thereby, the probability to detect the presence as well as differential expression of a given transcript varies strongly. Biological variance in gene expression due to genetic or environmental differences can further complicate RNA-Seq analyses (Todd et al. 2016). And, finally, bias can be introduced by technical differences when comparing different sequencing runs (or even lanes of a single flow cell) or different library preparations (McIntyre et al. 2011). To deal with these problems, gene expression experiments should be designed carefully. For example, increased sequence depth may help to uncover lowly expressed variants and alleviate problems related to transcript length, but at the same time also increases the number of false positives due to sequencing errors. As a rule of thumb, the larger the genome of the analysed species, as more complex is its transcriptome. For «simple» yeast transcriptomes, it was shown that with 30 million short (35 bp) reads the expression of >90% of the expected transcripts could be detected (Wang et al. 2009). For the more «complex» chicken transcriptome, similar numbers (~30 million) of medium-sized reads (75 bp) were enough to detect 90% of all annotated genes, and even with 10 million reads, 80% of the genes could be detected (Wang et al. 2011). By reviewing gene expression studies across diverse sets of eukaryotes, Todd et al. (2016) recommend that efforts in the range of 5 to 20 million mapped reads per sample seem a sufficient sequencing depth. There is also a trade-off in the number of biological replicates to be sequenced and their costs. Such replicates can improve estimates of variance for different sources of bias and are obviously necessary to quantify biological variation. It has been shown that the increase of number of biological replicates has a stronger positive effect on the statistical power of differential gene expression experiments than increasing the sequencing depth for each sample (Liu et al. 2014). Useful guidelines for the design of RNA-Seq experiments in the context of evolutionary and ecological research questions are given by Wolf (2013) and Todd et al. (2016).

4.5 Single-Cell Genomics and Transcriptomics

Single-cell genomics and transcriptomics aim to study genetic diversity on a cellular level (Tang et al. 2011; Shapiro et al. 2013). Using these approaches it is possible to study microbial ecosystems and cell lineage relationships or to connect genotypes with phenotypes on a single-cell level. However, the acquisition of high-quality single-cell sequencing data comes with major technical challenges: (1) physical isolation of individual cells, (2) amplification of the genome (or transcriptome) of single cells for downstream analyses and (3) analysing the data given the biases and errors introduced during the first two steps (Gawad et al. 2016). The isolation of individual cells can be facilitated by methods like serial dilution, microfluids, micromanipulation, laser-capture microdissection or fluorescence-activated cell sorting (FACS) (Yilmaz and Singh 2012). Single cells have to be transferred to reaction tubes for subsequent DNA or RNA extraction. In case of RNA, reverse transcription into cDNA is necessary. Currently, amplification of the DNA (or cDNA) of single cells is required to gain a sufficient amount of molecules for sequencing. However, in the near-future single-molecule sequencing as performed by third-generation sequencing, platforms (PacBio, Oxford Nanopore) should supersede this step. It is possible to sequence the transcriptome and genome of the same cell as demonstrated by Macaulay et al. (2015).

Single-cell genomics has emerged as a powerful tool to recover genomic information from uncultured, individual cells of environmental microorganisms (Stepanauskas 2012). As this method recovers all genomic information of a given cell, chromosomal and extrachromosomal elements are recovered, thereby also detecting possible infections by viruses. For example, Labonte et al. (2015) demonstrated the possibility to investigate host-virus relationships in marine microbial communities. Further on, single-cell genomics helps to link the genotype of so far unculturable prokaryotes with metabolic functions as derived from annotation of their genomes. For example, the investigation of ubiquitous but uncultured Proteobacteria lineages sampled in the dark oxygenated ocean revealed potential chemolithoautotrophy, thereby providing a new perspective on carbon cycling of this large oceanic habitat (Swan et al. 2011).

Single-cell transcriptomic approaches have been successfully implemented for evolutionary developmental research. Lee et al. (2014) developed fluorescent in situ RNA sequencing (FISSEQ), a method where cDNA is directly sequenced within biological samples (tissue sections, whole-mount embryos). Alternatively, Achim et al. (2015) proposed to compare transcriptomes from single-cell sequencing (of cells with unknown spatial locations) with available expression profiles from a gene expression atlas. Using this method >80% of cells could be allocated to precise locations in the brain of the model annelid *Platynereis dumerilii*. Ultimately, these methods will help to resolve the origin, features and fate of different cell types in complex tissues (Satija et al. 2015).

References

Achim K, Pettit J-B, Saraiva LR, Gavriouchkina D, Larsson T, Arendt D, Marioni JC (2015) High-throughput spatial mapping of single-cell RNA-seq data to tissue of origin. Nat Biotechnol 33:503–509

Adams M, Kelley J, Gocayne J, Dubnick M, Polymeropoulos M, Xiao H, Merril C, Wu A, Olde B, Moreno R, Kerlavage A, McCombie WR, Venter JC (1991) Complementary DNA sequencing: expressed sequence tags and human genome project. Science 252:1651–1656

Albert TJ, Molla MN, Muzny DM, Nazareth L, Wheeler D, Song X, Richmond TA, Middle CM, Rodesch MJ, Packard CJ, Weinstock GM, Gibbs RA (2007) Direct selection of human genomic loci by microarray hybridization. Nat Methods 4:903–905

Anderson S (1981) Shotgun DNA sequencing using cloned DNase I-generated fragments. Nucleic Acids Res 9:3015–3027

Andrews KR, Hohenlohe PA, Miller MR, Hand BK, Seeb JE, Luikart G (2014) Trade-offs and utility of alternative RADseq methods: reply to Puritz et al. Mol Ecol 23:5943–5946

Andrews KR, Good JM, Miller MR, Luikart G, Hohenlohe PA (2016) Harnessing the power of RADseq for ecological and evolutionary genomics. Nat Rev Genet 17:81–92

Arnold B, Corbett-Detig RB, Hartl D, Bomblies K (2013) RADseq underestimates diversity and introduces genealogical biases due to nonrandom haplotype sampling. Mol Ecol 22:3179–3190

Baird NA, Etter PD, Atwood TS, Currey MC, Shiver AL, Lewis ZA, Selker EU, Cresko WA, Johnson EA (2008) Rapid SNP discovery and genetic mapping using sequenced RAD markers. PLoS One 3:e3376

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

Briggs AW, Good JM, Green RE, Krause J, Maricic T, Stenzel U, Lalueza-Fox C, Rudan P, Brajković D, Kućan Ž, Gušić I, Schmitz R, Doronichev VB, Golovanova LV, de la Rasilla M, Fortea J, Rosas A, Pääbo S (2009) Targeted retrieval and analysis of five Neandertal mtDNA genomes. Science 325:318–321

Carpenter ML, Buenrostro JD, Valdiosera C, Schroeder H, Allentoft Morten E, Sikora M, Rasmussen M, Gravel S, Guillén S, Nekhrizov G, Leshtakov K, Dimitrova D, Theodossiev N, Pettener D, Luiselli D, Sandoval K, Moreno-Estrada A, Li Y, Wang J, Gilbert MTP, Willerslev E, Greenleaf WJ, Bustamante CD (2013) Pulling out the 1%: whole-genome capture for the targeted enrichment of ancient DNA sequencing libraries. Am J Hum Genet 93:852–864

Chaisson MJ, Brinza D, Pevzner PA (2009) De novo fragment assembly with short mate-paired reads: does the read length matter? Genome Res 19:336–346

Church G, Kieffer-Higgins S (1988) Multiplex DNA sequencing. Science 240:185–188

Cruaud A, Gautier M, Galan M, Foucaud J, Sauné L, Genson G, Dubois E, Nidelet S, Deuve T, Rasplus J-Y (2014) Empirical assessment of RAD sequencing for interspecific phylogeny. Mol Biol Evol 31:1272–1274

Dabney J, Knapp M, Glocke I, Gansauge M-T, Weihmann A, Nickel B, Valdiosera C, García N, Pääbo S, Arsuaga J-L, Meyer M (2013) Complete mitochondrial genome sequence of a Middle Pleistocene cave bear reconstructed from ultrashort DNA fragments. Proc Natl Acad Sci U S A 110:15758–15763

Davey JW, Hohenlohe PA, Etter PD, Boone JQ, Catchen JM, Blaxter ML (2011) Genome-wide genetic marker discovery and genotyping using next-generation sequencing. Nat Rev Genet 12:499–510

Dermitzakis ET, Reymond A, Antonarakis SE (2005) Conserved non-genic sequences – an unexpected feature of mammalian genomes. Nat Rev Genet 6:151–157

Deschamps S, Llaca V, May GD (2012) Genotyping-by-sequencing in plants. Biology 1:460

Dunn CW, Hejnol A, Matus DQ, Pang K, Browne WE, Smith SA, Seaver E, Rouse GW, Obst M, Edgecombe GD, Sorensen MV, Haddock SHD, Schmidt-Rhaesa A, Okusu A, Kristensen RM, Wheeler WC, Martindale MQ, Giribet G (2008) Broad phylogenomic sampling improves resolution of the animal tree of life. Nature 452:745–750

Dunn CW, Giribet G, Edgecombe GD, Hejnol A (2014) Animal phylogeny and its evolutionary implications. Annu Rev Ecol Syst 45:371–395

Ekblom R, Galindo J (2011) Applications of next generation sequencing in molecular ecology of non-model organisms. Heredity 107:1–15

Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, Mitchell SE (2011) A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS One 6:e19379

Etter PD, Preston JL, Bassham S, Cresko WA, Johnson EA (2011) Local de novo assembly of RAD paired-end contigs using short sequencing reads. PLoS One 6:e18561

Eytan RI, Evans BR, Dornburg A, Lemmon AR, Lemmon EM, Wainwright PC, Near TJ (2015) Are 100 enough? Inferring acanthomorph teleost phylogeny using anchored hybrid enrichment. BMC Evol Biol 15:113

Faircloth BC, McCormack JE, Crawford NG, Harvey MG, Brumfield RT, Glenn TC (2012) Ultraconserved elements anchor thousands of genetic markers spanning multiple evolutionary timescales. Syst Biol 61:717–726

Friis G, Aleixandre P, Rodríguez-Estrella R, Navarro-Sigüenza AG, Milá B (2016) Rapid postglacial diversification and long-term stasis within the songbird genus Junco: phylogeographic and phylogenomic evidence. Mol Ecol 24:6175-6195.

Fu Q, Meyer M, Gao X, Stenzel U, Burbano HA, Kelso J, Pääbo S (2013) DNA analysis of an early modern human from Tianyuan Cave, China. Proc Natl Acad Sci U S A 110:2223–2227
Futschik A, Schlötterer C (2010) The next generation of molecular markers from massively parallel sequencing of pooled DNA samples. Genetics 186:207–218
Gardner RC, Howarth AJ, Hahn P, Brown-Luedi M, Shepherd RJ, Messing J (1981) The complete nucleotide sequence of an infectious clone of cauliflower mosaic virus by M13mp7 shotgun sequencing. Nucleic Acids Res 9:2871–2888
Gawad C, Koh W, Quake SR (2016) Single-cell genome sequencing: current state of the science. Nat Rev Genet 17:175–188
Gilbert S (2013) Developmental biology, 10th edn. Sinauer Associates Inc., Sunderland
Gnirke A, Melnikov A, Maguire J, Rogov P, LeProust EM, Brockman W, Fennell T, Giannoukos G, Fisher S, Russ C, Gabriel S, Jaffe DB, Lander ES, Nusbaum C (2009) Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. Nat Biotechnol 27:182–189
Gonen S, Lowe NR, Cezard T, Gharbi K, Bishop SC, Houston RD (2014) Linkage maps of the Atlantic salmon (*Salmo salar*) genome derived from RAD sequencing. BMC Genomics 15:1–17
Green ED (2001) Strategies for the systematic sequencing of complex genomes. Nat Rev Genet 2:573–583
Herrera S, Reyes-Herrera PH, Shank TM (2015) Predicting RAD-seq marker numbers across the eukaryotic tree of life. Genome Biol Evol 7:3207–3225
Houston RD, Davey JW, Bishop SC, Lowe NR, Mota-Velasco JC, Hamilton A, Guy DR, Tinch AE, Thomson ML, Blaxter ML, Gharbi K, Bron JE, Taggart JB (2012) Characterisation of QTL-linked and genome-wide restriction site-associated DNA (RAD) markers in farmed Atlantic salmon. BMC Genomics 13:244
Huang H, Knowles LL (2014) Unforeseen consequences of excluding missing data from next-generation sequences: simulation study of RAD sequences. Syst Biol 65:357–365
International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. Nature 409:860–921
Jones MR, Good JM (2016) Targeted capture in evolutionary and ecological genomics. Mol Ecol 25:185–202
Labonte JM, Swan BK, Poulos B, Luo H, Koren S, Hallam SJ, Sullivan MB, Woyke T, Eric Wommack K, Stepanauskas R (2015) Single-cell genomics-based analysis of virus-host interactions in marine surface bacterioplankton. ISME J 9:2386–2399
Lam ET, Hastie A, Lin C, Ehrlich D, Das SK, Austin MD, Deshpande P, Cao H, Nagarajan N, Xiao M, Kwok P-Y (2012) Genome mapping on nanochannel arrays for structural variation analysis and sequence assembly. Nat Biotechnol 30:771–776
Lee JH, Daugharthy ER, Scheiman J, Kalhor R, Yang JL, Ferrante TC, Terry R, Jeanty SSF, Li C, Amamoto R, Peters DT, Turczyk BM, Marblestone AH, Inverso SA, Bernard A, Mali P, Rios X, Aach J, Church GM (2014) Highly multiplexed subcellular RNA sequencing in situ. Science 343:1360–1363
Lee H, Gurtowski J, Yoo S, Nattestad M, Marcus S, Goodwin S, McCombie W, Schatz M (2016) Third-generation sequencing and the future of genomics. BioRxiv. http://dx.doi.org/10.1101/048603
Lemmon EM, Lemmon AR (2013) High-throughput genomic data in systematics and phylogenetics. Annu Rev Ecol Syst 44:99–121
Lemmon AR, Emme SA, Lemmon EM (2012) Anchored hybrid enrichment for massively high-throughput phylogenomics. Syst Biol 61:727–744
Lepais O, Weir JT (2014) SimRAD: an R package for simulation-based prediction of the number of loci expected in RADseq and similar genotyping by sequencing approaches. Mol Ecol Resour 14:1314–1321
Li C, Hofreiter M, Straube N, Corrigan S, Naylor GJP (2013) Capturing protein-coding genes across highly divergent species. BioTechniques 54:321–326
Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO, Sandstrom R, Bernstein B, Bender MA, Groudine M, Gnirke A, Stamatoyannopoulos J, Mirny LA, Lander ES, Dekker J (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science 326:289–293
Lipshutz RJ, Fodor SPA, Gingeras TR, Lockhart DJ (1999) High density synthetic oligonucleotide arrays. Nat Genet 21:20–24
Liu Y, Zhou J, White KP (2014) RNA-seq differential expression studies: more sequence or more replication? Bioinformatics 30:301–304
Liu S, Wang X, Xie L, Tan M, Li Z, Su X, Zhang H, Misof B, Kjer KM, Tang M, Niehuis O, Jiang H, Zhou X (2016) Mitochondrial capture enriches mito-DNA 100 fold, enabling PCR-free mitogenomics biodiversity analysis. Mol Ecol Resour 16:470–479

Macaulay IC, Haerty W, Kumar P, Li YI, Hu TX, Teng MJ, Goolam M, Saurat N, Coupland P, Shirley LM, Smith M, Van der Aa N, Banerjee R, Ellis PD, Quail MA, Swerdlow HP, Zernicka-Goetz M, Livesey FJ, Ponting CP, Voet T (2015) G&T-seq: parallel sequencing of single-cell genomes and transcriptomes. Nat Methods 12:519–522

Mahairas GG, Wallace JC, Smith K, Swartzell S, Holzman T, Keller A, Shaker R, Furlong J, Young J, Zhao S, Adams MD, Hood L (1999) Sequence-tagged connectors: a sequence approach to mapping and scanning the human genome. Proc Natl Acad Sci U S A 96:9739–9744

Malone JH, Oliver B (2011) Microarrays, deep sequencing and the true measure of the transcriptome. BMC Biol 9:34

Marra MA, Kucaba TA, Dietrich NL, Green ED, Brownstein B, Wilson RK, McDonald KM, Hillier LW, McPherson JD, Waterston RH (1997) High throughput fingerprint analysis of large-insert clones. Genome Res 7:1072–1084

McCormack JE, Hird SM, Zellmer AJ, Carstens BC, Brumfield RT (2013) Applications of next-generation sequencing to phylogeography and phylogenetics. Mol Phylogenet Evol 66:526–538

McGee MD, Faircloth BC, Borstein SR, Zheng J, Darrin Hulsey C, Wainwright PC, Alfaro ME (2016) Replicated divergence in cichlid radiations mirrors a major vertebrate innovation. Proc R Soc Lond B Biol Sci 283:20151413

McIntyre LM, Lopiano KK, Morse AM, Amin V, Oberg AL, Young LJ, Nuzhdin SV (2011) RNA-seq: technical variability and sampling. BMC Genomics 12:293

Mendelowitz L, Pop M (2014) Computational methods for optical mapping. Gigascience 3:33

Meyer M, Fu Q, Aximu-Petri A, Glocke I, Nickel B, Arsuaga J-L, Martinez I, Gracia A, de Castro JMB, Carbonell E, Paabo S (2014) A mitochondrial genome sequence of a hominin from Sima de los Huesos. Nature 505:403–406

Michaeli Y, Ebenstein Y (2012) Channeling DNA for optical mapping. Nat Biotechnol 30:762–763

Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods 5:621–628

Mozo T, Dewar K, Dunn P, Ecker JR, Fischer S, Kloska S, Lehrach H, Marra M, Martienssen R, Meier-Ewert S, Altmann T (1999) A complete BAC-based physical map of the *Arabidopsis thaliana* genome. Nat Genet 22:271–275

Orlando L, Ginolhac A, Zhang G, Froese D, Albrechtsen A, Stiller M, Schubert M, Cappellini E, Petersen B, Moltke I, Johnson PLF, Fumagalli M, Vilstrup JT, Raghavan M, Korneliussen T, Malaspinas A-S, Vogt J, Szklarczyk D, Kelstrup CD, Vinther J, Dolocan A, Stenderup J, Velazquez AMV, Cahill J, Rasmussen M, Wang X, Min J, Zazula GD, Seguin-Orlando A, Mortensen C, Magnussen K, Thompson JF, Weinstock J, Gregersen K, Roed KH, Eisenmann V, Rubin CJ, Miller DC, Antczak DF, Bertelsen MF, Brunak S, Al-Rasheid KAS, Ryder O, Andersson L, Mundy J, Krogh A, Gilbert MTP, Kjaer K, Sicheritz-Ponten T, Jensen LJ, Olsen JV, Hofreiter M, Nielsen R, Shapiro B, Wang J, Willerslev E (2013) Recalibrating *Equus* evolution using the genome sequence of an early Middle Pleistocene horse. Nature 499:74–78

Oshlack A, Wakefield MJ (2009) Transcript length bias in RNA-seq data confounds systems biology. Biol Direct 4:14

Peterson BK, Weber JN, Kay EH, Fisher HS, Hoekstra HE (2012) Double Digest RADseq: an inexpensive method for de novo SNP discovery and genotyping in model and non-model species. PLoS One 7:e37135

Poland JA, Brown PJ, Sorrells ME, Jannink J-L (2012) Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. PLoS One 7:e32253

Prum RO, Berv JS, Dornburg A, Field DJ, Townsend JP, Lemmon EM, Lemmon AR (2015) A comprehensive phylogeny of birds (Aves) using targeted next-generation DNA sequencing. Nature 526:569–573

Puritz JB, Matz MV, Toonen RJ, Weber JN, Bolnick DI, Bird CE (2014) Demystifying the RAD fad. Mol Ecol 23:5937–5942

Putnam NH, O'Connell B, Stites JC, Rice BJ, Fields A, Hartley PD, Sugnet CW, Haussler D, Rokhsar DS, Green RE (2016) Chromosome-scale shotgun assembly using an in vitro method for long-range linkage. Genome Res 26:342–350

Rapaport F, Khanin R, Liang Y, Pirun M, Krek A, Zumbo P, Mason CE, Socci ND, Betel D (2013) Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. Genome Biol 14:1–13

Roberts RJ, Vincze T, Posfai J, Macelis D (2015) REBASE—a database for DNA restriction and modification: enzymes, genes and genomes. Nucleic Acids Res 43:D298–D299

Ruane S, Raxworthy CJ, Lemmon AR, Lemmon EM, Burbrink FT (2015) Comparing species tree estimation with large anchored phylogenomic and small Sanger-sequenced molecular datasets: an empirical study on Malagasy pseudoxyrhophiine snakes. BMC Evol Biol 15:1–14

References

Satija R, Farrell JA, Gennert D, Schier AF, Regev A (2015) Spatial reconstruction of single-cell gene expression data. Nat Biotechnol 33:495–502

Schena M, Shalon D, Davis RW, Brown PO (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science 270:467–470

Schuler DG (1997) Pieces of the puzzle: expressed sequence tags and the catalog of human genes. J Mol Med 75:694–698

Schulze A, Downward J (2001) Navigating gene expression using microarrays – a technology review. Nat Cell Biol 3:E190–E195

Schwarz A, Cabezas-Cruz A, Kopecky J, Valdes JJ (2014) Understanding the evolutionary structural variability and target specificity of tick salivary Kunitz peptides using next generation transcriptome data. BMC Evol Biol 14

Shapiro E, Biezuner T, Linnarsson S (2013) Single-cell sequencing-based technologies will revolutionize whole-organism science. Nat Rev Genet 14:618–630

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

Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S, Weinstock GM, Wilson RK, Gibbs RA, Kent WJ, Miller W, Haussler D (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res 15:1034–1050

Smith HO, Welcox KW (1970) A restriction enzyme from Hemophilus influenzae: I. Purification and general properties. J Mol Biol 51:379–391

Soderlund C, Longden I, Mott R (1997) FPC: a system for building contigs from restriction fingerprinted clones. Comput Appl Biosci CABIOS 13:523–535

Stepanauskas R (2012) Single cell genomics: an individual look at microbes. Curr Opin Microbiol 15:613–620

Studier FW, Moffatt BA (1986) Use of bacteriophage T7 RNA polymerase to direct selective high-level expression of cloned genes. J Mol Biol 189:113–130

Swan BK, Martinez-Garcia M, Preston CM, Sczyrba A, Woyke T, Lamy D, Reinthaler T, Poulton NJ, Masland EDP, Gomez ML, Sieracki ME, DeLong EF, Herndl GJ, Stepanauskas R (2011) Potential for chemolithoautotrophy among ubiquitous bacteria lineages in the dark ocean. Science 333:1296–1300

Tang F, Lao K, Surani MA (2011) Development and applications of single-cell transcriptome analysis. Nat Methods 8:S6–11

Tarazona S, García-Alcalde F, Dopazo J, Ferrer A, Conesa A (2011) Differential expression in RNA-seq: a matter of depth. Genome Res 21:2213–2223

Telford MJ, Budd GE, Philippe H (2015) Phylogenomic insights into animal evolution. Curr Biol 25:R876–R887

The Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature 408:796–815

The C. elegans Sequencing Consortium (1998) Genome sequence of the nematode *C. elegans:* a platform for investigating biology. Science 282:2012–2018

Todd EV, Black MA, Gemmell NJ (2016) The power and promise of RNA-seq in ecology and evolution. Mol Ecol 25:1224–1241

Toonen RJ, Puritz JB, Forsman ZH, Whitney JL, Fernandez-Silva I, Andrews KR, Bird CE (2013) ezRAD: a simplified method for genomic genotyping in non-model organisms. Peer J 1:e203

van Heesch S, Kloosterman WP, Lansu N, Ruzius F-P, Levandowsky E, Lee CC, Zhou S, Goldstein S, Schwartz DC, Harkins TT, Guryev V, Cuppen E (2013) Improving mammalian genome scaffolding using large insert mate-pair next-generation sequencing. BMC Genomics 14:257

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, Gocayne JD, Amanatides P, Ballew RM, Huson DH, Wortman JR, Zhang Q, Kodira CD, Zheng XH, Chen L, Skupski M, Subramanian G, Thomas PD, Zhang J, Gabor Miklos GL, Nelson C, Broder S, Clark AG, Nadeau J, McKusick VA, Zinder N, Levine AJ, Roberts RJ, Simon M, Slayman C, Hunkapiller M, Bolanos R, Delcher A, Dew I, Fasulo D, Flanigan M, Florea L, Halpern A, Hannenhalli S, Kravitz S, Levy S, Mobarry C, Reinert K, Remington K, Abu-Threideh J, Beasley E, Biddick K, Bonazzi V, Brandon R, Cargill M, Chandramouliswaran I, Charlab R, Chaturvedi K, Deng Z, Francesco VD, Dunn P, Eilbeck K, Evangelista C, Gabrielian AE, Gan W, Ge W, Gong F, Gu Z, Guan P, Heiman TJ, Higgins ME, Ji R-R, Ke Z, Ketchum KA, Lai Z, Lei Y, Li Z, Li J, Liang Y, Lin X, Lu F, Merkulov GV, Milshina N, Moore HM, Naik AK, Narayan VA, Neelam B, Nusskern D, Rusch DB, Salzberg S, Shao W, Shue B, Sun J, Wang ZY, Wang A, Wang X, Wang

J, Wei M-H, Wides R, Xiao C, Yan C, Yao A, Ye J, Zhan M, Zhang W, Zhang H, Zhao Q, Zheng L, Zhong F, Zhong W, Zhu SC, Zhao S, Gilbert D, Baumhueter S, Spier G, Carter C, Cravchik A, Woodage T, Ali F, An H, Awe A, Baldwin D, Baden H, Barnstead M, Barrow I, Beeson K, Busam D, Carver A, Center A, Cheng ML, Curry L, Danaher S, Davenport L, Desilets R, Dietz S, Dodson K, Doup L, Ferriera S, Garg N, Gluecksmann A, Hart B, Haynes J, Haynes C, Heiner C, Hladun S, Hostin D, Houck J, Howland T, Ibegwam C, Johnson J, Kalush F, Kline L, Koduru S, Love A, Mann F, May D, McCawley S, McIntosh T, McMullen I, Moy M, Moy L, Murphy B, Nelson K, Pfannkoch C, Pratts E, Puri V, Qureshi H, Reardon M, Rodriguez R, Rogers Y-H, Romblad D, Ruhfel B, Scott R, Sitter C, Smallwood M, Stewart E, Strong R, Suh E, Thomas R, Tint NN, Tse S, Vech C, Wang G, Wetter J, Williams S, Williams M, Windsor S, Winn-Deen E, Wolfe K, Zaveri J, Zaveri K, Abril JF, Guigó R, Campbell MJ, Sjolander KV, Karlak B, Kejariwal A, Mi H, Lazareva B, Hatton T, Narechania A, Diemer K, Muruganujan A, Guo N, Sato S, Bafna V, Istrail S, Lippert R, Schwartz R, Walenz B, Yooseph S, Allen D, Basu A, Baxendale J, Blick L, Caminha M, Carnes-Stine J, Caulk P, Chiang Y-H, Coyne M, Dahlke C, Mays AD, Dombroski M, Donnelly M, Ely D, Esparham S, Fosler C, Gire H, Glanowski S, Glasser K, Glodek A, Gorokhov M, Graham K, Gropman B, Harris M, Heil J, Henderson S, Hoover J, Jennings D, Jordan C, Jordan J, Kasha J, Kagan L, Kraft C, Levitsky A, Lewis M, Liu X, Lopez J, Ma D, Majoros W, McDaniel J, Murphy S, Newman M, Nguyen T, Nguyen N, Nodell M, Pan S, Peck J, Peterson M, Rowe W, Sanders R, Scott J, Simpson M, Smith T, Sprague A, Stockwell T, Turner R, Venter E, Wang M, Wen M, Wu D, Wu M, Xia A, Zandieh A, Zhu X (2001) The sequence of the human genome. Science 291:1304–1351

Vijay N, Poelstra JW, Künstner A, Wolf JBW (2013) Challenges and strategies in transcriptome assembly and differential gene expression quantification. A comprehensive in silico assessment of RNA-seq experiments. Mol Ecol 22:620–634

Wang Z, Gerstein M, Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet 10:57–63

Wang Y, Ghaffari N, Johnson CD, Braga-Neto UM, Wang H, Chen R, Zhou H (2011) Evaluation of the coverage and depth of transcriptome by RNA-Seq in chickens. BMC Bioinform 12(Suppl. 10):S5

Wang S, Meyer E, McKay JK, Matz MV (2012) 2b-RAD: a simple and flexible method for genome-wide genotyping. Nat Methods 9:808–810

Weber JL, Myers EW (1997) Human whole-genome shotgun sequencing. Genome Res 7:401–409

Wolf JBW (2013) Principles of transcriptome analysis and gene expression quantification: an RNA-seq tutorial. Mol Ecol Resour 13:559–572

Yilmaz S, Singh AK (2012) Single cell genome sequencing. Curr Opin Biotechnol 23:437–443

Zheng W-X, Zhang C-T (2008) Ultraconserved elements between the genomes of the plants *Arabidopsis thaliana* and Rice. J Biomol Struct Dyn 26:1–8

Assembly and Data Quality

C. Bleidorn, *Phylogenomics*, DOI 10.1007/978-3-319-54064-1_5

- The outputs of a sequencer are sequence reads, and each of its nucleotides receives a quality score, indicating the error probability.
- Overlapping sequence reads can be assembled into contiguous stretches of DNA called contigs, which can further on be ordered into scaffolds.
- Three main types of assembly strategies are in use, based on greedy algorithms, overlap-layout-consensus approaches or *k-mer* graphs.
- Different strategies are used for genome, transcriptome and metagenome assemblies, and all of them greatly benefit from the inclusion of long sequence reads.

5.1 Data Quality and Filtering

Sequence reads can be of either good or bad quality. To measure the error probability for a given base in a given sequence read, quality scores have been developed already back in the 1990s for Sanger sequencing. Based on sequence chromatograms, error probabilities were calculated for each position resulting in a quality score named *Phred* (Ewing and Green 1998).

$$Q_{\text{Phred}} = -10\log_{10}(P) \quad \textbf{(5.1)}$$

In this formula, P is the expected error probability for a given base call and Q_{Phred} specifies the according, logarithmically linked *Phred* score (Table 5.1). For example, a base call having a probability of 1/1000 to be wrong receives a *Phred* score of 30. High *Phred* scores correspond to low base-calling error probabilities, whereas low scores indicate higher ones. *Phred* quality values are always rounded to the nearest integer. *Phred* scores can handily be used to estimate the number of expected errors in sequence projects. Let us assume we sequenced a 70 Kb insert of a BAC clone with an average *Phred* score of 32 for every base. Given the formula above, this translates to an error probability of 0.00063, and one would have to expect ~44 wrongly called bases in this sequence.

Phred qualities are predicted without reference to a «true» sequence, but they were shown to correspond well with observed error rates. Moreover, it has been demonstrated that *Phred* scores have a high sensitivity to discriminate between correct and incorrect base calls. Due to their usefulness, these scores have been incorporated into Sanger sequencing machine analysis software early on. As such, *Phred* scores were routinely used to make decision regarding double peaks in the chromatogram or for trimming the ends of sequences to get rid of low-quality regions.

Table 5.1 *Phred* score and error probabilities

Phred score	Probability of incorrect base calls	Accuracy of base calling (%)
10	1 in 10	90
20	1 in 100	99
30	1 in 1000	99,9
40	1 in 10000	99,99
50	1 in 100000	99,999

Next-generation sequencing (NGS) techniques like Illumina, 454 or PacBio usually produce sequence reads with much higher error rates than Sanger sequencing. For each of these sequencing techniques, separate quality scores to assess sequencing reads were introduced. Unfortunately, the first quality scores introduced for Illumina (then called Solexa) sequencing were different from the established *Phred* scores and calculated as follows:

$$Q_{\text{Solexa}} = -10\log_{10}(p/1-p) \tag{5.2}$$

Using a rearrangement of this formula and an equation of the estimated error, it is possible to transform Q_{Solexa} values into *Phred* values:

$$Q_{\text{Phred}} = -10\log_{10}(10^{\text{QSolexa}/10}+1) \tag{5.3}$$

In 2009, Illumina also introduced the commonly used *Phred* score system as quality scores, which is also used for 454 and Ion Torrent. A *Phred* mimicking quality score is available for PacBio (Hackl et al. 2014). The usual format to store this quality information is the fastq format (Cock et al. 2010). Nanopore sequencing data using the MinION does not follow *Phred* scores (Laver et al. 2015).

Sequence reads in the fastq format are organized in four lines (Fig. 5.1). The first line always starts with an @ and contains information regarding the sequencing description and a unique identification. The second line contains the raw sequence. Usually the third line only contains the «+» sign or the «+» sign and a repetition of the sequence identification. The fourth line contains the quality values and must contain the same number of characters as the sequence line. As *Phred* scores are often values higher than 10 (and therefore needing two characters for description), they are encoded using the ASCII code, which comprises 128 specified characters including alphabetic letters, numbers and symbols. Each of these characters specifies one integer of the *Phred* score. To make things complicated, different ranges of values had been historically used for Sanger and some

Fig. 5.1 Example of the fastq format. **a** Sequence reads are organized in four lines. The first line indicates a name, the second line comprises the sequence data, the third line only contains a + plus (sometimes also the name again), and the fourth line bears the *Phred* quality information coded in ASCII. **b** ASCII code encoding *Phred* scores ordered from low to high scores

Illumina fastq formats, which has to be taken into account when analyzing this quality data. For Sanger and starting with Illumina 1.8 (and recent versions), the ASCII characters 33–126 are used, indicating a range of the *Phred* quality score from 0 to 93. In contrast, for some Illumina versions (1.3 to 1.8) the ASCII characters 64–126 are used, indicating a *Phred* quality score range from 0 to 62. Obviously, using the wrong ASCII translation might result in the strong over- or underestimation of error rates.

Base calling can be conducted using various methods. Moreover, different sequencing strategies are prone to different error types. As such, 454 and nanopore sequencing often produce errors due to the misspecification of the number of bases in homopolymers. PacBio sequences seem to be especially prone to contain chimeras (Hackl et al. 2014). In contrast, for Illumina sequence data, it might be difficult to distinguish between A and C and G and T, as both pairs of bases show similar emission spectra. Another problem for Illumina sequencing is generated due to a phenomenon called phasing. In this case, the incomplete removal of the 3′-blocking and fluorophore leads to the detection of the wrong signal during the next cycle (Kircher et al. 2011). Moreover, inverted repeats seem to be a problem caused by PCR amplification of single-stranded DNA during library preparation (Nakamura et al. 2011). All these error sources have to be taken into account by base-calling programs. Several programs exist besides software distributed with the analysis pipeline of the Illumina machines, like IBIS and FREEIBIS (Kircher et al. 2009; Renaud et al. 2013). For MinION nanopore sequencing, NANOCALL is a freely available open source base caller (David et al. 2017).

After base calling, a first quality filtering of the sequence data is usually conducted. In this step, adapter sequences and barcodes (or index primer regions) are removed. In the next step, it is recommended to remove low-quality regions or discard such reads completely. For Illumina reads, often an exponential increase in error probabilities from the 5′- to 3′-end is observed. The stringency of the filter procedure is chosen by the user and different parameters might be exploited as part of the analysis. For example, using filtering sequence reads which contain more than 5% of bases under a specified quality score (e.g. *Phred* 20) could be removed. As also known for Sanger sequences, the 5′- and 3′-ends of reads often show lower quality than the rest of the sequence read and might be completely discarded. This process is called trimming. Initial rigorous quality checks of sequence reads clearly increase the quality and minimize the number of artefacts of subsequent analysis steps, as assembly or mapping. Several programs are available to visualize the distribution of error probabilities across reads, as, for example, the freely available software fastqc report (► http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). TRIMMOMATIC is a widely used software for Illumina read trimming (Bolger et al. 2014), a package of perl scripts called CONDETRI is another easy to use for trimming and read filtering tools (Smeds and Kunstner 2011), and error correction can be conducted with the program QUAKE (Kelley et al. 2010). As long reads from PacBio or nanopore sequencing are often more error prone, they can be corrected in a hybrid approach using Illumina short reads (Salmela and Rivals 2014; Koren et al. 2012; Goodwin et al. 2015).

5.2 Assembly Strategies

Usually two different main strategies analyzing sequence reads are employed: assembly and mapping. Mapping describes a procedure to align sequence reads or assembled contigs on a given reference sequence. This reference might be a genome or transcriptome of

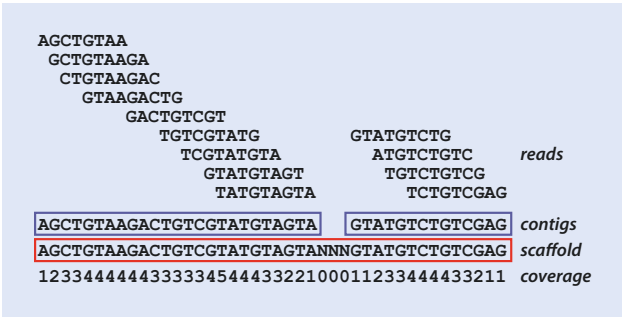


Fig. 5.2 Important terms for understanding sequence assembly methods. Reads are assembled into contigs, which can be ordered into scaffolds. The coverage shows the number of reads covering a certain position in the contig

the target species or from closely related species. As this strategy is basically an application of alignment methods, it will be introduced in the corresponding chapter. Assembly refers to the procedure of generating longer, continuous stretches of sequences by using combinations of shorter sequence reads (Miller et al. 2010). Some basic terms are important to know before digging deeper into how different assembly strategies and methods work (Fig. 5.2).

An assembly is a set of contigs computed from sequence reads. A sequence derived from assembling several sequence reads is called contig. Some methods further work with unitigs, which are basically high-confidence contigs (Myers et al. 2000). Evolutionary studies based on NGS data usually work with contigs, and it is important to remember that these do not refer to observations derived from a sequencing technique but are products of the ongoing analysis. Using different assembly strategies, different assembling parameters or even different quality procedures as described above may lead to different contigs. In the ideal case for whole-genome shotgun data, a single contig refers to a single chromosome, which might already constitute the complete genome in case of bacteria or organelle genomes. However, usually more and smaller contigs compared with the number of sequenced chromosomes are the result of the assembly. By using additional information (e.g. positional information from paired-end or mate pair reads), these contigs can be ordered into scaffolds, which further resolve the orientation of contigs to each other. In scaffolds, nonoverlapping stretches between contigs are marked by stretches of N's (unspecified bases). As part of the analysis, sequence reads can be mapped onto contigs. How often a given sequence position is covered by sequence reads is called coverage. A coverage of 10x means that any sequence position of an assembly is covered on average by ten sequencing reads. Before starting a sequencing project, it is useful to estimate the genome size of the target organism. For example, in case of the genome size of humans (~3 Gb), a single lane of paired-end sequencing (100 cycles) by the Illumina HiSeq sequencer would already produce an expected data volume of ~60 Gb, corresponding with a theoretical coverage of 20x (note that the practical coverage will be considerably lower). An oversampling of the genome in terms of coverage is important to have overlapping reads for assembly.

Assemblies are often compared with solving puzzles, and in case of de novo assemblies, even the desired picture is unknown. If a genome of the target species or from a closely related species is already available, it can be used to guide the assembly. Three main types of assembly methods are currently in use: greedy, overlap-layout-consensus methods and *k*-mer assemblies (Miller et al. 2010). Numerous assemblers are available for all these methods (Table 5.2).

Table 5.2 Overview of some current widely used sequence assemblers

Program name	Citation/source	Strategy	Remarks
ABruijn	Lin et al. (2016)	*k*-mer	Long-read assembly for PacBio and nanopore data
AbySS	Simpson et al. (2009) Robertson et al. (2010)	*k*-mer	Versions for genome and transcriptomes available
ALLPATHS	MacCallum et al. (2009)	*k*-mer	Hybrid assemblies using short and long reads
ARACHNE	Batzoglou et al. (2002)	OLC	Whole-genome assembly
Canu	Koren et al. (2016)	OLC	Long-read assembly for PacBio and nanopore data
CAP3	Huang and Madan (1999)	greedy	Useful for Sanger data
Celera	Myers et al. (2000)	OLC	Strictly a variant of OLC using so-called string graphs. Used to assemble the genomes of *Drosophila melanogaster* and humans
CLC	► https://www.qiagenbioinformatics.com/products/clc-genomics-workbench/	*k*-mer	Commercial software package for genomic applications. Easy to use and memory efficient
Edena	Hernandez et al. (2008)	OLC	Fast and memory efficient
Euler	Pevzner et al. (2001)	*k*-mer	First *k*-mer assembler
FALCON	► https://github.com/PacificBiosciences/FALCON-integrate	OLC	Long-read assembly of PacBio data
MEGAHIT	Li et al. (2016)	*k*-mer	Fast and memory-efficient metagenome assembler
Minia	Chikhi and Medvedev (2014)	*k*-mer	Memory efficient, usable in low memory environments
Miniasm	Li (2016)	OLC	Ultrafast long-read assembly for PacBio and nanopore data
MIRA	Chevreux et al. (2004)	hybrid	Swiss army knife of sequence assembly, useful for combining different technologies
Newbler	Distributed with 454 (Roche) sequencing platforms	OLC	Standard for 454 data

Table 5.2 (continued)

Program name	Citation/source	Strategy	Remarks
IDBA	Peng et al. (2010) Peng et al. (2011) Peng et al. (2013)	*k*-mer	Iterative *k*-mer size, versions for genomes, transcriptomes and metagenomes
Oases	Schulz et al. (2012)	*k*-mer	De novo transcriptome assembly, splice variants
PHRAP	▶ http://www.phrap.org/	greedy	Useful for Sanger data
SOAPdenovo	Luo et al. (2012)	*k*-mer	Assembly of the first eukaryotic genome solely based on short reads, different modules
SPAdes	Bankevich et al. (2012)	*k*-mer, hybrid	Assembler for bacterial genomes, hybrid module to include PacBio reads
TruSPAdes	Bankevich and Pevzner (2016)	*k*-mer	Assembler for synthetic long reads (e.g. TruSeq, 10x Genomics)
Trinity	Grabherr et al. (2011)	*k*-mer	De novo transcriptome assembler, splice variant detection
Velvet	Zerbino and Birney (2008)	*k*-mer	For genomes, included in some transcriptome assemblers

5.2.1 Greedy Assemblies

Assemblers using the greedy algorithm represent the most simple and intuitive approaches. In this case, sequence reads are iteratively joined to build contigs, starting with those showing the highest score for an overlap. These scores measure the amount of matching bases and the length of the overlap region, and parameters can usually be defined by the user. The operation of joining reads and/or contigs is repeated using the same rules till no more steps are possible. The term greedy refers to the fact that due to the search for best overlaps, only local optimal solutions are analysed, which leads to a result comparatively fast. However, the best overall (global) assembly might be missed using this strategy. An assembly of sequences from PCR experiments consisting of several overlapping fragments were usually constructed with this method, as implemented in widely distributed software, e.g. CAP3 (Huang and Madan 1999).

5.2.2 Overlap-Layout-Consensus (OLC) Assemblies

The OLC assembly can be divided into three steps. In the first step, pairwise alignments of all sequence reads are conducted, where overlaps for each pair of sequence reads are analysed. This overlap might be perfect (a match of corresponding nucleotides in every overlapping position) or contains few mismatches. However, only overlaps of ends of sequencing reads are allowed (Fig. 5.3).

The information of overlapping reads is stored in graphs, so-called overlap graphs (Fig. 5.4). Mathematical graphs are of central importance of many genomic methods and will appear in several chapters of this book. In the case of overlap graphs, the nodes in the graph represent sequence reads, whereas the branches (or alternatively called edges) indicate which reads are connected by an overlap (Fig. 5.4b). Due to the fact that edges should only be traversed in one direction (as indicated by arrows), the result is a directional graph. The number of subgraphs produced in this step will correspond to the number of contigs which are resolved.

The second step is the layout. During this step, the relative position of sequence reads (nodes) of every overlap graph is determined and arranged accordingly into an alignment. This is conducted by searching for a mathematical path describing a way to go over every

Fig. 5.3 Overlap of sequence reads. **a** Only overlaps of ends of sequence reads will be included in the subsequent graph. The minimum number of overlapping nucleotides has to be specified by the user and mismatches might be allowed. **b** When overlapping regions are in the middle of one of the sequences, they will not be used in subsequent analysis

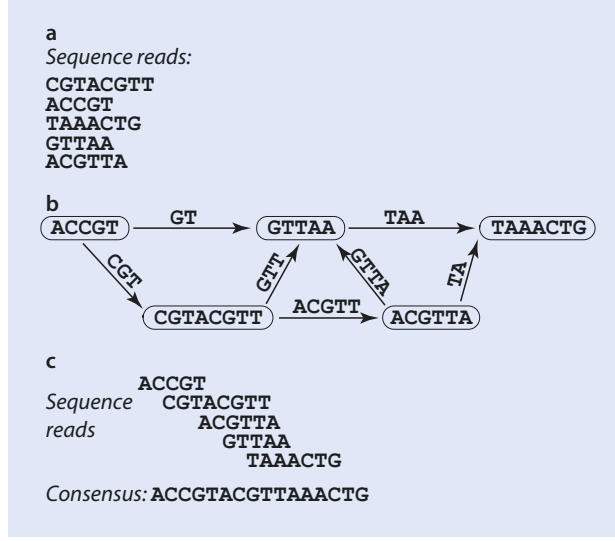


Fig. 5.4 Overlap-layout-consensus. **a** Sequence reads for assembly. **b** Overlap graph. **c** Alignment of reads after layout step, in which a Hamiltonian path was searched for in the overlap graph. The consensus sequence is the resulting contig

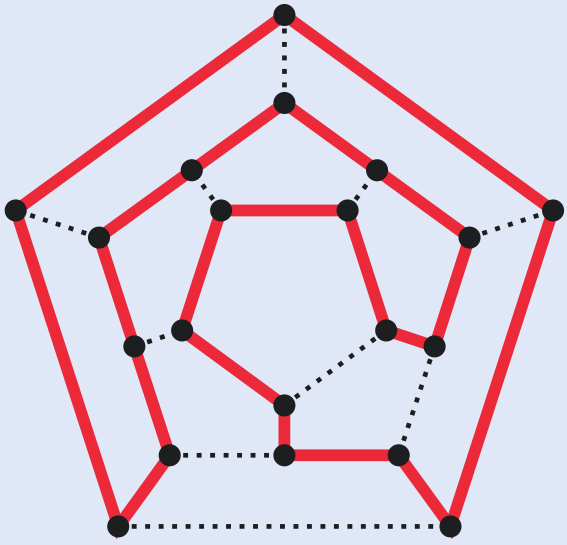
node exactly once, the Hamiltonian path (▶ see also Infobox 5.1). In the last step, the resulting alignments are used to determine consensus sequences which represent contigs (◘ Fig. 5.4c). Coverage information can be used to correct base calling or sequencing errors. For example, in case of resolving a certain nucleotide for a given position in a contig, the alternative supported by most of the reads covering this position is chosen. For example, when for a certain sequence position with a 10x coverage eight times an A is read, but only two times a C, then A will be chosen.

Infobox 5.1

The Century-Old Origin of Short-Read Genome Assembly Algorithms

Solutio problematis ad geometriam situs pertinentis was the name of an article by the Swiss mathematician and physicist Leonhard Euler (1707–1783) which described a solution for the so-called Bridges of Königsberg problem presented to the St. Petersburg Academy in 1735. This ground-breaking work not only solved an old mathematical problem but also was one of the first contributions to graph theory, including an idea that is now part of *k*-mer assembly strategies. The former Eastern Prussian and now Russian city of Königsberg (Kaliningrad) is located at the opposing sites of the river Pregel, as well as on two river islands. The four parts of Königsberg were joined by seven bridges. The «Bridges of Königsberg problem» asked the question if it would be possible two visit all four parts of the city by crossing every bridge exactly once, while returning to the starting point. Euler's brilliant idea to solve this problem was to represent each part of the city as a node and each bridge of the city as an edge and to connect them appropriately within a graph. Euler described a method that finds a path traversing the graph while visiting each edge exactly once, and this path is still known as the Euler path. Unfortunately, there is no way to cross the seven bridges of Königsberg exactly once and visiting all parts of the city.

Another important mathematical path is named after William R. Hamilton (1805–1865), an Irish mathematician and physicist. The Hamiltonian path visits each node of a graph exactly once. A graph that contains a Hamiltonian path that forms a cycle is called Hamiltonian cycle. Hamilton used such cycles to invent the icosian game. The aim of this game is to find a Hamiltonian path along the edges of a dodecahedron, a geometrical figure which might also be described as a cube with 12 flat faces (◘ Fig. 5.5).



◘ **Fig. 5.5** Hamiltonian Path through a Dodecahedron by Christoph Sommer - Own work. Licensed under Creative Commons Attribution-Share Alike 3.0 via Wikimedia Commons

```
true genome sequence:
AAGACTGTCGTATGTATATATACCAAGGTTCCTATATATATGTCTGTCGAGCGTC
AAGACTGTCGTATGTATATATA                          read_1
              TATATATATGTCTGTCGAGCGTC           read_2
AAGACTGTCGTATGTATATATATGTCTGTCGAGCGTC           assembly
```

Fig. 5.6 Example of a wrong assembly of a repetitive region. The repeat motive is given in *red*, a stretch of the true sequence which is missing in the resulting assembly is given in *blue*

OLC assemblers were originally developed for the analysis of Sanger sequence data. Accordingly, they are well suited when analysing moderate amounts of larger sequence reads (>500). The first sequenced animal and plant genomes were assembled with methods based on this strategy (Myers et al. 2000). A widely used assembler for 454 data, a sequencing technique which usually generates longer but less reads than Illumina sequencing, is NEWBLER (Table 5.2). In this software, two subsequent OLC steps are performed. In the first step, so-called unitigs are generated by the process described above. Unitigs are high-confidence contigs composed of reads which do not bear overlap with reads in any other unitig. In the second step, unitigs are joined into longer contigs based on pairwise overlap between unitigs.

However, some problems remain with OLC methods. Firstly, finding the Hamiltonian path is a mathematically NP-hard problem. Nondeterministically, polynomial-time hard problems are those which are not efficiently solvable by algorithms. This basically means that computers are and will always be too slow to calculate this kind of mathematical paths. As always in these cases, heuristic solutions are used in hope to find the best path for the problem. Secondly, the first step of finding overlaps by pairwise alignments becomes too time and memory intensive with NGS data. Originally developed for hundreds to rarely up to millions of longer sequence reads of very high quality, OLC becomes problematic to unusable when dealing with millions to billions of short reads of often lower quality. Moreover, usage of OLC methods can be problematic to resolve long repetitive regions and may produce misassemblies in this case (Fig. 5.6).

To avoid these problems, low-complexity regions (e.g. long stretches of a single nucleotide) and repetitive regions are often masked and discarded before assembly. Alternatively, an error correction step can be performed before starting the assembly. However, the availability of long-read sequences (e.g. from PacBio and nanopore sequencing) revived the OLC approach, and many assemblers dealing with this kind of data have been recently published (see Table 5.2).

5.2.3 K-mer Assemblies Using de Bruijn Graphs

Assemblies using de Bruijn graphs based on *k*-mers are composed of two steps: In the first step, the sequence reads are fragmented into smaller pieces called *k*-mers, which are used to construct a de Bruijn graph. In the second step, the contigs are derived from the de Bruijn graph (Schatz et al. 2010).

Every sequence, reads from a sequencer as well as complete genomes downloaded from GenBank, can be fragmented into *k*-mers. The *k* in *k*-mers denotes the size of the fragment, and after choosing this, the sequence is fragmented in all possible *k*-1 overlapping fragments of this size (Fig. 5.7).

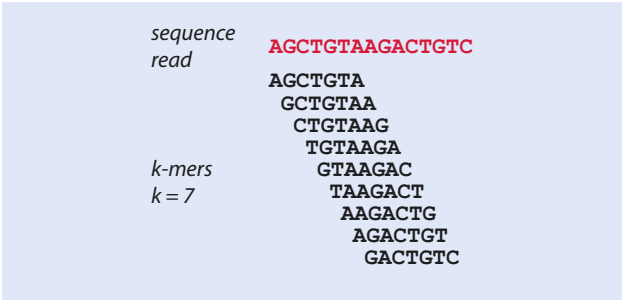


Fig. 5.7 An example sequence and all its *k*-mers of the size 7

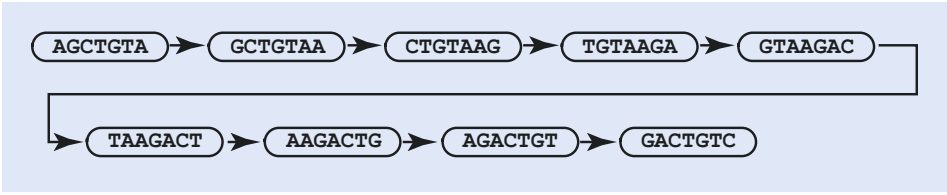


Fig. 5.8 De Bruijn graph of *k*-mers from Fig. 5.7

Surprisingly, the most common method to deal with de novo assemblies of short-read data is to fragment these short reads into even smaller pieces. The resulting *k*-mers are then connected via a de Bruijn graph. In the case of assemblies, the de Bruijn graph is a graph where the nodes represent sequences (*k*-mers) which are connected by edges in case they show a *k*-1 overlap. Arrows are used to indicate the direction of the overlap from the *k*-mer where the last *k*-1 nucleotides overlap to the *k*-mer with the first *k*-1 nucleotides (Fig. 5.8).

To reconstruct contigs, the de Bruijn graph has to be traversed by finding a Euler path (▶ see also Infobox 5.1). The Euler path goes exactly once over every edge of the graph. Reconstructing the contig derived from perfect *k*-mers of a single short sequence is a simple problem. However, real genomic data is usually more complex, including repetitive regions. Further on, real sequencing data usually contains errors. Both errors and repeat regions lead to more complex de Bruijn graphs which are much more difficult to resolve. The presence of repeats can introduce loops into the graph as illustrated by a simple example (Fig. 5.9).

Likewise, sequencing errors lead to more complex graphs. Errors in the middle of a sequence can lead to bubbles in the graph, whereas errors at the end of sequences may introduce dead ends (tips) into the graph (Fig. 5.10).

Both repeats and number of errors introduced are directly influenced by the chosen *k*-mer value. Unfortunately, this choice represents a trade-off (Chikhi and Medvedev 2014). Larger number of *k* reduces the number of repeats which can tangle the graph and break up contigs. Obviously, a repetitive region or sequence motive which is longer than the chosen *k* cannot be resolved. This would argue for choosing the highest possible value for *k*, which is limited by the length of the sequence reads. However, with longer *k*-mers, the probability that these *k*-mers contain sequencing errors increases. For example, given an error in the middle of a 100 bp sequencing read and a chosen *k*-mer size of 25, up to 25 erroneous *k*-mers are included into the analysis. When choosing a *k*-mer size of 13 for the same data, only up to13 erroneous *k*-mers are created. The choice of *k* also directly influences the size of the de Bruijn graph, as lower *k*-mer values decrease the number of edges stored in the graph which at the same time reduces the amount of memory needed to store

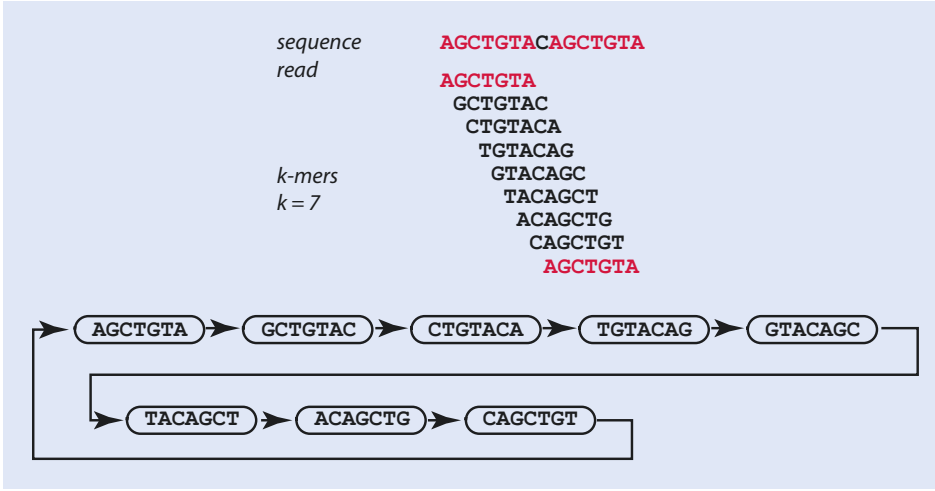


Fig. 5.9 Repetitive sequences can lead to loops in a de Bruijn graph. The repetitive motive is indicated in *red*

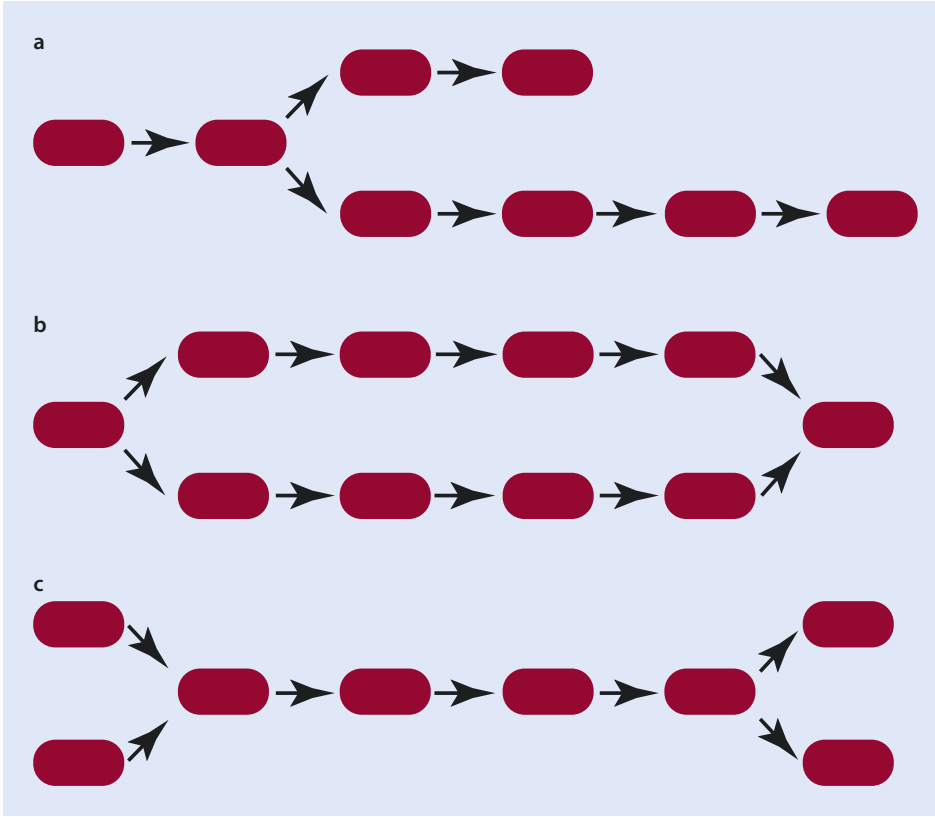


Fig. 5.10 Sequence errors and repeats lead to more complex *k*-mer graphs. Nodes representing *k*-mers are indicated by red boxes. **a** Errors at the end of sequence introduce dead ends into the graph **b**. Errors in the middle of sequences introduce bubbles into the graph. **c** Repeat sequences lead to a pattern of convergent and divergent paths (After Miller et al. (2010))

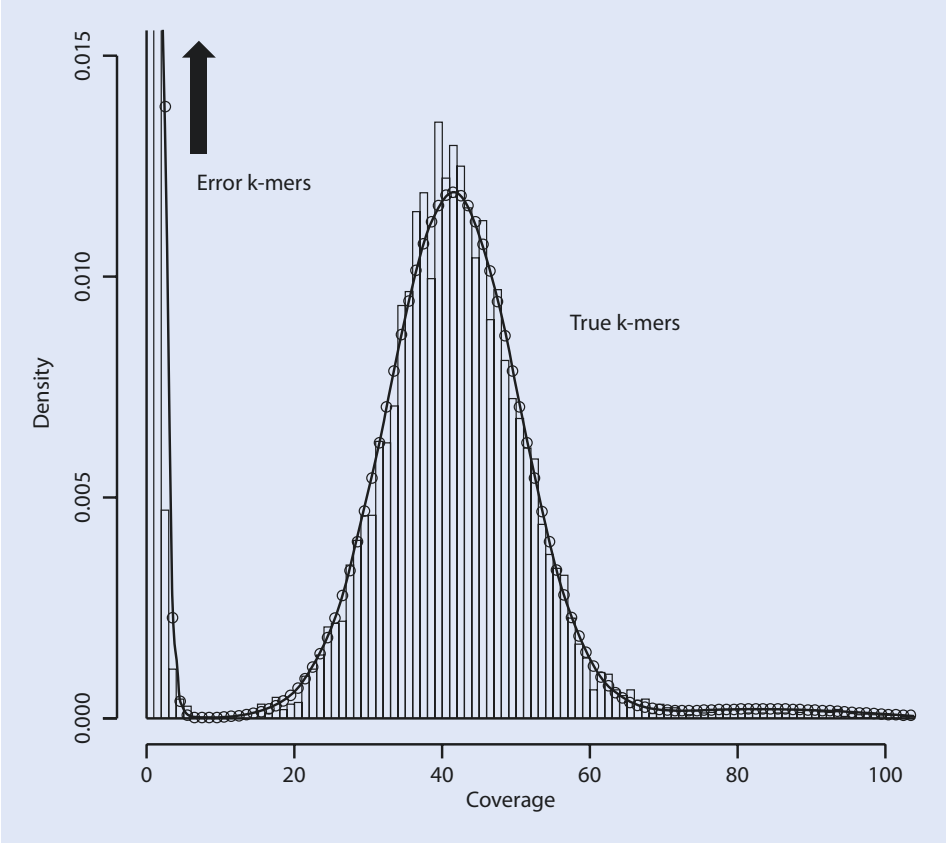


Fig. 5.11 Typical *k*-mer distribution for genomic data originating from one individual/species. The coverage of *k*-mers is plotted against its frequency (density). Low-coverage *k*-mers likely represent sequencing errors (Reprinted from Kelley et al. (2010))

this information. For this reason, several *k*-mer-based assemblers have an upper limit for *k*, as otherwise the computation becomes too memory intensive. On the other hand, larger *k*-mers are more informative and the numbers of nodes in the graph which can be traversed are decreased, making it easier finding paths through it. Consequently, the first step of any *k*-mer assembly is the careful choice for *k*. The number of *k*-mers generated per read can be estimated by a simple formula.

$$\text{Formula 5.4: } N_{\text{kmers}} = L_{\text{read}} - k + 1$$

In this formula, N_{kmers} refers to the calculated number of *k*-mers per read, with the read length defined by L_{read}. However, when counting *k*-mers for subsequent assemblies, only unique *k*-mers are stored together with the information how often they occurred. This can be done for a range of different *k*-values. Usually only uneven integers are used for *k*, as in case of even-numbered *k*-mers the occurrence of palindromic sequences can introduce further complexity into the graph, leading to shorter contigs. Plotting the coverage of *k*-mers against its frequency (Fig. 5.11) can be used to choose the optimal *k*-mer value for assembly. These plots are generated for many *k*-mers sizes, and the one leading to the highest number of distinct non-erroneous *k*-mers is chosen (Chikhi and Medvedev 2014).

Moreover, *k*-mer counting is used for error correction before assembly and can be used to detect repeated sequences, e.g. transposons (Marçais and Kingsford 2011).

K-mer frequencies can further be used to get a rough estimate of the genome size. A first step is to plot the *k*-mer coverage against the frequency as indicated in Fig. 5.11. Such outputs can be quite easily generated, for example, with the software JELLYFISH (Marçais and Kingsford 2011). The histograms can be used to distinguish between erroneous *k*-mers and potentially true *k*-mers (Fig. 5.11). The peak of the true *k*-mer distribution gives an estimate of the coverage of the genome (~40× in the example in Fig. 5.11). In the last step, the total number of true *k*-mers is divided by the coverage estimate to get an estimate of the total genome size. However, this number can be a huge underestimation if the genome bears a high percentage of repeat regions.

Several approaches can be used to choose the best *k*-value. An obvious way would be to generate assemblies for every possible *k* and choose the best after comparison of the assemblies. However, assemblies are usually very memory-intensive computations, and especially in case of large *k*'s, this way is not suitable. An intuitive (and heuristic) way to choose the best *k* is to generate abundance histograms (as shown in Fig. 5.11) for many values of *k* and to choose the value which generates the highest number non-erroneous *k*-mers (Chikhi and Medvedev 2014). A different approach is used by the IDBA assembler, which uses an iterative *k*-mer optimization, thereby de facto using different *k*-mer size for one assembly (Peng et al. 2010).

After choosing a value for *k* and fragmentation of sequence reads into *k*-mers, rare *k*-mers should be discarded. The logic is that in case of high-coverage genome sequencing, the abundance of *k*-mers should correlate with the expected coverage. Rare *k*-mers likely arose from sequencing errors. Likewise, overabundant *k*-mers are assumed to originate from high copy number regions of the genome (e.g. ribosomal cluster, transposons). This assumption might not work for transcriptome assemblies (▶ see Sect. 5.4). The resulting *k*-mer graph will be finally used to generate contigs by finding the Euler paths. Many software applications are available for generating *k*-mer assemblies (Table 5.2).

Compared with the OLC strategy, *k*-mer approaches bear some advantages for assembling huge numbers of short reads. Firstly, as no step for initial pairwise alignments is involved, *k*-mer strategies are much more time and memory efficient. Moreover, efficient algorithms are available for finding the Euler path within a de Bruijn graph. However, *k*-mer assemblies are usually less robust against sequencing errors, and the number of potential Euler paths is exponential to the number of repeats in the genome. Not surprisingly, especially the de novo assembly of eukaryotic genomes remains a challenge.

5.3 Comparing Assemblies

Different assembly strategies, programs or parameters can lead to vastly different sets of contigs. This raises the important question of how to judge different assemblies. A straightforward assessment of the accuracy of an assembly would be to compare it with an independent sequencing project of the same organism. This has been partly conducted for the assembly of the panda genome, which was the first assembled complex eukaryotic genome solely based on short-read data (Li et al. 2010). For validation, extra Sanger sequencing was performed to assess the quality of the short-read assembly. However, such an experimental setup is time intensive and costly and usually not practicable. When a reference genome is available, this can be also used for comparison, and BLAST searches

could be conducted to find assembled contigs. However, in many cases no reference of the same or even closely related species is available. Some metrics are available to describe assemblies without referring to a reference, and the most widely used is the N50 (Fig. 5.12). The N50 of an assembly is a weighted median and means that half of the assembled bases of all contigs are represented by contigs of this length or longer (Salzberg et al. 2012). To calculate the N50, after assembly all contigs are ordered according to their size, starting with longest one. In the next step, the overall contig size summing up all contigs is calculated. Lastly, starting with the longest contig, the next longest one is added until the sum of these combined contigs reaches 50% or more of the size of all contigs. The length of the contig added in this last step defines the N50 value for an assembly. For example, if the N50 is 15 mb, it means that contigs which contain 50% of the nucleotides of the complete assembly are at least 15 mb or longer. For genome assembly, it is usually desired to have higher N50 values. In an equal way, values for N75 or N80 could be calculated, containing the information for the respective percentages of the total contig size. The N50 is part of the output of many assemblers, but for comparison, it is important that the total contig size is calculated comparable, as often contigs with a size under a certain threshold (e.g. 500 bp) are not included in the calculation. A convenient way for multiple assembly comparison is to use the *QUAST* (quality assessment tool for genome assemblies) software, which could be either installed locally or used on an online server (Gurevich et al. 2013).

How well does the N50 describe the quality of a genome assembly? And which is the best assembler for my problem? To get an answer for these questions, the scientific community organized a competition for de novo short-read assemblers, the Assemblathon (► www.assemblathon.org). For the first competition, interested groups of software developers got sequence reads of a simulated eukaryotic genome, including contaminations and a typical sequence error profile. The genome had to be assembled de novo (even though information of a simulated reference genome could have been included), and the

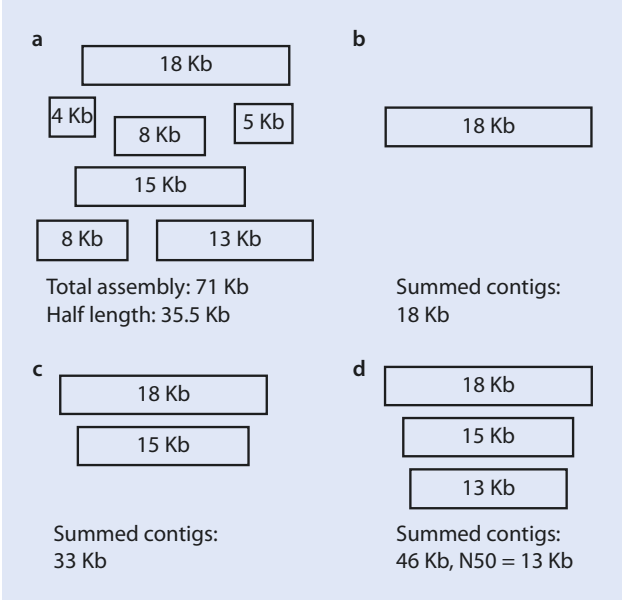


Fig. 5.12 Calculation of the N50 metric for sequence assemblies. **a** The assembly size if all contigs is 71 Kb. The cut-off value for the N50 is 35.5 Kb. **b** Sequences are summed step by step, starting with the longest. **c** Adding the second longest contig sums up to a total size of 33 Kb, still under the cut-off for the N50. **d** Adding the third longest contig result in a total sum of 46 Kb, which is over the cut-off (35.5 Kb). The length of this contig equals the N50 (13 Kb)

outcome was compared to the «true» simulated genome. Using this data, metrics relying on a reference genome could be compared with those working without any reference. One of the most important outcomes was that the N50 indeed seems to be a good way to describe assembly quality (Earl et al. 2011). Moreover, large differences could be shown between different assemblers. Consequently, the second iteration of the competition targeted this issue. This time, real genomic sequence data of three vertebrates was analysed. Overall, the tested genome assemblers produced useful assemblies, providing a significant representation of genes and overall genome structure. However, it was found that approaches which work well in assembling the genome of one species may not necessarily work well for another (Bradnam et al. 2013). The practical advice is to use different assemblers (Table 5.2) and to choose the best assembly based on available metrics like the N50 afterwards.

It is well known for phylogenetic tree reconstruction that many equally or similarly good solutions can be the outcome of the analysis. The same is true for assemblies, where due to uncertainty alternative solutions for contig building may be present. In a Bayesian framework, each assembly alternative could be given a probability, making it possible to evaluate different assemblies in a statistical framework (Howison et al. 2014; Howison et al. 2013). The development of software implementing these strategies is at the beginning, but ideas like this will open new future directions for choosing the best assembly.

5.4 De Novo Assembly of Genomes

Genomes can differ hugely in size and content of repetitive regions and so differ in their degree of difficulty to be assembled. Genomes (or chromosomes) also dramatically exceed the length of sequence reads generated by any sequencing technique actually used. Therefore, strategies like whole-genome shotgun sequencing are used, where the genome is fragmented in small pieces, and later these sequenced fragments are puzzled into the complete genome by assembly programs. The most widely used technique today is Illumina sequencing. However, as the recovered reads are usually not longer than 150 bp (HiSeq) or 250 bp (MiSeq), especially the assembly of complex eukaryotic genomes including many repeat regions remains challenging. Different strategies are applied to improve the initial assembly and to combine contigs into longer pieces.

5.4.1 Scaffolding

Contigs can be linked together into longer pieces via scaffolding. The information to bring contigs into an order usually comes from paired-end reads or mate pairs. As for assembly, graph theory can be used to solve scaffolding. In this case, the assembled contigs represent the nodes of the graph, and they are linked through read pairs as represented by edges (Hunt et al. 2014). Moreover, whereas the ordering information (and orientation) of contigs can be retained from such graphs, the expected length can be deduced from the approximate distance of read pairs known due to library preparation. This information can be included in the length of the edges connecting nodes in the graph. In scaffold sequences, this distance is given by N's inserted between two linked contigs. The first step for scaffolding is always the mapping of the actual sequence reads onto the contigs, to know the location of paired-end or mate pair reads. In the second step, this information is

comprised into a graph as described above. The available scaffolders use different strategies to resolve the graph into contigs, from exact mathematical solutions to heuristic approaches simplifying the graph into subgraphs. Some fast solutions are based on greedy algorithms where those contigs are joined first which are linked by the highest number of edges. Comparable to what has been found for assembly programs, scaffolding software can vary strongly in their performance and the required analysis time, based on the complexity of the analysed genome (Hunt et al. 2014). Several assembly programs also include scaffolding modules. Stand-alone scaffolding tools are, for example, SSPACE (Boetzer et al. 2011), SCARPA (Donmez and Brudno 2013) or OPERA (Gao et al. 2011). As in the case of assemblers, the performance of several scaffolders should be compared to choose the best suited for the task at hand.

5.4.2 Hybrid Assemblies

Assembling complex genomes solely with short reads is a difficult task, especially when no reference genome from a closely related organism is available. Not surprisingly, the difficulty of genome assembly is reduced with increasing sequence read length. Whereas Illumina sequencing is by far the most widely used technique, other high-throughput sequencing techniques generating considerably longer reads are available (e.g. PacBio and nanopore sequencing). The caveat with long reads of current sequencing techniques is their high error rate. For example, by using PacBio, an error rate of approximately 15% is expected, which can additionally vary dramatically across positions (Chin et al. 2013). Current assembly strategies are not equipped to directly deal with these high error rates (Sović et al. 2016). For example, when using OLC it is easily conceivable that perfect overlaps are hard to find. In contrast, *k*-mer-based assemblers need to find exact-matching k-mers between reads, which is an limiting factor when dealing with high error rates. To deal with these problems, hybrid methods have been developed taking advantage from the fact that Illumina short-read data, which has a much lower error rate than PacBio or nanopore long reads, is perfectly suited for error correction of long reads (Koren et al. 2012). Using hybrid assembly strategies, high-coverage (50x and higher coverage of the genome) short-read data (100–150 bp per read) is combined with low-coverage (10–20x) long reads (reads >5000 bp). The first step would be to assemble all short reads into contigs. These contigs are mapped onto the long reads for error correction. Error-corrected long reads can then be overlapped to get long contigs. In the last step, scaffolding as described above using short-read paired-end or mate-pair data might be performed. Alternatively, Illumina short reads can be assembled first, and then long reads are incorporated to bridge coverage gaps and resolve repeats, e.g. using the assembler ALLPATHS-LG (MacCallum et al. 2009).

5.5 De Novo Assembly of Transcriptomes and Metagenomes

Transcriptomes comprise the total RNA or mRNA expression data of isolated cells or tissue. Genes with a high expression will be represented by many sequence reads, whereas genes with low expression will yield few reads and non-expressed genes will obviously be missed totally (Wang et al. 2009). As such, a mixture of full-length and partial transcripts at various levels of abundance is expected. Consequently, huge differences in coverage of

different transcripts will be found in the final assembly. To further complicate things, different forms of alternatively spliced genes might be recovered (at least in eukaryotes). This is a challenge for all currently used assembly algorithms, and assembly quality decreases as transcriptome complexity increases (Chang et al. 2014).

As for genome assemblies, major strategies for transcriptome assemblies are reference based, de novo or a mixture of both. Reference-based assemblies consists of three steps: first, alignment/mapping of reads against a reference genome is conducted; second, overlapping reads of one locus are used to build a graph; and, third, the resulting graph is traversed to resolve isoforms (Martin and Wang 2011). De novo transcriptome assembly of short-read data is usually facilitated by de Bruijn graph-based assembly methods. It has been shown that lower *k*-mer sizes yield more sensitive assemblies, recovering also lowly expressed variants, whereas higher *k*-mer sizes result in more specific assemblies, leading to a more accurate assembly of abundant transcripts (Nagarajan and Pop 2013). It is important to keep in mind that the interpretation of *k*-mer abundance is less straight forward than for genome assemblies. Whereas in the latter case rare *k*-mers are probably originated from sequencing errors, rare *k*-mers may alternatively originate from lowly expressed transcripts in the case of transcriptome sequencing. A common idea of many transcriptome assemblers is to use different *k*-mer values. The resulting assemblies are merged afterwards, with redundant contigs being removed (Robertson et al. 2010). For many phylogenetic studies, researchers are only interested in the most reliable transcripts, which will be used for further analyses, and such approaches are well suited in this case. However, if isoforms and splice variants are targets of the study, more refined methods should be used. Two widely used transcriptome assemblers for this task are OASES (Schulz et al. 2012) and TRINITY (Grabherr et al. 2011). Both these methods firstly reconstruct contigs using *k*-mer-based de Bruijn graphs and subsequently explore transcript variants by connecting contigs or by retrieving contigs which represent different paths through the graph, but share the same starting and end point. Assessment of the quality of de novo transcriptome assembly is less established as for de novo genome assemblies. The N50 can be still used as a measure, but as transcriptome assemblies usually represent a set of thousands of medium-sized contigs (transcripts), the ultimate goal is not to get as few and large contigs as possible. Some approaches to assess transcriptome assemblies are in use which do not depend on a closely related reference. Completeness of transcripts can be described using reference alignments with homologous genes and check for start codons and – if applicable – presence of signal peptides. Further on, all organisms rely on a core set of housekeeping genes which are generally expected to be expressed in most cells. For eukaryotes, a set of such core proteins is well established and can be automatically detected using the BUSCO pipeline (Simão et al. 2015). Missing genes of this set could indicate improper assembly or lack of sequencing depth. A similar approach is implemented in the software DOGMA, which performs a fast and easy quality assessment of transcriptome assemblies based on conserved protein domains (Dohmen et al. 2016). Other proposed quality metrics can be derived from read mapping and include descriptions of accuracy, fragmentation, incompleteness, redundancy or chimerism (Fig. 5.13), e.g. implemented in TRANSRATE (Smith-Unna et al. 2016). If a reference genome is available, several metrics can be inferred based on its comparison, e.g. completeness or contiguity (Martin and Wang 2011). However, the biggest hope for the future is the availability of low-error long reads originating from single transcripts which could solve the assembly problem in this field completely.

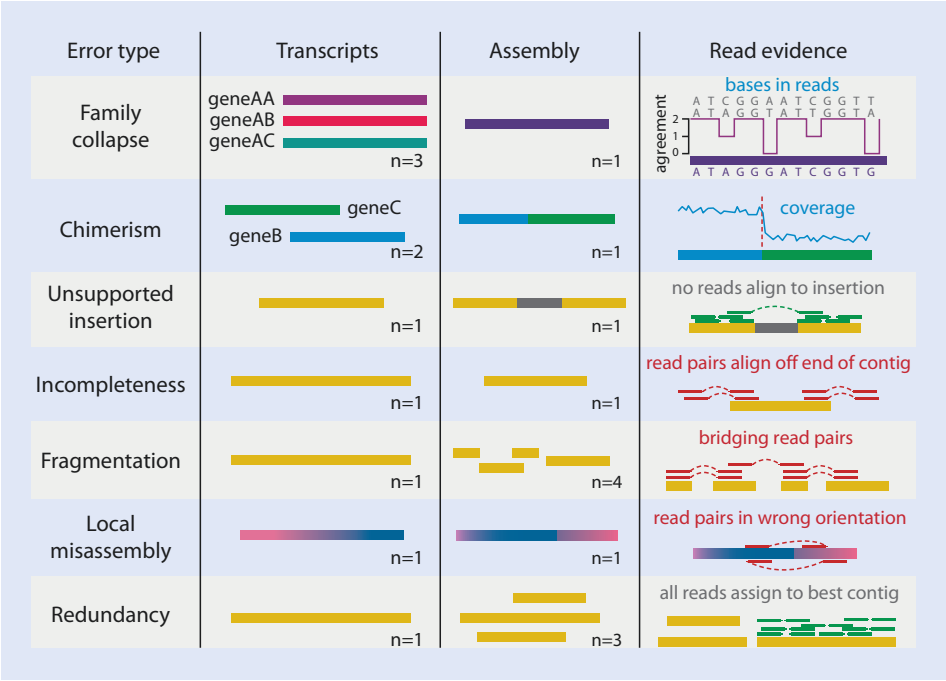


Fig. 5.13 Typical errors in transcriptome assemblies which can be assessed based on read evidence (Reprinted from (Smith-Unna et al. 2016))

The properties of metagenomic data are similar in some aspects to those of transcriptomes. Metagenomes comprise data originated from many different genomes from different individuals (Coughlan et al. 2015). As for transcriptomes, differences in abundance and therefore coverage through sequence reads are expected. Moreover, different organisms may harbour identical sequences, e.g. in the case of easily horizontally transmitted retrotransposons. Some assemblers are available which have been optimized for metagenome assembly, e.g. METAVELVET (Namiki et al. 2012) or METAIDBA (Peng et al. 2011). Additionally to the steps of «normal» *k*-mer-based genome assemblers, *k*-mer abundance information is used during assembly. It is assumed that sequences from different organisms differ in their coverage due to the individual abundance of the organisms in the sample. For example, highly abundant bacteria from a soil sample will be covered by more sequence reads than rare bacteria. This information can be used to refine the resolution of the *k-mer* graph. Nodes of these graphs which are included in the final contig should be joined by *k-mers* of similar coverage. A path through chimeric nodes, which represent the identical *k*-mers from different species, can be resolved according to this information (Fig. 5.14). The main problem of NGS-based metagenomic studies is to trace the origin of short reads back to different organisms (especially when there are no reference genomes available). In the future, techniques generating long-read data with lower error probabilities will be a key to enhance the accuracy of metagenomic assemblies.

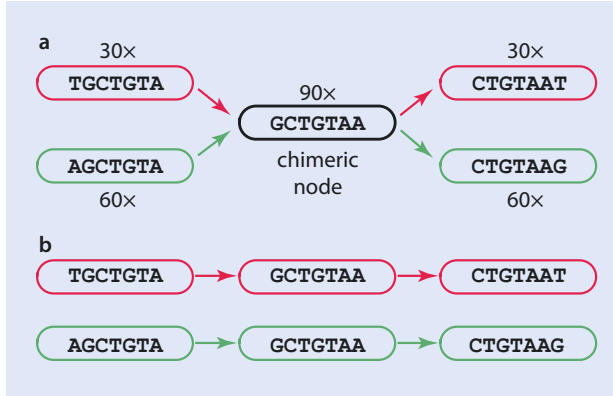


Fig. 5.14 Metagenomic assembly and chimeric nodes. **a** A *k*-mer assembly produces a chimer node, as indicated by the coverage information given by the number above the nodes. **b** Assembly of contigs can be resolved using the coverage information

5

References

Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA (2012) SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol 19:455–477

Bankevich A, Pevzner PA (2016) TruSPAdes: barcode assembly of TruSeq synthetic long reads. Nat Methods 13:248–250

Batzoglou S, Jaffe DB, Stanley K, Butler J, Gnerre S, Mauceli E, Berger B, Mesirov JP, Lander ES (2002) ARACHNE: a whole-genome shotgun assembler. Genome Res 12:177–189

Boetzer M, Henkel CV, Jansen HJ, Butler D, Pirovano W (2011) Scaffolding pre-assembled contigs using SSPACE. Bioinformatics 27:578–579

Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30:2114–2120

Bradnam K, Fass J, Alexandrov A, Baranay P, Bechner M, Birol I, Boisvert S, Chapman J, Chapuis G, Chikhi R, Chitsaz H, Chou W-C, Corbeil J, Del Fabbro C, Docking T, Durbin R, Earl D, Emrich S, Fedotov P, Fonseca N, Ganapathy G, Gibbs R, Gnerre S, Godzaridis E, Goldstein S, Haimel M, Hall G, Haussler D, Hiatt J, Ho I, Howard J, Hunt M, Jackman S, Jaffe D, Jarvis E, Jiang H, Kazakov S, Kersey P, Kitzman J, Knight J, Koren S, Lam T-W, Lavenier D, Laviolette F, Li Y, Li Z, Liu B, Liu Y, Luo R, MacCallum I, MacManes M, Maillet N, Melnikov S, Naquin D, Ning Z, Otto T, Paten B, Paulo O, Phillippy A, Pina-Martins F, Place M, Przybylski D, Qin X, Qu C, Ribeiro F, Richards S, Rokhsar D, Ruby J, Scalabrin S, Schatz M, Schwartz D, Sergushichev A, Sharpe T, Shaw T, Shendure J, Shi Y, Simpson J, Song H, Tsarev F, Vezzi F, Vicedomini R, Vieira B, Wang J, Worley K, Yin S, Yiu S-M, Yuan J, Zhang G, Zhang H, Zhou S, Korf I (2013) Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. GigaScience 2:10

Chang Z, Wang Z, Li G (2014) The impacts of read length and transcriptome complexity for *De Novo* assembly: a simulation study. PLoS One 9:e94825

Chevreux B, Pfisterer T, Drescher B, Driesel AJ, Müller WEG, Wetter T, Suhai S (2004) Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs. Genome Res 14:1147–1159

Chikhi R, Medvedev P (2014) Informed and automated k-mer size selection for genome assembly. Bioinformatics 30:31–37

Chin C-S, Alexander DH, Marks P, Klammer AA, Drake J, Heiner C, Clum A, Copeland A, Huddleston J, Eichler EE, Turner SW, Korlach J (2013) Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. Nat Methods 10:563–569

Cock PJA, Fields CJ, Goto N, Heuer ML, Rice PM (2010) The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic Acids Res 38:1767–1771

Coughlan L, Cotter P, Hill C, Alvarez-Ordóñez A (2015) Biotechnological applications of functional metagenomics in the food and pharmaceutical industries. Front Microbiol 6:672

David M, Dursi LJ, Yao D, Boutros PC, Simpson JT (2017) Nanocall: an open source basecaller for Oxford nanopore sequencing data. Bioinformatics 33:49–55

Dohmen E, Kremer LPM, Bornberg-Bauer E, Kemena C (2016) DOGMA: domain-based transcriptome and proteome quality assessment. Bioinformatics 32:2577–2581

Donmez N, Brudno M (2013) SCARPA: scaffolding reads with practical algorithms. Bioinformatics 29:428–434

Earl D, Bradnam K, St John J, Darling A, Lin D, Fass J, Hung On Ken Y, Buffalo V, Zerbino DR, Diekhans M, Ngan N, Ariyaratne PN, Sung W-K, Ning Z, Haimel M, Simpson JT, Fonseca NA, Birol I, Docking TR, Ho IY, Rokhsar DS, Chikhi R, Lavenier D, Chapuis G, Naquin D, Maillet N, Schatz MC, Kelley DR, Phillippy AM, Koren S, Yang S-P, Wu W, Chou W-C, Srivastava A, Shaw TI, Ruby JG, Skewes-Cox P, Betegon M, Dimon MT, Solovyev V, Seledtsov I, Kosarev P, Vorobyev D, Ramirez-Gonzalez R, Leggett R, MacLean D, Xia F, Luo R, Li Z, Xie Y, Liu B, Gnerre S, MacCallum I, Przybylski D, Ribeiro FJ, Yin S, Sharpe T, Hall G, Kersey PJ, Durbin R, Jackman SD, Chapman JA, Huang X, DeRisi JL, Caccamo M, Li Y, Jaffe DB, Green RE, Haussler D, Korf I, Paten B (2011) Assemblathon 1: a competitive assessment of de novo short read assembly methods. Genome Res 21:2224–2241

Ewing B, Green P (1998) Base-calling of automated sequencer traces using *Phred*. II. Error probabilities. Genome Res 8:186–194

Gao S, Sung W-K, Nagarajan N (2011) Opera: reconstructing optimal genomic scaffolds with high-throughput paired-end sequences. J Comput Biol 18:1681–1691

Goodwin S, Gurtowski J, Ethe-Sayers S, Deshpande P, Schatz MC, McCombie WR (2015) Oxford nanopore sequencing, hybrid error correction, and de novo assembly of a eukaryotic genome. Genome Res 25:1750–1756

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, Chen Z, Mauceli E, Hacohen N, Gnirke A, Rhind N, di Palma F, Birren BW, Nusbaum C, Lindblad-Toh K, Friedman N, Regev A (2011) Full-length transcriptome assembly from RNA-seq data without a reference genome. Nat Biotechnol 29:644–U130

Gurevich A, Saveliev V, Vyahhi N, Tesler G (2013) QUAST: quality assessment tool for genome assemblies. Bioinformatics 29:1072–1075

Hackl T, Hedrich R, Schultz J, Förster F (2014) *proovread*: large-scale high-accuracy PacBio correction through iterative short read consensus. Bioinformatics 30:3004–3011

Hernandez D, François P, Farinelli L, Østerås M, Schrenzel J (2008) *De novo* bacterial genome sequencing: millions of very short reads assembled on a desktop computer. Genome Res 18:802–809

Howison M, Zapata F, Dunn CW (2013) Toward a statistically explicit understanding of de novo sequence assembly. Bioinformatics 29:2959–2963

Howison M, Zapata F, Edwards EJ, Dunn CW (2014) Bayesian genome assembly and assessment by Markov chain Monte Carlo sampling. PLoS One 9:e99497

Huang X, Madan A (1999) CAP3: a DNA sequence assembly program. Genome Res 9:868–877

Hunt M, Newbold C, Berriman M, Otto T (2014) A comprehensive evaluation of assembly scaffolding tools. Genome Biol 15:R42

Kelley D, Schatz M, Salzberg S (2010) Quake: quality-aware detection and correction of sequencing errors. Genome Biol 11:R116

Kircher M, Heyn P, Kelso J (2011) Addressing challenges in the production and analysis of Illumina sequencing data. BMC Genomics 12:382

Kircher M, Stenzel U, Kelso J (2009) Improved base calling for the Illumina genome analyzer using machine learning strategies. Genome Biol 10:R83

Koren S, Schatz M, Walenz B, Martin J, Howard J, Ganapathy G, Wang Z, Rasko D, McCombie W, Jarvis E, Phillippy A (2012) Hybrid error correction and de novo assembly of single-molecule sequencing reads. Nat Biotechnol 30:693–700

Koren S, Walenz BP, Berlin K, Miller JR, Phillippy AM (2016) Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. bioRxiv. doi.org/10.1101/071282.

Laver T, Harrison J, O'Neill PA, Moore K, Farbos A, Paszkiewicz K, Studholme DJ (2015) Assessing the performance of the Oxford nanopore technologies MinION. Biomol Detect Quantif 3:1–8

Li D, Luo R, Liu C-M, Leung C-M, Ting H-F, Sadakane K, Yamashita H, Lam T-W (2016) MEGAHIT v1.0: a fast and scalable metagenome assembler driven by advanced methodologies and community practices. Methods 102:3–11

Li H (2016) Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. Bioinformatics 32:2103–2110

Li R, Fan W, Tian G, Zhu H, He L, Cai J, Huang Q, Cai Q, Li B, Bai Y, Zhang Z, Zhang Y, Wang W, Li J, Wei F, Li H, Jian M, Nielsen R, Li D, Gu W, Yang Z, Xuan Z, Ryder O, Leung F-C, Zhou Y, Cao J, Sun X, Fu Y (2010) The sequence and de novo assembly of the giant panda genome. Nature 463:311–317

Lin Y, Yuan J, Kolmogorov M, Shen MW, Pevzner PA (2016) Assembly of long error-prone reads using de Bruijn graphs. Proc Natl Acad Sci USA 113:E8396-E8405 (In press)

Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, He G, Chen Y, Pan Q, Liu Y (2012) SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. GigaScience 1:18

MacCallum I, Przybylski D, Gnerre S, Burton J, Shlyakhter I, Gnirke A, Malek J, McKernan K, Ranade S, Shea TP, Williams L, Young S, Nusbaum C, Jaffe DB (2009) ALLPATHS 2: small genomes assembled accurately and with high continuity from short paired reads. Genome Biol 10:R103

Marçais G, Kingsford C (2011) A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics 27:764–770

Martin JA, Wang Z (2011) Next-generation transcriptome assembly. Nat Rev Genet 12:671–682

Miller J, Koren S, Sutton G (2010) Assembly algorithms for next-generation sequencing data. Genomics 95:315

Myers EW, Sutton GG, Delcher AL, Dew IM, Fasulo DP, Flanigan MJ, Kravitz SA, Mobarry CM, Reinert KHJ, Remington KA, Anson EL, Bolanos RA, Chou H-H, Jordan CM, Halpern AL, Lonardi S, Beasley EM, Brandon RC, Chen L, Dunn PJ, Lai Z, Liang Y, Nusskern DR, Zhan M, Zhang Q, Zheng X, Rubin GM, Adams MD, Venter JC (2000) A whole-genome assembly of *Drosophila*. Science 287:2196–2204

Nagarajan N, Pop M (2013) Sequence assembly demystified. Nat Rev Genet 14:157–167

Nakamura K, Oshima T, Morimoto T, Ikeda S, Yoshikawa H, Shiwa Y, Ishikawa S, Linak MC, Hirai A, Takahashi H, Altaf-Ul-Amin M, Ogasawara N, Kanaya S (2011) Sequence-specific error profile of Illumina sequencers. Nucleic Acids Res 39:e90

Namiki T, Hachiya T, Tanaka H, Sakakibara Y (2012) MetaVelvet: an extension of velvet assembler to de novo metagenome assembly from short sequence reads. Nucleic Acids Res 40:e155

Peng Y, Leung HCM, Yiu S-M, Chin FYL (2010) IDBA—a practical iterative de Bruijn graph de novo assembler. In: Berger B (ed) Research in computational molecular biology, vol 6044. Springer, Berlin, pp 426–440

Peng Y, Leung HCM, Yiu S-M, Lv M-J, Zhu X-G, Chin FYL (2013) IDBA-Tran: a more robust de novo de Bruijn graph assembler for transcriptomes with uneven expression levels. Bioinformatics 29:326–334

Peng Y, Leung HCM, Yiu SM, Chin FYL (2011) Meta-IDBA: a de Novo assembler for metagenomic data. Bioinformatics 27:i94–i101

Pevzner P, Tang H, Waterman M (2001) An Eulerian path approach to DNA fragment assembly. Proc Natl Acad Sci 98:9748–9753

Renaud G, Kircher M, Stenzel U, Kelso J (2013) freeIbis: an efficient basecaller with calibrated quality scores for Illumina sequencers. Bioinformatics 29:1208–1209

Robertson G, Schein J, Chiu R, Corbett R, Field M, Jackman SD, Mungall K, Lee S, Okada HM, Qian JQ, Griffith M, Raymond A, Thiessen N, Cezard T, Butterfield YS, Newsome R, Chan SK, She R, Varhol R, Kamoh B, Prabhu A-L, Tam A, Zhao Y, Moore RA, Hirst M, Marra MA, Jones SJM, Hoodless PA, Birol I (2010) *De novo* assembly and analysis of RNA-seq data. Nat Methods 7:909–912

Salmela L, Rivals E (2014) LoRDEC: accurate and efficient long read error correction. Bioinformatics 30:3506–3514

Salzberg S, Phillippy A, Zimin A, Puiu D, Magoc T, Koren S, Treangen T, Schatz M, Delcher A, Roberts M, Marcais G, Pop M, Yorke J (2012) GAGE: a critical evaluation of genome assemblies and assembly algorithms. Genome Res 22:557–567

Schatz MC, Delcher AL, Salzberg SL (2010) Assembly of large genomes using second-generation sequencing. Genome Res 20:1165–1173

Schulz MH, Zerbino DR, Vingron M, Birney E (2012) *Oases*: robust de novo RNA-seq assembly across the dynamic range of expression levels. Bioinformatics 28:1086–1092

Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM (2015) BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31:3210–3212

Simpson J, Wong K, Jackman S, Schein J, Jones S, Birol I (2009) ABySS: a parallel assembler for short read sequence data. Genome Res 19:1117–1123

Smeds L, Kunstner A (2011) CONDETRI - A content dependent read trimmer for Illumina data. PLoS One 6:e26314

References

Smith-Unna R, Boursnell C, Patro R, Hibberd J, Kelly S (2016) TransRate: reference free quality assessment of de novo transcriptome assemblies. Genome Res 26:1134–1144

Sović I, Križanović K, Skala K, Šikić M (2016) Evaluation of hybrid and non-hybrid methods for de novo assembly of nanopore reads. Bioinformatics 32:2582–2589

Wang Z, Gerstein M, Snyder M (2009) RNA-seq: a revolutionary tool for transcriptomics. Nat Rev Genet 10:57–63

Zerbino D, Birney E (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res 18:821–829

Alignment and Mapping

C. Bleidorn, *Phylogenomics*, DOI 10.1007/978-3-319-54064-1_6

- Alignments are hypotheses of positional homologies between nucleotides or amino acids of sequences.
- The Needleman and Wunsch algorithm finds the optimal pairwise alignments of two sequences, which can contain matches, mismatches and gaps.
- Local alignments optimize the positional homology for substrings of sequences and are widely used in database searches.
- Multiple sequence alignments can only be retrieved using heuristic approaches, e.g. progressive alignments.
- Alignment masking is the exclusion of unreliably aligned positions to improve the signal-to-noise ratio of the data.
- Mapping of sequence reads to reference sequences is a special case of alignments; most mapping algorithms are either based on a seed-and-extend approach or Burrows-Wheeler transform-related methods.

6.1 Pairwise Alignment

Homology is broadly defined as a character that arises as a result of common ancestry (Thornton and DeSalle 2000). Homologies can be hypothesized at different levels. For phylogenomics it is important to establish the homology of genes (or genomic regions), but also the homology of nucleotide or amino acid positions within genes (or genomic regions). Alignments are hypotheses of positional homologies between the nucleotides or amino acids of sequences (Rosenberg 2009) and can be either global or local (Phillips et al. 2000). In a global alignment, positional homology across all positions of two aligned sequences is determined. Global alignments are used for phylogenetic analyses or to detect patterns of selection. In contrast, for local alignments positional homology is optimized only for fragments (substrings) of two sequences. Local alignments are widely used for database searches as, for example, implemented in the BLAST algorithm (see below). In general, it is possible to align any two sequences and there are many possibilities to do this. To compare different sequence alignments, it is necessary to use a metric to estimate the quality of each alignment.

In an alignment the horizontal rows are sequences, whereas the vertical rows represent characters which refer to positions in a sequence. The residues of the sequence itself are used as character states. There are four different character states for nucleotide sequences and 20 different character states for amino acid sequences. If a character of two aligned sequences shows the same character state, it is called a match, whereas the presence of different character states within a character is called mismatch (Fig. 6.1). Additionally it is

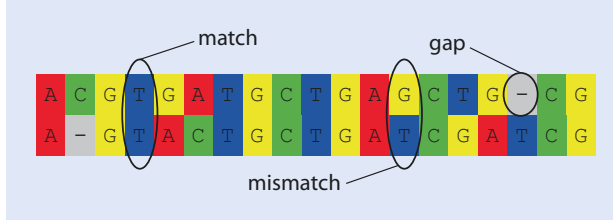


Fig. 6.1 Global alignment of two sequences. Matched base pairs, mismatches and gaps are exemplified. Scoring matches with +1 and mismatches and gaps with −1 results in a total score of +3

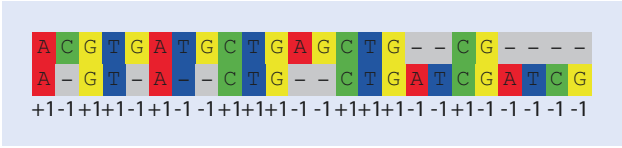


Fig. 6.2 A different pairwise sequence alignment with the same sequences as in Fig. 6.1. Scores for matches and gaps are given below each sequence position; the combined score is 0

possible that gaps are inserted into alignments (Fig. 6.1). These gaps represent either events of insertions in one sequence or a deletion in the other sequence. Often it is neither simple nor necessary to determine which of the events took place, and they are together summarized as indels (Simmons and Ochoterena 2000).

To estimate the quality of a pairwise alignment, a simple score can be developed, where the number of matched base pairs is scored as a benefit, whereas the number of mismatches and gap positions induces costs. Generally, the goal is to maximize the benefits while minimizing the costs. For example, scores could be arbitrarily set as follows: match +1, mismatch −1 and gap −1. The alignment in Fig. 6.1 has ten matches, five mismatches and two gap positions, which results into a score of +3. An alternative alignment of the same two sequences is given in Fig. 6.2. This alignment only contains matches and gap positions. Even though the number of matches is higher than the alignment given in Fig. 6.1, the total score of 0 is lower. Comparing these two alignments, the alternative in Fig. 6.1 would be chosen as the better one, as it has the higher score given our introduced scoring system.

Of course the used scoring system is arbitrary, and a different one may support the choice of an alternative alignment. Especially the scoring of gap characters has been debated (Giribet and Wheeler 1999). Gaps have obviously to be introduced when aligning two sequences of different lengths. Gaps are resulting from a different biological process than mismatches. Whereas mismatches (mostly) trace back to mutations, gaps are the result of indels. Possible mechanism for indels are errors during DNA replication (e.g. slipped-strand mispairing), unequal crossing over during recombination or introduction of mobile elements (McGuffin 2009; Levinson and Gutman 1987). All these mechanisms usually result in the simultaneous insertion (or deletion) of sequences, which implies that multiple neighbouring gaps stem from a single event. Using a scoring system that treats all gaps independently would therefore introduce an over-penalization for them, as implicitly separate events would be assumed for their origin (McGuffin 2009). As a solution to this problem, the use of affine gap costs has been introduced. This type of penalty differentiates between opening a gap and extending it. For example, using gap opening costs of −1 and gap extension costs of 0.1 for Fig. 6.2 would result in a total score of +5.4, whereas the alignment in Fig. 6.1 remains at +3. Similarly, it is possible to introduce different scores for mismatches. For example, in case of aligning protein sequences, scores are usually based on matrices that incorporate the evolutionary preferences for certain substitutions over other kinds of substitutions. Widely used matrices are BLOSUM and PAM (Henikoff and Henikoff 1992). Figure 6.3 shows the BLOSUM62 matrix (Henikoff and Henikoff 1992), which is used by all BLAST searches (see below) on an amino acid level. Scores in these matrices are given as log-odds, which can be directly used as parameters of alignment scoring schemes. Positive scores mean that we find amino acid pairings more often than expected by chance (conservative substitutions); negative values indicate those

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
Ala (A)	4																			
Alg (R)	-1	5																		
Asn (N)	-2	0	6																	
Asp (D)	-2	-2	1	6																
Cys (C)	0	-3	-3	-3	9															
Gln (Q)	-1	1	0	0	-3	5														
Glu (E)	-1	0	0	2	-4	2	5													
Gly (G)	0	-2	0	-1	-3	-2	-2	6												
His (H)	-2	0	1	-1	-3	0	0	-2	8											
Ile (I)	-1	-3	-3	-3	-1	-3	-3	-4	-3	4										
Leu (L)	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4									
Lys (K)	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	6								
Met (M)	-1	-2	-2	-3	-1	0	-2	-3	-2	1	2	-1	5							
Phe (F)	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6						
Pro (P)	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7					
Ser (S)	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4				
Thr (T)	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5			
Trp (W)	-3	-3	-4	-3	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11		
Tyr (Y)	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-2	-2	-2	2	7	
Val (V)	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4

Fig. 6.3 BLOSUM62 matrix giving log-odd scores for each possible amino acid substitution derived from pairwise sequence alignments of at least 62% identity

occurring less often as expected (non-conservative substitutions) (Eddy 2004). Alternatively, a matrix counting the steps for amino acid substitutions inferred from the genetic code can be implied. In this case, costs are either −1 (one change in the codon triplet needed), −2 (two changes needed) or −3 (three changes needed). Obviously, choice of the scoring function and its parameters has a huge influence on selecting the best pairwise alignment.

To find the best optimal alignment, it would be necessary to compare all possible pairwise alignments, which can be a giant number given that it grows faster than exponentially with increasing sequence length. However, a simple solution finding the optimal pairwise global alignment was published by Needleman and Wunsch (1970). This method is a dynamic programming approach, where to solve a complex problem is broken down into more simple and thereby easy-to-solve subproblems (Cooper and Cooper 1981). The Needleman and Wunsch algorithm consists of three steps: matrix initialization, matrix filling and traceback. In the first step, a matrix is initialized, containing the two sequences along an axis (Fig. 6.4). Additionally, an empty row is added to the top and an empty column to the left of the matrix. Next, a zero is placed in the upper left corner, and the top

		A	C	G	T	G	A	T	G
	0	← -1	← -2	← -3	← -4	← -5	← -6	← -7	← -8
A	↑-1								
G	↑-2								
T	↑-3								
A	↑-4								
C	↑-5								
T	↑-6								
G	↑-7								
C	↑-8								

Fig. 6.4 Initialization of a matrix for the Needleman and Wunsch (1970) global alignment algorithm. The first row and column are filled with increasing multiples of the gap cost and arrows pointing to 0

row and the left column are filled with increasing multiples of the costs for a gap. Moreover, arrows are introduced into each of these cells, pointing to the zero in the upper left.

The second step is filling the matrix. Given the chosen scoring system, three values are calculated for every single cell (Fig. 6.5): match/mismatch score, vertical gap score and horizontal gap score. The match/mismatch cost (M) equals the sum of the value of the cell that is diagonally to the upper left plus costs for a match or mismatch (whatever applies). The horizontal gap score equals the sum of the value of the cell to the left plus the gap score. The vertical gap score equals the sum of the value of the cell above it plus the gap score. The highest value is chosen to fill the box, and an arrow indicates where it comes from. In the case of equally high values, multiple arrows can be introduced or one of the solutions is chosen randomly.

The last step is the traceback, where starting with field in the bottom right, the path of the arrows is followed to the upper left (Fig. 6.6). Following a diagonal arrow means that residues from the row and the column of this field should be aligned. In case of following a vertical arrow, a residue of the vertical (upper sequence) is aligned with a gap, whereas in the case of following a horizontal arrow, the gap is placed in the other sequence. If multiple arrows were introduced during the filling, equally optimal alignments can be retrieved.

The dynamic programming for global alignments can be formalized (when using linear gap costs) as a recursion, where the maximal score *F(i,j)* is calculated between the first i residues of sequence X and the first j residues of sequence Y. The recursion of the Needleman and Wunsch (1970) algorithm looks as follows:

$$F(i,j) = \max \begin{cases} F(i-1,j-1) + s(X_i, Y_j) \\ F(i-1,j) - g \\ F(i,j-1) - g \end{cases} \quad (6.1)$$

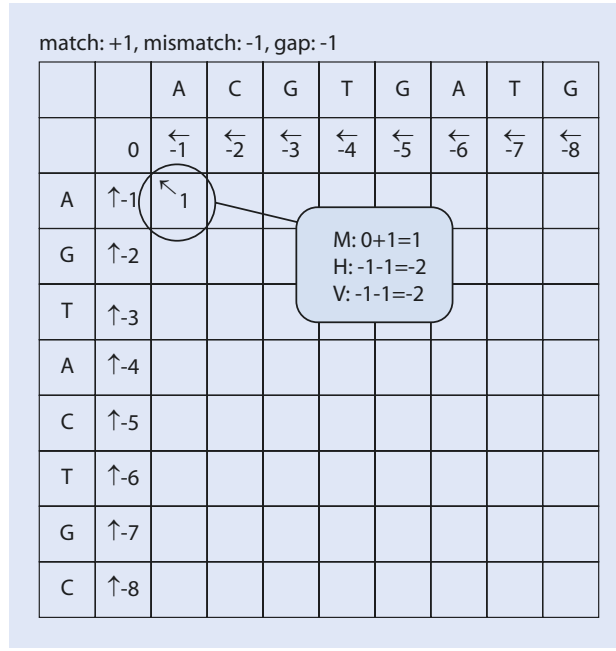


Fig. 6.5 Filling of the matrix. Using the chosen scoring system, three values are calculated for each box. The match/mismatch cost (M) is the sum of the value of the cell that is diagonally to the upper left (0) plus costs for a match (in this example, it can be also mismatch) (+1) which totals +1. The horizontal gap score is the sum of the value of the cell to the left (−1) plus the gap score (−1), which totals −1. The vertical gap score is the sum of the value of the cell above it (−1) plus the gap score (−1). The highest value is chosen to fill the box, and an *arrow* indicates where it comes from

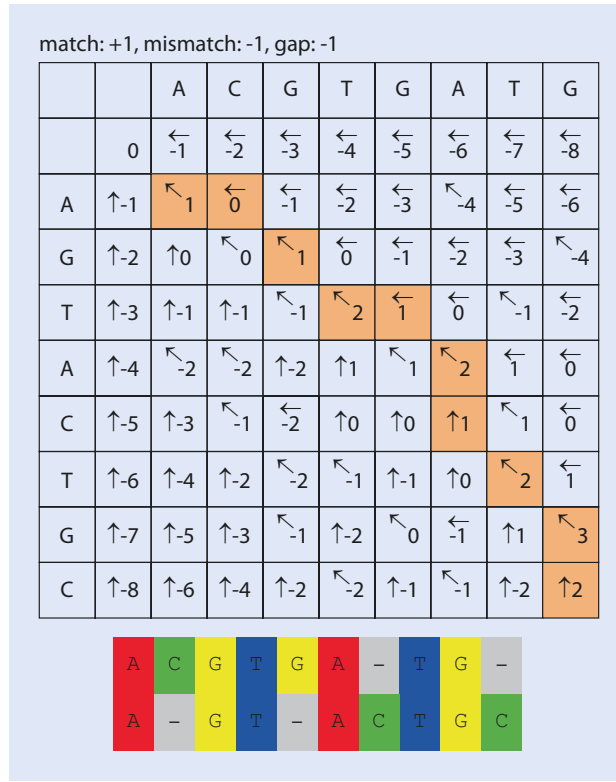


Fig. 6.6 The traceback uncovers the (or one) optimal alignment. Starting with field in the bottom right, the path of the arrows is followed to the upper left. The arrows indicate if bases should be matched (*diagonal arrows*), gaps should be included in the upper sequence (*arrow pointing upwards*) or gaps should be introduced in the sequence to the left (*arrow pointing left*)

First, the score for the last alignment column is calculated, which is either *s(Xi,Yj)* in case of matching (mismatching) base pairs or *–g* when a gap is included in either of the sequences. The score of each of the remaining alignment columns is *F(i,j)*, *F(i-1,j)* or *F(i,j-1)*, depending on which of the alternatives applies. The score of the optimal alignment is the sum of the scores of the alignment columns (Morgenstern 2009). Many computational tools for pairwise sequence alignment are available. For example, an online tool for DNA and protein alignments based on this algorithm can be found on the website of the EMBL-EBI (► http://www.ebi.ac.uk/Tools/psa/) (Rice et al. 2000).

6.2 Local Alignment and BLAST Searches

Local alignments can be used to find similarities (and putative homologies) between fragments (substrings) of two sequences. Typical applications are database searches to retrieve most similar sequences (sequence fragments) for the input sequence. At the beginning, the task for a local pairwise alignment looks to be more complex as for global pairwise alignments, as it basically means performing many different global alignments for different starts and ends of the compared substrings. Fortunately, Smith and Waterman (1981) proposed a computationally easy solution for this problem based on an adaptation of the Needleman and Wunsch (1970) algorithm (NWA). Similarly to the latter algorithm, a matrix is created based on the length of the sequences, and all cells are filled based on a scoring system (■ Fig. 6.7). However, the extra row and column directly at the upper and

match: +1, mismatch: -1, gap: -1

		A	T	G	C	T	T	A	A
	0	0	0	0	0	0	0	0	0
A	0	0	0	0	1	0	0	0	0
C	0	0	0	0	1	0	0	0	0
T	0	0	1	0	0	↖2	1	0	0
G	0	0	0	1	0	0	↖1	0	0
A	0	1	0	0	0	0	0	↖2	←1
T	0	0	↖2	←1	0	0	0	0	↖1
G	0	0	0	↖3	←2	←1	0	0	0
C	0	0	0	↑2	↖4	←3	←2	←1	0

■ **Fig. 6.7** Completed matrix using the Smith and Waterman (1981) algorithm. The traceback starts at the highest value of the matrix and only substrings of the sequences are aligned

left border of the matrix are now filled with zeros. During the fill-in, the cells of the matrix are filled with the same rules as in the NWA, with the exception that always when a negative value is calculated, the cell is filled with a zero instead. Moreover, arrows are only assigned in case they point towards a positive value. The final traceback starts at the highest values within the matrix, following the arrows till a zero is reached (Fig. 6.7).

The computation time of the Smith and Waterman (1981) algorithm (SWA) grows linearly with the product of the length of the two compared sequences (Cristianini and Hahn 2007). While this is a relatively fast approach, it is still computationally too resource intensive for standard database search applications. For example, a common task is to find the most similar sequences of a given query in a public database. Usually, all published sequence data are stored in one of the three main databases: NCBI GenBank, EMBL-Bank of the European Molecular Biology Laboratory or in the DNA Data Bank of Japan (DDBJ) (Pevsner 2015). All these databases share their data daily. NCBI GenBank is hosted by the National Institutes of Health (NIH) in the USA, which keeps an annotated collection of all publicly available DNA sequences (Benson et al. 2013). In February 2016 (Release 212.0), GenBank comprised 207,018,196,067 bases from 190,250,235 reported sequences in its sequence database. Additionally, billions of sequences from NGS high-throughput platforms are stored in the sequence read archive. Faster database search algorithms are needed to handle these huge numbers of sequences. Two prominent algorithms which have been developed are FASTA and BLAST. Both methods use heuristics to identify regions of high similarity before calculating pairwise alignment scores. FASTA (Lipman and Pearson 1985) is nowadays mostly known for the underlying sequence format, which became a standard in molecular sequence analyses. However, the by far most popular method to search in extremely large databases is the BLAST algorithm (Altschul et al. 1990). BLAST is an acronym for Basic Local Alignment Search Tool. In contrast to dynamic programming, it does not guarantee to find the optimal alignment, as it uses a heuristic approach. However, it is by two orders of magnitudes faster than the Smith and Waterman algorithm, which is achieved by only searching within the sequence space of high similarity.

BLAST searches start by finding all words (*k*-mers) of a length *k* (typically 3 for amino acids and 11 for nucleotides), which exist in the query sequence (Fig. 6.8a). Additionally, based on a substitution matrix, similar high-scoring words (neighbourhood words) are listed for each word of the query matrix. For example, in the example in Fig. 6.8, the word LEH is derived from the query sequence. Similar words from its «neighbourhood» are aligned (e.g. LKH, CEH, QEH, etc.) and ordered according to its alignment score as calculated by using a substitution matrix. For amino acid substitutions, the BLOSUM62 matrix (Fig. 6.3) is usually used. A list of all words retrieved by this procedure is stored, and exact matches of these words in the database sequences are searched for (Fig. 6.8b). Every match is called a «high-scoring sequence pair» (HSP), which is used as «seed» for local sequence alignment (Fig. 6.8c). The alignment is extended to the left and the right of the seed, and the alignment score is calculated after every extension based on the substitution matrix. The algorithm stops extending the alignment once the score decreases by a fixed value X from the maximum score found at any point during alignment. The final score for each local alignment is kept, and all alignments with a score below a threshold value S are discarded. BLAST has been initially developed for un-gapped alignments (Altschul et al. 1990), but is also available for alignments including gaps (Altschul et al. 1997).

Based on the type of query sequence and the type of chosen database, there are five variations of BLAST searches. BLASTN uses nucleotide sequences as query to search within a nucleotide database. BLASTP uses amino acid sequences as query to search within a

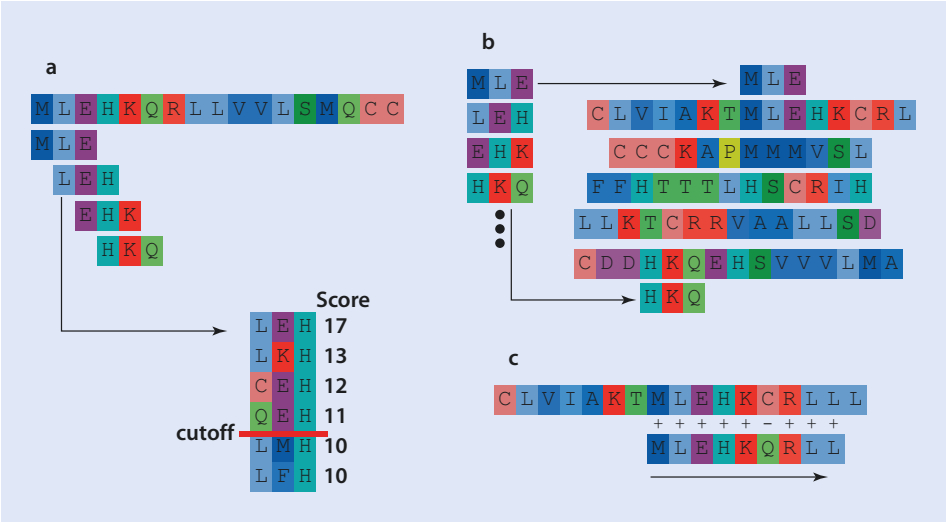


Fig. 6.8 Schematic representation of the workflow of the BLAST algorithm. **a** A list of words is derived from a query sequence. Additionally, for each word high-scoring similar words based on a distance matrix (e.g. BLOSUM62 for amino acids) are stored. **b** The complete list of words is used to find exact matches in the database sequences. **c** Starting from the exact match, the alignment is extended (in both directions) to find alignments with scores above a given threshold

protein database. For BLASTX a nucleotide query is translated in all six possible reading frames to be compared with a protein database. TBLASTN compares a protein query against a nucleotide database which is translated in all six reading frames. And TBLASTX uses a nucleotide query translated in all six reading frames to compare it on the amino acid level against a nucleotide database, which is also translated in all six reading frames. When targeting protein-coding genes, it is in most cases advisable to use a BLAST algorithm that compares sequences on an amino acid level. The BLAT algorithm is an alignment tool similar to BLAST (Kent 2002). It can be used to search genome assemblies for sequences of high similarity. BLAT of DNA is designed to quickly find sequences with a similarity greater than 95% and of a length of 40 bases or more. Therefore it is commonly used to identify the location of a sequence in the genome or determine the exon structure of an mRNA.

Hits retrieved from BLAST searches can be ordered according to their alignment score or using an expectation value (e-value). The alignment score is calculated based on pairwise alignments of the retrieved similar sequence fragments. The e-value describes the number of hits we would expect to find by chance (Cristianini and Hahn 2007). The lower the number, the more significant is the hit. For example, an e-value of 1 means that at least one hit of similar sequence length and sequence similarity is to be expected by chance. Because the size of the database itself is included in the calculation, very short sequences will always have quite high e-values, even if there are 100% identical hits in the database. BLAST searches can be performed by stand-alone software using a command line or in a browser by several web applications (e.g. ▶ http://blast.ncbi.nlm.nih.gov/Blast.cgi). The output of BLAST searches are lists of sequences, ordered by similarity. Typical output formats are pairwise outputs including alignments (Fig. 6.9a) or tabular outputs (Fig. 6.9b). The latter are more practical for phylogenomic applications, as all relevant information can be easily parsed using scripts or command line tools.

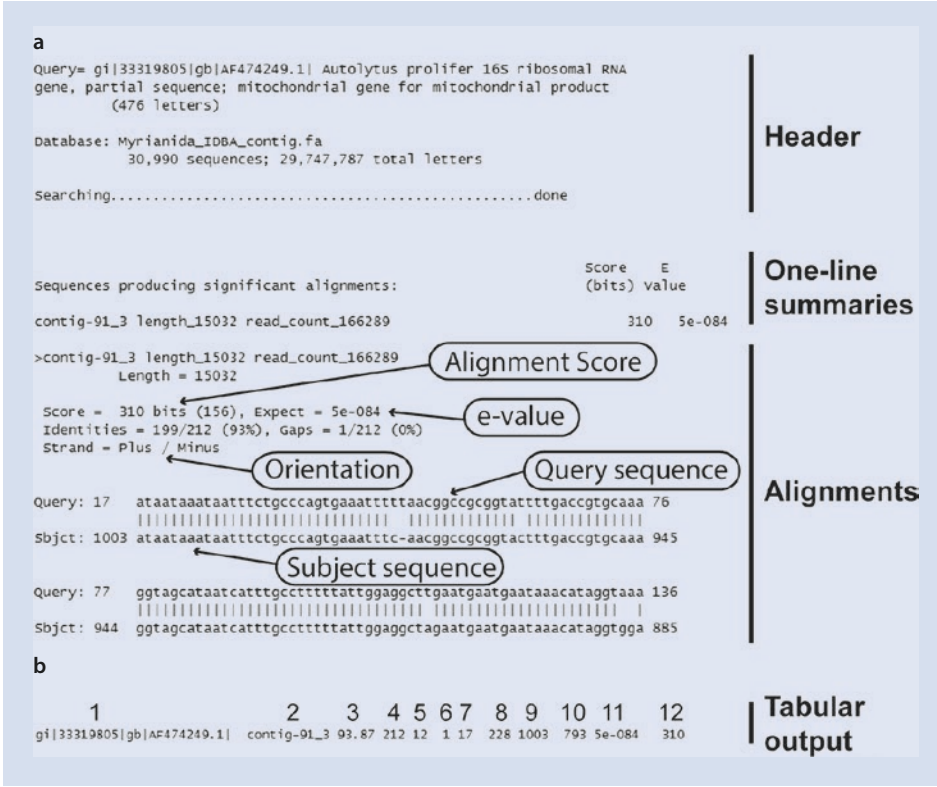


Fig. 6.9 Typical output of local BLAST searches. **a** Pairwise output consisting of header, one-line summary and alignments. **b** Tabular output in 12 columns: 1, query sequence ID; 2, subject sequence ID; 3, percent identity; 4, alignment length; 5, mismatches; 6, gaps; 7, query sequence alignment start; 8, query sequence alignment end; 9, subject sequence alignment start; 10, subject sequence alignment end; 11, e-value; 12, alignment score

6.3 Multiple Sequence Alignment

Alignments of more than two sequences are needed to resolve phylogenies of genes or species. Principally, the Needleman and Wunsch algorithm introduced in 6.1 could be extended to the problem of multiple sequence alignments (MSAs) (Chan et al. 1992). In this case the matrix would become multidimensional, and the algorithm would work successively through each dimension. This approach is an exhaustive method and would guarantee finding an optimal alignment. However, the costs in terms of computation time increase exponentially with the number of sequences and sequence length, thereby limiting the usefulness of such an approach to cases with very few sequences (Edgar and Batzoglou 2006). Instead, heuristic approaches with reduced computational time are normally used for MSA. The most popular approach is known as progressive alignment, developed by Feng and Doolittle (1987). This approach decomposes MSA into a series of pairwise sequence alignment operations. Using a phylogenetic guide tree (e.g. a neighbour-joining tree based on the pairwise distances derived from pairwise alignments), the MSA is constructed by adding sequences individually. Each node of the guide tree represents a separate pairwise alignment, and the most similar sequences are added first, and more

distant sequences are added gradually (Phillips et al. 2000). By using this approach, incongruent placement of gaps in pairwise alignments can severely affect the quality of the corresponding MSA. Several methods performing an iterative refinement have been developed to correct placement of inconsistent gap positions and other problems in the final MSA.

A phylogeny-aware method treating indels as evolutionary distinct events was developed to increase alignment quality (Löytynoja and Goldman 2008). However, this method seems to be rarely used for phylogenomic studies, and alignment lengths are usually greatly inflated by excessively introducing gaps. Whereas recent phylogenomic analyses are mostly based on protein-coding genes (and therefore amino acid alignments), phylogenetic studies using single genes are often performed using the small (eukaryotes) or large (Bacteria and Archaea) ribosomal subunit, which are structural genes exhibiting a secondary (and tertiary) structure (Cole et al. 2009). For this case, several alignment programs using secondary structure models of ribosomal genes to guide the alignment are available (Gardner et al. 2005).

The most widely used software for MSA has been CLUSTAL W, which was introduced in the mid-1990s (Thompson et al. 1994). However, several newer alignment programs are not only faster, but often also more accurate: MAFFT (Katoh and Standley 2013), MUSCLE (Edgar 2004) or T-COFFEE (Notredame et al. 2000), to name the most popular. Some benchmark datasets based on alignments of different complexity (domain organization, mixture of conserved and non-conserved regions) have been constructed and used to test the speed and accuracy of different aligners (Thompson et al. 2005; Thompson et al. 2011). Generally, the tested alignment programs work well. However, often different programs excel for different problems. For example, some programs are better suited to align conserved sequence blocks, whereas others are better in aligning strongly diverging sequences (Thompson et al. 2011).

6.4 Alignment Masking

For sequence alignments it is not unusual that some regions are aligned with more confidence than others. For example, protein-coding genes often comprise one or more conserved domains which are easier to align than flanking regions. In the case of ribosomal genes, conserved regions and more variable expansion regions differ in their degree of variability and thereby in the confidence how regions can be aligned. Different alignments mean different hypotheses of positional homology, and it is long known that this can affect the resulting phylogenies (Morrison and Ellis 1997; Thorne and Kishino 1992). Likewise, also other estimates as, for example, model parameters or inference of positive selection might be heavily influenced by the accuracy of the underlying alignment (Privman et al. 2012; Wong et al. 2008). As already mentioned every set of characters can be aligned somehow. Alignments of random data have been shown to bear phylogenetic signal resulting in supported tree topologies (Hillis and Huelsenbeck 1992). Moreover, different alignment methods seem to differ in their bias of creating artificial phylogenetic resolution from random sequence data (Simmons et al. 2010). Furthermore, several similar optimal solutions to the recovered alignment exist. In case of multiple alignments using heuristics, it is not even guaranteed to find the optimal solution. Not surprisingly, the most used alignment algorithms differ in ~20% of the aligned positions when aligning the same set of sequences in normal and reverse order (Landan and Graur 2007). In summary, difficult

sequence alignments will usually always contain parts of ambiguous positions and random similarity. As a solution to all these problems, several studies proposed to mask and exclude unreliably aligned positions of sequence alignments and thereby improve the signal-to-noise ratio of the data.

Initially, alignment masking has been often performed manually, which, however, has been strongly criticized as irreproducible (Landan and Graur 2007) and is also not possible when dealing with hundreds of genes. Several programs for automatic and reproducible alignment masking have been published, and some of the most widely used are GBLOCKS (Castresana 2000; Talavera and Castresana 2007), SOAP (Löytynoja and Milinkovitch 2001), AL2CO (Pei and Grishin 2001), MUMSA (Lassmann and Sonnhammer 2005), ALISCORE (Misof and Misof 2009), GUIDANCE (Penn et al. 2010a) and ZORRO (Wu et al. 2012).

GBLOCKS was one of the first available alignment maskers and is still widely used. By calculating the degree of conservation of every single alignment position, conserved «blocks» are identified. These «blocks» are retained for further analyses based on a set of rules that can be modified by the user. For example, a higher number of gap positions are allowed or poorly conserved regions which are flanked by conserved ones can be kept. Even though GBLOCKS has been criticized for using arbitrary rules without theoretical justification, comparisons with other alignment maskers based on simulated datasets show that this software works well when parameters are carefully chosen (Kück et al. 2010; Talavera and Castresana 2007). In contrast, other alignment maskers explicitly use hidden Markov models or resampling techniques to identify noisy alignment positions for exclusion. ALISCORE uses parametric Monte Carlo resampling to identify positions with random signal in multiple sequence alignments. Therefore, an expected similarity score is generated for pairwise alignments of randomized sequences within a sliding window. In the case of nucleotide data, a scoring function based on matches and mismatches is used to generate the similarity score, whereas for amino acid, data scores of randomized sequences are derived from an empirical matrix (e.g. BLOSUM62, see above). For nucleotide sequences scores are adapted to varying base composition along sequences and among sequences, whereas for amino acid data, this is only calculated once based on the composition of the original data. For the defined sliding window (e.g. 5 bps) of the original alignment, the observed score is calculated as the sum of all single-position comparisons, thereby calculating scores for all sequence pairs. Finally, the observed score of the selected window of the sequence alignment is compared with the expected score from randomized sequences. For this comparison a frequency distribution of random scores is generated, where randomness is assumed if the observed score fails to be better than 95% of the scores from the random sequences, generated by Monte Carlo resampling (Kück et al. 2010; Misof and Misof 2009). ZORRO measures the quality of each individual alignment position by using a pair-hidden Markov model (pair-HMM) (Wu et al. 2012). Using this approach the quality of two aligned residues is estimated in the context of all possible pairwise alignments. The rational of the ZORRO algorithm is that if two residues are truly homologous, they should also align in most of the alternative pairwise alignments. Using pair-HMM (Bradley et al. 2009), the posterior probability of two positions being aligned in all possible alignments is calculated. If the posterior probability is close to 1, the alignment of this position is highly reliable, whereas a posterior probability close to 0 identifies ambiguous positions. To assess confidence for positions of multiple sequence alignments, a weighted sum of pairs scheme to sum up the posterior probability of all pairs in the column is calculated (Wu et al. 2012). All alignment positions with a confidence score under

a certain threshold (e.g. >0.95) are excluded. GUIDANCE is a method where alignment uncertainty is calculated by comparing alignment positions across bootstrapped guide trees (Penn et al. 2010a; Penn et al. 2010b). This is based on the idea that the guide tree, which is used by progressive alignment methods (see above), introduces uncertainty. For example, different guide trees will lead to different multiple sequence alignments. By using a simple bootstrapping approach, multiple guide trees are generated and used for alignment. Finally, the occurrence of every single position of the original alignment is inspected in the alignments from the perturbated trees. As more often a position occurs in the alternative alignments, it is regarded to be more reliable. According to a user-defined value, unreliable positions are discarded. Alignments seem to be especially unreliable for sequence regions containing many gaps. In an updated version called GUIDANCE2, different gap opening costs are used to create further alternative alignments which are inspected regarding consistency of every single alignment position (Sela et al. 2015). Simulation studies comparing the here described alignment maskers show that all of them improve the accuracy of subsequent phylogenetic analyses based on the masked alignment. Based on the analysed datasets, ZORRO and GUIDANCE outperform ALISCORE and GBLOCKS, resulting in more significant improvements of the alignment quality (Wu et al. 2012). This might be due to the fact that both ZORRO and GUIDANCE calculate scores for every single position, whereas «blocks» or «windows» of ambiguously aligned positions are identified by ALISCORE and GBLOCKS. Finally, GUIDANCE2 seem to outperform all here discussed methods (Sela et al. 2015).

6.5 Mapping Sequence Reads

A specific alignment application is the mapping of sequence reads to already known reference sequences (e.g. genomes, transcripts). Mapping is widely used to study gene expression, DNA-protein interaction, RNA splicing, SNP detection, or genome resequencing (Li et al. 2009b; Mortazavi et al. 2008; Nagalakshmi et al. 2008). Furthermore, mapping of sequence reads has been successfully used for the discovery and genotyping of transposable elements (Ewing 2015). The typical problem of read mapping is to resolve the exact origin (location) of a sequence read in a given reference sequence. This problem is complicated due to the occurrence of repetitive sequences (and thereby several equally likely locations), sequencing errors and genetic variation. Even more challenging is the mapping of mRNA transcripts onto reference genomes for the discovery of introns and splice variants, as huge gaps are expected separating the ends of the sequencing read. The BLAST algorithm described above could basically be used for read mapping, but as the output of next-generation sequencing technologies literally produce billions of short reads, more efficient and less time- and memory-consuming methods have to be explored. Nowadays several read mappers are available that are able to map millions of sequence reads onto large genomes within reasonable time using standard desktop computer resources.

Most mapping algorithms are either based on a seed-and-extend approach (hash table indexing) or are using methods based on the Burrows-Wheeler transform and specific indexing forms (Li and Homer 2010). The seed-and-extend approach is basically the same algorithm as used for BLAST searches. These approaches trace the position of each *k*-mer (or word) of a predefined length (e.g. 11 bps) of a query sequence and store them in a so-called hash table. By referring to the hash table, the reference sequence is scanned for exact matches of these *k*-mers, which are called seed, which are then attempted to be

elongated (▶ see 6.2). Retrieving all *k*-mers of a sequence and storing them in a table is called indexing. Several modified versions of this approach enhancing speed and sensitivity are implemented in read mapping software. For example, the software MAQ uses spaced speed indexing, where every read is divided into four segments which are used as seeds (Li et al. 2008a). By aligning all possible pairs of seeds against the reference sequence the list of possible locations where the full read maps can be limited quickly. The sensitivity of the mapping can be controlled by defining the number of possible mismatches of the seeds (spaced seeds). Other programs applying this strategy are indexing the reference sequence instead of the sequence reads, e.g. as implemented in SOAP (Li et al. 2008b) or BFAST (Homer et al. 2009). BFAST is first indexing the reference sequence, and in a second step, all candidate alignment locations are identified by using the stored *k*-mers. In a last step, a local alignment allowing gaps is performed. However, seed-and-extend approaches are intensive in the use of memory and computational time.

Much more memory-efficient and less time-consuming approaches of read mapping use an indexing scheme of the reference sequence based on Burrows-Wheeler transform (BWT) and FM-index (Ferragina and Manzini 2001), as, for example, implemented in the software BOWTIE (Langmead et al. 2009; Langmead and Salzberg 2012) and BWA (Li and Durbin 2009). BWT has been initially developed for data compression, e.g. to create zip files (Burrows and Wheeler 1994). Using BWT, a character string (in our case a sequence) is transformed by sorting all permutations of the string into lexical order and using the last column of this table as output (◘ Fig. 6.10). Due to the lexical ordering, transformed outputs will possess many repeated characters, which make them easily compressible. Without any extra information, it is possible to reverse the transformation of this output into the original string (sequence) (◘ Fig. 6.11).

The FM-index is a compressed suffix-array-like index based on BWT. It was created as a data structure that allows to locate and find a pattern within compressed text (Ferragina and Manzini 2000). To create the FM-index, the lexically sorted BWT of the sequence data is used. The transformed matrix can be used for so-called last first (LF) mapping. This means the i^{th} occurrence of a character in the last column corresponds to the i^{th} occurrence of the same character in the first column (Langmead et al. 2009). Using this lookup,

Original sequence	All permutations	Alphabetical ordering of rows	Output of last column
>BONOBO*	>BONOBO*	BONOBO*>	>
	*>BONOBO	BO*>BON**O**	**O**
	O*>BONOB	NOBO*>B**O**	**O**
	BO*>BONO	OBO*>BO**N**	**N**
	OBO*>BON	ONOBO*>**B**	**B**
	NOBO*>BO	O*>BONO**B**	**B**
	ONOBO*>B	>BONOBO*	*
	BONOBO*>	*>BONOB**O**	**O**

◘ **Fig. 6.10** Burrows-Wheeler transform of the string >BONOBO*. The original sequence is permutated in all possible orders, the rows are alphabetically ordered and last column is used as output. > denotes the beginning of the sequence, * denotes the end

Inverse transformation using Burrows-Wheeler transform			
Add cycle 1	Sort cycle 1	Add cycle 2	Sort cycle 2
>	B	>B	BO
O	B	OB	BO
O	N	ON	NO
N	O	NO	OB
B	O	BO	ON
B	O	BO	O*
*	>	*>	>B
O	*	O*	*>
Add cycle 3	Sort cycle 3	Add cycle 4	Sort cycle 4
>BO	BON	>BON	BONO
OBO	BO*	OBO*	BO*>
ONO	NOB	ONOB	NOBO
NOB	OBO	NOBO	OBO*
BON	ONO	BONO	ONOB
BO*	O*>	BO*>	O*>B
*>B	>BO	*>BO	>BON
O*>	*>B	O*>B	*>BO
Add cycle 3	Sort cycle 3	Add cycle 4	Sort cycle 4
>BONO	BONOB	>BONOB	BONOBO
OBO*>	BO*>B	OBO*>B	BO*>BO
ONOBO	NOBO*	ONOBO*	NOBO*>
NOBO*	OBO*>	NOBO*>	OBO*>B
BONOB	ONOBO	BONOBO	ONOBO*
BO*>B	O*>BO	BO*>BO	O*>BON
*>BON	>BONO	*>BONO	>BONO B
O*>BO	*>BON	O*>BON	*>BONO
Add cycle 3	Sort cycle 3	Add cycle 4	Sort cycle 4
>BONOBO	BONOBO*	>BONOBO*	BONOBO*>
OBO*>BO	BO*>BON	OBO*>BON	BO*>BONO
ONOBO*>	NOBO*>B	ONOBO*>B	NOBO*>BO
NOBO*>B	OBO*>BO	NOBO*>BO	OBO*>BON
BONOBO*	ONOBO*>	BONOBO*>	ONOBO*>B
BO*>BON	O*>BONO	BO*>BONO	O*>BONOB
*>BONOB	>BONOBO	*>BONOBO	>BONOBO*
O*>BONO	*>BONOB	O*>BONOB	*>BONOBO

Fig. 6.11 Inverse transformation of the output from Fig. 6.10 using Burrows-Wheeler transform. Starting with the output column, columns are added and lexically ordered. These steps are cyclically repeated till the original size of the string is recovered. The row with the symbol (*) denoting the sequence end at its end represents the original sequence (*shaded*)

exact matches of a read in the reference can be traced by subsequently tracing (aligning) the position of successively growing suffixes of the read starting from its end. For example, a read AGCT would be located along the rows of the BWT matrix in the order T, CT, GCT and AGCT (Trapnell and Salzberg 2009). Whereas exact matches are working well to find occurrences of words in a compressed book, it might be problematical to locate sequence reads, as they may not match exactly due to genetic variation or sequencing errors. Consequently, an algorithm implementing so-called backtracking is used to find inexact matches. This search is similar to that for exact matches and calculates matrix locations (ranges of possible rows) for successively longer suffixes of the query read. However, in case a suffix is not found in the text, an already matched suffix position is chosen and a substitution of a different base is introduced, thereby allowing a mismatch in the alignment (Langmead et al. 2009). Using a BWT approach for mapping has some key advantages. First, BWT approaches are memory efficient. The index for the complete human genome can be stored into less than 2 Gb of RAM memory (usually available for desktop computers), whereas by using a spaced seed approach, more than 50 Gb RAM is needed (access to a high-performance computer cluster necessary) (Trapnell and Salzberg 2009). Second, BWT approaches are also more time efficient. For example, the BWT-based BOWTIE runs around 30 times faster than MAQ, which uses a seed-and-extend approach.

Alternatively, alignment-free approaches are available for read mapping, e.g. as implemented in the software KALLISTO (Bray et al. 2016). In this case only the target sequence where a read is originating from is stored – but not the exact alignment position. In a first step, a de Bruijn graph of the reference sequences (e.g. transcriptome data) is created as index, where the nodes represent *k*-mers. Then, intersecting sets of *k*-mer matches of the reads are searched for in the graph to create pseudoalignments of the reads. By doing this, the information of the order of all *k*-mers of each single read remains intact. A similar approach described as «lightweight alignment» is used by the software SALMON (Patro et al. 2016). The advantage of these approaches is that they are order of magnitudes faster than alignment-based read mapping software, while being as accurate. Both methods are especially useful for RNA-Seq quantification, where only the information how many reads map on a specific transcript is important, but not its exact position.

Fonseca et al. (2012) counted more than 60 available read mappers in their review and even more have been published since then. These programs differ not only in the underlying algorithms, speed or memory efficiency but also in the ability to perform specific mapping problems. As such it is possible to map DNA on DNA, RNA on DNA or microRNAs back to the genome. For the detection of methylation patterns, it is possible to map reads from sequencing of bisulphite-treated DNA, where unmethylated C's are converted to T's (Chen et al. 2010). Another typical read mapping problem is the detection of splice junctions by mapping RNA-Seq reads onto a reference genome. A commonly used pipeline for this application is TOPHAT (Trapnell et al. 2009), which combines two of the above discussed methods. First, all reads are mapped onto the genome using BOWTIE. All reads that successfully map are used to generate consensus assemblies of possible exons, whereas reads which do not map onto the genome are collected for a second step. The exact limits of the identified exon regions are further refined based on the knowledge that most introns of eukaryotic genes begin with GT (splice donor) and end with AG (splice acceptor) (Mount 1982). Less frequent splice donor and acceptor pairs are also recognized, e.g. GC-AG and AT-AC introns. Identified exons represent possible splice sites. In a second step, the remaining reads are mapped onto these splice site candidates by a seed-and-extend approach using MAQ to find possible splices. Recently, with HISAT (Kim et al. 2015), a replacement of TOPHAT has been published. However, RNAs which are products

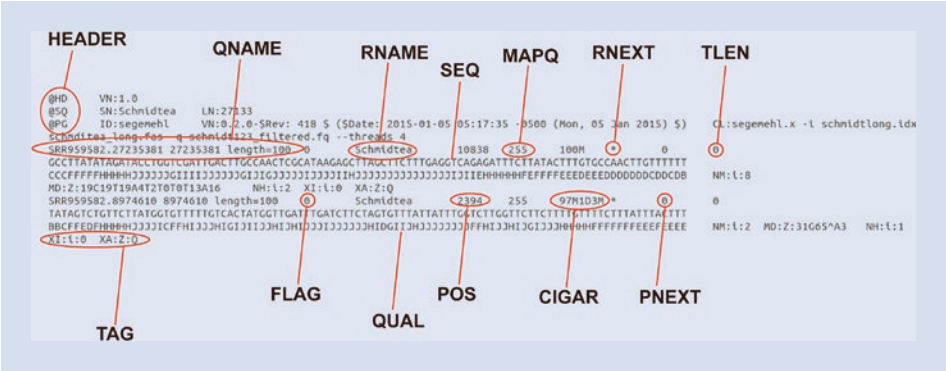


Fig. 6.12 Example of the SAM-format. The first four rows comprise the header section (HEADER) in this case including a header line (@HD), a reference sequence dictionary (@SD) and information about the used program (@PG). The next rows are the alignment section, always starting with the name of the read to map (QNAME), followed by the FLAG containing information about the sequence read. RNAME denotes the name of the reference sequence and the first position where the read starts to align in the reference (POS). The quality of the mapping can be indicated (MAPQ), and the CIGAR describes how the read maps. For example, 97M1D3M means that the first 97 bases are matching the reference, 1D describes a deletion in the reference, and the last 3 bases of the read match again. RNEXT gives information about the next mapping read (e.g. reverse read) and its starting position (PNEXT), TLEN describes the length of the template (e.g. read pair). This is followed by a string of characters which represent the actual mapped sequence (SEQ) and a string with its according quality values (QUAL). Additional information can be given as optional tags

of gene fusion, circularization or trans-splicing are difficult to detect with this approach. The read mapping software SEGEMEHL uses specific algorithms to find these more unusual RNAs (Hoffmann et al. 2014). Several read mappers for RNA-Seq data have been published, and their performance was evaluated by Engström et al. (2013).

The output of read mapping is usually stored in SAM- or BAM-format. The SAM-format consists of two sections: a header section and an alignment section. Every line of the header section starts with the character «@», whereas lines in the alignment section do not have this characteristic (Li et al. 2009a). The SAM-format can store plenty of information, e.g. quality scores, parameters of the used software, etc. (Fig. 6.12). The BAM-format is the binary equivalent of the SAM-format, making it more compressed and less memory intensive. BAM-files are not only used to store alignment information but also in the submission of raw-sequencing data to NCBI GenBank. Conversion of SAM-files to BAM-files and vice versa can be conducted by SAM-TOOLS (Li et al. 2009a). As SAM/BAM-files contain sequence and quality data, they can be easily converted into FASTQ- or FASTA-format, which are widely used formats for assembly or multiple sequence alignments.

6.6 Whole-Genome Alignments

For many comparative genomic analyses, it is necessary to align complete genomes. By comparing two genomes, differences can be found locally, but also at large scale (Darling et al. 2010; Feuk et al. 2006). For example, at local scales mutations will occur between two compared genome sequences, but also insertions and deletions. These are basically the same processes that have to be resolved as in alignments of sequences. However, at the genome level, also large-scale changes have to be taken into account, as genes or genomic regions can be either gained or lost. Some regions will be completely missing, whereas for

duplicated regions, homology has to be inferred. Moreover, the order of genes or genomic regions can be massively rearranged. Local mutations do not change the order of sequence positions and can be inferred by collinear alignment methods as introduced in this chapter. In contrast, large-scale changes can lead to noncollinear changes and need to be addressed by alignment approaches that focus on many different kinds of evolutionary changes of the genome. Whereas in collinear alignments, positional homology is inferred, the detection of noncollineary changes is basically the prediction of orthology of genes or larger genomic regions (Dewey 2012; Dewey and Pachter 2006).

Most whole-genome alignment methods can be broadly classified into hierarchical and local approaches (Dewey 2012). Using the hierarchical approach, collinear and homologous (ideally orthologous) segments are identified first. In a second step, global sequence alignments on a nucleotide level of these collinear segments are conducted. A widely used software implementing such an approach is progressiveMAUVE (Darling et al. 2010). Local approaches firstly conduct large sets of nucleotide alignments of genomic regions, which in subsequent steps are filtered and merged to produce alignments of homologous (ideally orthologous) genomic regions. MUMMER (Delcher et al. 1999) is among the most frequently used software solutions based on this approach.

Preservation of the order of genes or genomic regions along the chromosomes is called synteny (Bentley and Parkhill 2004). By conducting whole-genome alignments, syntenic regions across compared genomes can be identified and visualized. Synteny is used to identify conserved regions across compared genomes which are often interpreted as functional regions. Synteny of large regions of vertebrate genomes was already noticed in the pre-genomic era of molecular biology and interpreted as «frozen accidents» (Ohno 1973). It was first assumed that chromosomal rearrangements, which are able to break up larger syntenic regions, are randomly distributed within the genome of eukaryotes (Nadeau and Taylor 1984). Based on this idea, syntenic regions are basically relicts in the eukaryote genome (Kikuta et al. 2007). However, the increasing availability of completely sequenced genomes led to the discovery of syntenic blocks across deeply diverged lineages, which clearly suggest evolutionary conservation of these genomic regions. Earliest known examples are represented by clustering of Hox-genes among most investigated Metazoa (Ferrier and Holland 2001). However, as this gene family arose by tandem duplication, it might be an exception. Interestingly, large-scale genomic studies revealed that co-expressed genes are statistically more often clustered within the genome than expected, which has been demonstrated for all major eukaryotic lineages (Hurst et al. 2004). Furthermore, large genomic regulatory blocks including developmental regulatory genes and highly conserved non-coding sequences have been identified in vertebrate and insect genomes (Kikuta et al. 2007; Engström et al. 2007). Synteny is even more pronounced across bacterial and archaeal genomes, where co-expressed genes under the control of the same promoter are organized in operons.

References

Altschul S, Gish W, Miller W, Myers E, Lipman D (1990) Basic local alignment search tool. J Mol Biol 215:403–410

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402

Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW (2013) GenBank. Nucleic Acids Res 41:D36–D42

Bentley SD, Parkhill J (2004) Comparative genomic structure of prokaryotes. Annu Rev Genet 38:771–791
Bradley RK, Roberts A, Smoot M, Juvekar S, Do J, Dewey C, Holmes I, Pachter L (2009) Fast statistical alignment. PLoS Comput Biol 5:e1000392
Bray NL, Pimentel H, Melsted P, Pachter L (2016) Near-optimal probabilistic RNA-seq quantification. Nat Biotechnol 34:525–527
Burrows M, Wheeler DJ (1994) A block-sorting lossless data compression algorithm. Digital Equipment Corporation Technical Report 124, Palo Alto
Castresana J (2000) Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol 17:540–552
Chan SC, Wong AKC, Chiu DKY (1992) A survey of multiple sequence comparison methods. Bull Math Biol 54:563–598
Chen P-Y, Cokus SJ, Pellegrini M (2010) BS seeker: precise mapping for bisulfite sequencing. BMC Bioinformatics 11:203
Cole JR, Wang Q, Cardenas E, Fish J, Chai B, Farris RJ, Kulam-Syed-Mohideen AS, McGarrell DM, Marsh T, Garrity GM, Tiedje JM (2009) The ribosomal database project: improved alignments and new tools for rRNA analysis. Nucleic Acids Res 37:D141–D145
Cooper L, Cooper MW (1981) Introduction to dynamic programming. Pergamon Press, New York
Cristianini N, Hahn MW (2007) Introduction to computational genomics. Cambridge University Press, Cambridge, UK, A case studies approach
Darling AE, Mau B, Perna NT (2010) progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. PLoS One 5:e11147
Delcher AL, Kasif S, Fleischmann RD, Peterson J, White O, Salzberg SL (1999) Alignment of whole genomes. Nucleic Acids Res 27:2369–2376
Dewey CN (2012) Whole-genome alignment. In: Anisimova M (ed) Evolutionary genomics: statistical and computational methods, vol 1. Humana Press, Totowa, pp 237–257
Dewey CN, Pachter L (2006) Evolution at the nucleotide level: the problem of multiple whole-genome alignment. Hum Mol Genet 15:R51–R56
Eddy SR (2004) Where did the BLOSUM62 alignment score matrix come from? Nat Biotechnol 22:1035–1036
Edgar RC (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics 5:1–19
Edgar RC, Batzoglou S (2006) Multiple sequence alignment. Curr Opin Struct Biol 16:368–373
Engström PG, Ho Sui SJ, Drivenes Ø, Becker TS, Lenhard B (2007) Genomic regulatory blocks underlie extensive microsynteny conservation in insects. Genome Res 17:1898–1908
Engström PG, Steijger T, Sipos B, Grant GR, Kahles A, The RC, Ratsch G, Goldman N, Hubbard TJ, Harrow J, Guigo R, Bertone P (2013) Systematic evaluation of spliced alignment programs for RNA-seq data. Nat Methods 10:1185–1191
Ewing AD (2015) Transposable element detection from whole genome sequence data. Mob DNA 6:24
Feng D-F, Doolittle RF (1987) Progressive sequence alignment as a prerequisite to correct phylogenetic trees. J Mol Evol 25:351–360
Ferragina P, Manzini G (2000) Opportunistic data structures with applications. In: 41st annual symposium on Foundations of Computer Science, Washington, DC
Ferragina P, Manzini G (2001) An experimental study of an opportunistic index. Paper presented at the proceedings of the twelfth annual ACM-SIAM symposium on Discrete Algorithms, Washington, DC
Ferrier DEK, Holland PWH (2001) Ancient origin of the Hox gene cluster. Nat Rev Genet 2:33–38
Feuk L, Carson AR, Scherer SW (2006) Structural variation in the human genome. Nat Rev Genet 7:85–97
Fonseca NA, Rung J, Brazma A, Marioni JC (2012) Tools for mapping high-throughput sequencing data. Bioinformatics 28:3169–3177
Gardner PP, Wilm A, Washietl S (2005) A benchmark of multiple sequence alignment programs upon structural RNAs. Nucleic Acids Res 33:2433–2439
Giribet G, Wheeler WC (1999) On Gaps. Mol Phylogenet Evol 13:132–143
Henikoff S, Henikoff JG (1992) Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci U S A 89:10915–10919
Hillis DM, Huelsenbeck JP (1992) Signal, noise, and reliability in molecular phylogenetic analyses. J Hered 83:189–195
Hoffmann S, Otto C, Doose G, Tanzer A, Langenberger D, Christ S, Kunz M, Holdt L, Teupser D, Hackermuller J, Stadler P (2014) A multi-split mapping algorithm for circular RNA, splicing, trans-splicing and fusion detection. Genome Biol 15:R34

Homer N, Merriman B, Nelson SF (2009) BFAST: an alignment tool for large scale genome resequencing. PLoS One 4:e7767
Hurst LD, Pal C, Lercher MJ (2004) The evolutionary dynamics of eukaryotic gene order. Nat Rev Genet 5:299–310
Katoh K, Standley DM (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol 30:772–780
Kent WJ (2002) BLAT—the BLAST-like alignment tool. Genome Res 12:656–664
Kikuta H, Laplante M, Navratilova P, Komisarczuk AZ, Engström PG, Fredman D, Akalin A, Caccamo M, Sealy I, Howe K, Ghislain J, Pezeron G, Mourrain P, Ellingsen S, Oates AC, Thisse C, Thisse B, Foucher I, Adolf B, Geling A, Lenhard B, Becker TS (2007) Genomic regulatory blocks encompass multiple neighboring genes and maintain conserved synteny in vertebrates. Genome Res 17:545–555
Kim D, Langmead B, Salzberg SL (2015) HISAT: a fast spliced aligner with low memory requirements. Nat Methods 12:357–360
Kück P, Meusemann K, Dambach J, Thormann B, von Reumont BM, Wägele JW, Misof B (2010) Parametric and non-parametric masking of randomness in sequence alignments can be improved and leads to better resolved trees. Front Zool 7:10
Landan G, Graur D (2007) Heads or tails: a simple reliability check for multiple sequence alignments. Mol Biol Evol 24:1380–1383
Langmead B, Salzberg SL (2012) Fast gapped-read alignment with bowtie 2. Nat Methods 9:357–359
Langmead B, Trapnell C, Pop M, Salzberg S (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10:R25
Lassmann T, Sonnhammer ELL (2005) Automatic assessment of alignment quality. Nucleic Acids Res 33:7120–7128
Levinson G, Gutman GA (1987) Slipped-strand mispairing: a major mechanism for DNA sequence evolution. Mol Biol Evol 4:203–221
Li H, Durbin R (2009) Fast and accurate short read alignment with burrows-wheeler transform. Bioinformatics 25:1754–1760
Li H, Homer N (2010) A survey of sequence alignment algorithms for next-generation sequencing. Brief Bioinform 11:473–483
Li H, Ruan J, Durbin R (2008a) Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res 18:1851–1858
Li R, Li Y, Kristiansen K, Wang J (2008b) SOAP: short oligonucleotide alignment program. Bioinformatics 24:713–714
Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Subgroup GPDP (2009a) The Sequence Alignment/Map format and SAMtools. Bioinformatics 25:2078–2079
Li R, Li Y, Fang X, Yang H, Wang J, Kristiansen K, Wang J (2009b) SNP detection for massively parallel whole-genome resequencing. Genome Res 19:1124–1132
Lipman D, Pearson W (1985) Rapid and sensitive protein similarity searches. Science 227:1435–1441
Löytynoja A, Goldman N (2008) Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. Science 320:1632–1635
Löytynoja A, Milinkovitch MC (2001) SOAP, cleaning multiple alignments from unstable blocks. Bioinformatics 17:573–574
McGuffin L (2009) Insertion and deletion events, their molecular mechanisms, and their impact on sequence alignments. In: Rosenberg M (ed) Sequence alignment: methods, models, concepts and strategies. Universtiy of California Press, Berkeley, pp 23–38
Misof B, Misof K (2009) A Monte Carlo approach successfully identifies randomness in multiple sequence alignments: a more objective means of data exclusion. Syst Biol 58:21–34
Morgenstern B (2009) Local versus global alignments. In: Rosenberg M (ed) Sequence alignment: methods, models, concepts and strategies. Universtiy of California Press, Berkeley, pp 39–53
Morrison DA, Ellis JT (1997) Effects of nucleotide sequence alignment on phylogeny estimation: a case study of 18S rDNAs of apicomplexa. Mol Biol Evol 14:428–441
Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B (2008) Mapping and quantifying mammalian transcriptomes by RNA-seq. Nat Methods 5:621–628
Mount SM (1982) A catalogue of splice junction sequences. Nucleic Acids Res 10:459–472
Nadeau JH, Taylor BA (1984) Lengths of chromosomal segments conserved since divergence of man and mouse. Proc Natl Acad Sci U S A 81:814–818

Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, Snyder M (2008) The transcriptional landscape of the yeast genome defined by RNA sequencing. Science 320:1344–1349
Needleman SB, Wunsch CD (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol 48:443–453
Notredame C, Higgins DG, Heringa J (2000) T-coffee: a novel method for fast and accurate multiple sequence alignment1. J Mol Biol 302:205–217
Ohno S (1973) Ancient linkage groups and frozen accidents. Nature 244:259–262
Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C (2016) Salmon provides accurate, fast, and bias-aware transcript expression estimates using dual-phase inference. bioRxiv. doi.org/10.1101/021592.
Pei J, Grishin NV (2001) AL2CO: calculation of positional conservation in a protein sequence alignment. Bioinformatics 17:700–712
Penn O, Privman E, Ashkenazy H, Landan G, Graur D, Pupko T (2010a) GUIDANCE: a web server for assessing alignment confidence scores. Nucleic Acids Res 38:W23–W28
Penn O, Privman E, Landan G, Graur D, Pupko T (2010b) An alignment confidence score capturing robustness to guide tree uncertainty. Mol Biol Evol 27:1759–1767
Pevsner J (2015) Bioinformatics and functional genomics, 3rd edn. Wiley-Blackwell, Hoboken
Phillips A, Janies D, Wheeler W (2000) Multiple sequence alignment in phylogenetic analysis. Mol Phylogenet Evol 16:317–330
Privman E, Penn O, Pupko T (2012) Improving the performance of positive selection inference by filtering unreliable alignment regions. Mol Biol Evol 29:1–5
Rice P, Longden I, Bleasby A (2000) EMBOSS: The European Molecular Biology Open Software Suite. Trends Genet 16:276–277
Rosenberg M (2009) Sequence alignment: concepts and history. In: Rosenberg M (ed) Sequence alignment: methods, models, concepts and strategies. Universtiy of California Press, Berkeley, pp 1–22
Sela I, Ashkenazy H, Katoh K, Pupko T (2015) GUIDANCE2: accurate detection of unreliable alignment regions accounting for the uncertainty of multiple parameters. Nucleic Acids Res 43:W7–W14
Simmons MP, Ochoterena H (2000) Gaps as characters in sequence-based phylogenetic analyses. Syst Biol 49:369–381
Simmons MP, Müller KF, Norton AP (2010) Alignment of, and phylogenetic inference from, random sequences: the susceptibility of alternative alignment methods to creating artifactual resolution and support. Mol Phylogenet Evol 57:1004–1016
Smith TF, Waterman MS (1981) Identification of common molecular subsequences. J Mol Biol 147:195–197
Talavera G, Castresana J (2007) Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. Syst Biol 56:564–577
Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22:4673–4680
Thompson JD, Koehl P, Ripp R, Poch O (2005) BAliBASE 3.0: latest developments of the multiple sequence alignment benchmark. Proteins: Structure, Function, and Bioinformatics 61:127–136
Thompson JD, Linard B, Lecompte O, Poch O (2011) A comprehensive benchmark study of multiple sequence alignment methods: current challenges and future perspectives. PLoS One 6:e18093
Thorne JL, Kishino H (1992) Freeing phylogenies from artifacts of alignment. Mol Biol Evol 9:1148–1162
Thornton JW, DeSalle R (2000) Gene family evolution and homology: genomics meets phylogenetics. Annu Rev Genomics Hum Genet 1:41–73
Trapnell C, Salzberg SL (2009) How to map billions of short reads onto genomes. Nat Biotechnol 27:455–457
Trapnell C, Pachter L, Salzberg SL (2009) TopHat: discovering splice junctions with RNA-seq. Bioinformatics 25:1105–1111
Wong KMA, Suchard MA, Huelsenbeck JP (2008) Alignment uncertainty and genomic analysis. Science 319(5862):473–476
Wu M, Chatterji S, Eisen JA (2012) Accounting for alignment uncertainty in phylogenomics. PLoS One 7:e30288

Finding Genes

C. Bleidorn, *Phylogenomics*, DOI 10.1007/978-3-319-54064-1_7

- Genes are broadly defined as a union of genomic sequences encoding a coherent set of potentially overlapping functional products.
- Gene numbers between species are highly variable and gene loss and gain are common.
- Homologous genes are derived from a common ancestor; copies which arose due to speciation events are called orthologs, and those that arose by duplication events are paralogs.
- Orthology inference methods are classified into graph-based methods and tree-based methods.
- The ortholog conjecture predicts that orthologs are more likely to indicate conserved function than paralogs, an assumption used for genome annotations.
- Functional predictions of gene products are cataloged according to cellular component, biological processes and molecular function in the Gene Ontology project using a controlled vocabulary.

7.1 What Is a Gene?

The concept of a «gene» is central to molecular biology in general. Surprisingly, there seems to be no consensus of what a gene is – and widely used textbooks or databases define genes differently (Orgogozo et al. 2016). The term gene as initially introduced by Johannsen (1909) abstractly described the units of inheritance as described in the work of Gregor Mendel. Later on, the definition changed over the course of time and was updated to reflect new scientific insights, as, for example, the discovery that genes play an important role for biochemical pathways or the discovery of the genetic code. First, it was postulated that genes carry the instructions for proteins, leading to the paradigm that one gene encodes for one protein. Subsequently, a more general definition of genes defined them as an annotated open reading frame in the genome (Doolittle 1986). However, the picture became more complicated after it has been discovered that the same DNA sequence can be transcribed into different proteins (e.g., alternative splicing) and that overlap between genes is not unusual. To deal with all these complexities, Gerstein et al. (2007) came up with the following definition:

» The gene is a union of genomic sequences encoding a coherent set of potentially overlapping functional products.

In this definition, functional products refer to either transcribed RNA molecules or proteins.

7.2 Gene Gain and Loss

The number of genes between species is highly variable. There seems to be no correlation between the complexity of an organism (e.g., as measured by the number of different cell types) and its number of genes. For example, a comparative analysis across 12 closely related *Drosophila* species, which share a last common ancestor around 60 mya, found that the number of genes among these relatively closely related species varies between 14,000 and 18,000 (Hahn et al. 2007; Drosophila 12 Genomes Consortium 2007). This

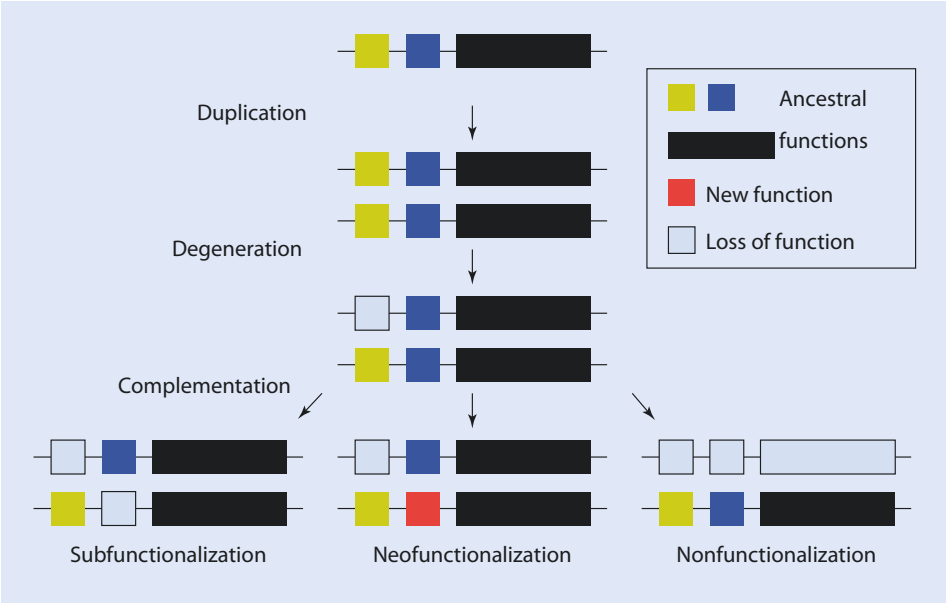


Fig. 7.1 The DDC model describing alternative evolutionary fates of duplicated genes after Force et al. (1999). An ancestral gene is duplicated and both copies are retained. After some time, degenerative mutations can lead to loss of function for part of the gene (e.g., regulatory sequence or protein domain). A second mutation defines the fate of the gene duplicate. Under subfunctionalization, ancestral functions are retained in different copies of the gene. The acquisition of new functions is described as neofunctionalization. Nonfunctionalization describes the fate when one copy loses all functional ability

illustrates that gene gain and loss is frequent over evolutionary timescales. Duplication events also explain why most genes are organized in gene families in which only the smaller part is represented by single-copy genes. As typical for organisms in general, most genes are protein-coding genes, whereas only a small fraction is represented by different RNA genes (Drosophila 12 Genomes Consortium 2007). Whereas for many genes homologues can be found in many species, some genes are restricted to single lineages (orphan genes).

Copies of genes arise either by whole-genome duplications or by gene duplication events. Many different models of gene duplication evolution have been postulated (Innan and Kondrashov 2010). A popular model describing the evolutionary fate of genes during evolution is the duplication-degeneration-complementation (DDC) model (Fig. 7.1) (Force et al. 1999). According to this model, three different fates of a copy of a gene are distinguished: neofunctionalization, subfunctionalization and nonfunctionalization. The most frequent outcome of duplication events is nonfunctionalization, where one copy ends up as a pseudogene due to the accumulation of mutations (e.g., causing frame shifts, internal stop codons). Over time pseudogenes are removed from the genome or are so divergent that similarity with functional copies is undetectable (Zhang 2003). Two different ways to retain functional copies of genes are distinguished. When one copy acquires a new function, it is termed neofunctionalization. Subfunctionalization describes the persistence of partial ancestral functions in different copies of the gene. Gene copies that are retained via subfunctionalization might acquire new functions over time and are thereby neofunctionalized (Rastogi and Liberles 2005).

7.3 Homology of Genes

Homology is a concept central to comparative biology in general. The term homology was first used by Richard Owen to describe anatomical similarities (*the same organ under every variety of form and function*). Owen (1843) used topological correspondence to support different types of homology (general, special, serial) of anatomical similarities. This definition predates the publication of Darwin's theory of evolution. Lankester (1870) coined the first definition of homology with an evolutionary background where he distinguishes between shared similarities based on common ancestry (homogeny) and similarities which can be traced back to a common function (homoplasy). In modern evolutionary biology, homology broadly defines the relationship of two characters that have descended from a common ancestral character (Fitch 2000). These characters can refer to nucleotide or amino acid positions, genes, morphology or behavioral features of an organism. It is important to remember that homologies represent hypotheses. Homologies can be either defined using similarities or referring to inheritance from a common ancestor. In a cladistic framework, primary and secondary homology has been distinguished (de Pinna 1991). Hypotheses of primary homology are based on similarities, which can be inferred, for example, using the criteria of homology proposed by Remane (1952): position, structure and transitions. In a second step, a phylogeny is reconstructed using the defined homologies to investigate if the hypotheses of primary homology are congruent with the resulting phylogenetic tree. This way of analyzing homologies was established for morphological characters, but seems less usable for molecular characters (but see Brower and Schawaroch (1996)), which are rather simple and therefore prone to convergence. Moreover, this approach cannot be used to resolve homology between genes.

Consequently, in molecular systematics and genomics, refined criteria for homology are in use. The positional homology of nucleotide or amino acid positions within a sequence is used to describe the relationship of sites within gene alignments. Analysing homology of genes is more complex. To distinguish between different types of homology of genes, Fitch (1970) introduced the concepts of orthology and paralogy (Fig. 7.2). Homologous genes are derived from a common ancestor and are usually identified through similarity. Genes are either homologous or not, and statements describing the percentage of homology should be avoided or in this case rather referred to identity or similarity (Webber and Ponting 2004). Orthologous genes (or orthologs) are defined as a pair of homologous genes that have emerged through a speciation event. In contrast, paralogous

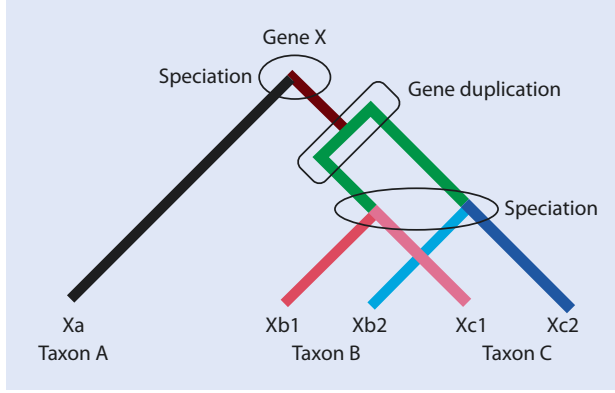


Fig. 7.2 Following the history of gene x through time. Two speciation events and one duplication event are illustrated. Last common node of genes Xb1 and Xc1 (or Xb2 and Xc2) refers to a speciation event, which makes them orthologs. Last common node of genes Xb1 and Xb2 (or, e.g., Xb1 and Xc2) refers to a duplication event, which makes the paralogs. Gene Xa is an ortholog to all other genes, as their last common node refers to a speciation event

genes (or paralogs) can be traced back to duplication events within a lineage. Relative to the timing of the duplication event, two types of paralogs are distinguished. Inparalogs are paralogs that arose by duplication after the speciation event separating the lineages which are compared. Outparalogs are those paralogs where the duplication event happened before the speciation event (Sonnhammer and Koonin 2002). This definition always relates to the level of comparison. Comparing taxa A and B in ◘ Fig. 7.2, genes Xb1 and Xb2 are inparalogs, as the duplication event of these genes happened after the lineages of taxon A and B split. In this case both genes are regarded as co-orthologs to gene Xa. By comparing taxa B and C, genes Xa1 and Xa2 are regarded as outparalogs, as the duplication even took place after the speciation event of the compared lineages (◘ Fig. 7.2). A special case of homology is defined as xenology. Xenologs are a pair of homologous genes where its common evolutionary history involves horizontal gene transfer of at least one of these genes (Fitch 2000). The concepts of orthology, paralogy and xenology are not restricted to genes and can also be applied to larger genomic regions (Webber and Ponting 2004).

Distinguishing orthologous from paralogous genes is central to phylogenomics. Phylogenetic relationships of species can be inferred by analyzing alignments of orthologous genes, as the gene trees should in the ideal case coincide with the species tree. However, if several copies of inparalogs exist in one of the investigated taxa, it will not matter for phylogenetic analyses which copy is chosen. Moreover, orthology is used to infer function of genes (Eisen 1998). According to the «ortholog conjecture» (► see Sect. 7.7), it is assumed that orthologous genes are more likely to be functionally similar than paralogous genes (Koonin 2005).

7.4 Inferring Orthology

Orthology inference methods can be broadly classified into graph-based methods and tree-based methods (Kuzniar et al. 2008; Altenhoff and Dessimoz 2012). In graph-based methods, genes represent the nodes of the graph and edges connect these nodes according to their sequence similarity. Clusters of genes representing putative orthologs can be resolved using these graphs. Graph-based methods are typically used to infer orthology of genes in two genomes. Based on the assumption that orthologs in the two compared genomes share the highest similarity, local alignments (e.g., BLAST) are used to infer the best hit for every gene in the other genome. As this relation should be symmetrical for orthologs, local alignments of the best hit with the genome containing the originally queried gene should exactly find this gene as reciprocal best hit (◘ Fig. 7.3a). Several refinements of the approach to find reciprocal best hits have been developed. It has been noted that the accuracy of ortholog prediction depends on gene length when using BLAST scores. This is due to the fact that pairwise comparisons of short sequences cannot produce high-scoring similarity values or low e-values, whereas long sequences may always generate high scores. Subsequently, short genes might be missed in ortholog predictions, whereas long genes are prone to produce false positives. This bias is taken into account using a normalization step for all scores by the software ORTHOFINDER (Emms and Kelly 2015). The program INPARANOID not only searches for orthologs in a pair of genomes (or proteomes) but also distinguishes between inparalogs and outparalogs (Remm et al. 2001). In a first step, an all-against-all BLAST search is conducted to find pairs of sequences with reciprocal best hits. These pairs of sequences are regarded as main orthologs. Additionally, potential co-orthologs are detected for each ortholog group based on BLAST similarity

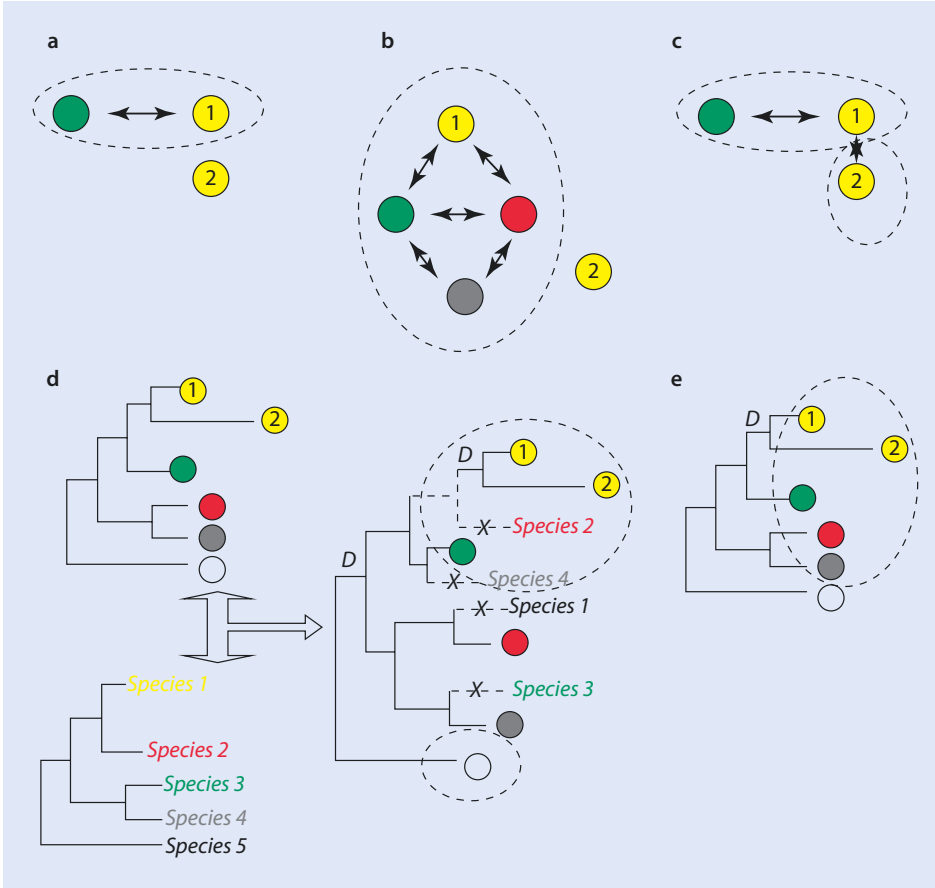


Fig. 7.3 Graph-based and tree-based prediction of orthology. Circles in different colors represent proteins from different species. *Dashed circles* surround orthologous groups. **a** The *yellow* protein 1 is the reciprocal best hit of the *green* protein. **b** Triangulation is used in the establishment of COGs. *Yellow* 1, *green*, *red* and *grey* are part of two overlapping triangles of reciprocal best hits, which are merged into an ortholog cluster. **c** The INPARANOID approach also considers potential co-orthologs, besides the identified reciprocal best hits. Co-orthologs are more similar to each other than to any protein of other compared species. **d** Gene tree (*top*) and species tree are reconciled for the prediction of orthology. **e** All proteins derived from a common ancestor are regarded as orthologs in this phylogenetic approach. A recent duplication event led to two paralogs in the yellow species (Figure reprinted from Gabaldón (2008))

scores, under the assumption that sequences from the same species that are more similar to the main ortholog than to any sequence from the other species are inparalogs (Fig. 7.3c). A database containing orthologs and inparalogs of pairwise comparisons of 273 (mostly eukaryotic) organisms is available (Sonnhammer and Östlund 2015). A similar approach to INPARANOID uses Markov clustering (MCL) to detect groups of orthologs/co-orthologs. First, orthologs and potential inparalogs are detected as described for INPARANOID. The recovered information is stored in a graph, with the sequences as nodes and similarity values as connecting edges. Using the MCL algorithm, this graph is resolved into clusters of orthologs and inparalogs (Li et al. 2003). The MCL algorithm uses

random walks along the graph to identify clusters (van Dongen 2000). A random walk is a path of random steps on a graph created by a Markov process. In a Markov process, stochastic changes of states (e.g., steps along the graph) are only depending on the current state (e.g., current position in the graph) and given transition rules. By using random walks, a flow is simulated within the graph, and clusters are delineated as groups of nodes between which the current of the flow changes from strong to weak. A database with MCL ortholog clusters for 150 sequenced genomes (July 2016) is available (Chen et al. 2006). Another widely used collection of orthologs can be found in the COG (for prokaryotes) and KOG (for eukaryote) databases. COGs (or KOGs) are clusters of orthologs which are constructed based on BLAST similarities. First, sets of proteomes (all annotated proteins in a genome) from different species are compared using all-against-all BLAST. Based on the similarity scores, obvious paralogs are detected as sequences from a single species which are more similar to each other than to sequences of other proteomes and subsequently merged. In a second step, triangles (three-way comparisons) of consistent reciprocal best hits across genomes (including detected paralog groups) are collected and merged into larger COGs (or KOGs) using multiple triangle comparisons (◘ Fig. 7.3b) (Tatusov et al. 2000, 2003). A similar approach, but based on all-against-all Smith and Waterman local alignment distances, is used by eggNOG (Jensen et al. 2008). The eggNOG database includes ortholog groups and functional annotation of more than 2000 organisms and also >350 viral proteomes (Huerta-Cepas et al. 2016). A comparison of the here-mentioned databases hosting ortholog collections was conducted by Altenhoff and Dessimoz (2009), which further includes the also widely used OMA database (Altenhoff et al. 2015).

Tree-based methods use phylogenetic trees to distinguish orthologs from paralogs. In contrast to graph-based methods, multiple sequence alignments of a set of homologues sequences are a prerequisite for this approach (Kuzniar et al. 2008). Based on the alignment, a phylogenetic reconstruction is performed to recover the gene tree. Using a concept that is called reconciliation, nodes in the tree referring to speciation or duplication events are identified based on species relationships (◘ Fig. 7.3d). Reconciliation methods mostly focus on gene duplication and gene loss events to explain differences between gene trees and species trees (Arvestad et al. 2003). However, other biological mechanisms such as horizontal gene transfer, hybridization or incomplete lineage sorting may also contribute to these differences, thereby leading to wrong orthology assignments. Some methods are available to distinguish between orthology and paralogy in the absence of a species tree (◘ Fig. 7.3e) (van der Heijden et al. 2007). However, tree-based approaches have been shown to be less accurate for orthology prediction, as they are sensitive to multiple sequence alignment errors and the availability of (correctly rooted) species trees. Additionally, tree-based inference is strongly influenced by massive gene loss events, which seem to be difficult to reconcile (Gabaldón 2008; Kuzniar et al. 2008). Several hybrid methods using a mixture of both approaches have been proposed, which usually start with graph-based orthology prediction which is subsequently refined using phylogenetic approaches (Kuzniar et al. 2008; Kristensen et al. 2011).

7.5 Hidden Markov Profiles

A typical task in phylogenomics is to find a set of predefined orthologs (e.g., recovered by the methods introduced above) in a transcriptomic dataset. Besides searches for the best BLAST hit, approaches based on profile hidden Markov models (pHMM) became widely

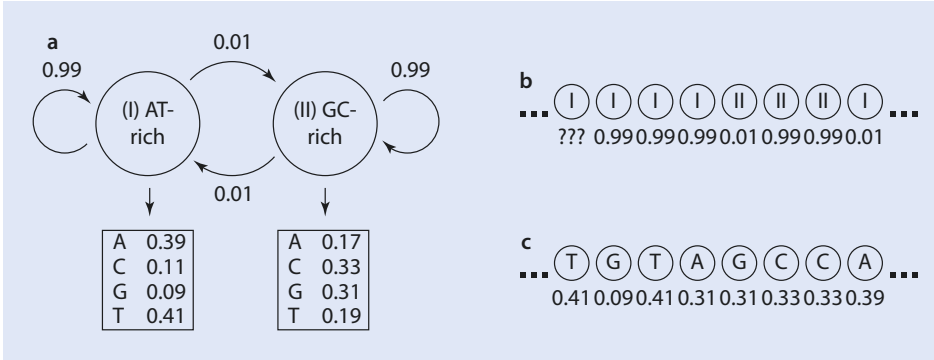


Fig. 7.4 **a** Two-state hidden Markov model describing states to distinguishing AT- and GC-rich base composition. State transition probabilities are indicated by curved arrows. Symbol emission frequencies for A, C, G and T are given for each state. **b** Sequence of hidden states (I, AT rich; II GC rich) and their corresponding transition frequencies. **c** Sequence of observable emitted symbols and their corresponding emission frequency

7

used to address this question. By using BLAST searches, it is assumed that all alignment positions are equally important and scoring parameters independent of the alignment position are used. However, multiple alignments of gene families usually show that some positions are more conserved than others, as well as in some regions indels might be more frequent. By defining a hidden Markov profile, position-specific information can be harnessed for a refined search of potential orthologs. Especially for the detection of homology between sequences which diverged long ago and where structural and functional diversity is obscure, profile HMM methods have been demonstrated to clearly outperform pairwise methods like BLAST (Park et al. 1998). Moreover, profile HMMs can be also used to generate multiple sequence alignments or to add sequences to existing multiple sequence alignments (Sievers et al. 2011).

HMMs are probabilistic models which model a system under the assumption that it can be represented by a Markov process (see above, ▶ Sect. 7.3) with hidden (unobserved) states. The theory behind HMMs was already developed back in the 1960s and has been especially applied for speech recognition problems. For example, HMMs have been used recognizing words within recorded and digitized sequences of human speech (Rabiner 1989). This example can be used to also understand how HMMs are used in a phylogenomic context. Instead of words in a digitized sequence of speech, sequence motives are searched for in a DNA or protein sequence. In both cases, noise is a problem to recognize the words or sequence motive. In the case of the latter, the noise is generated due to mutations over time. The use of HMMs for genomics was introduced by Krogh et al. (1994). HMMs are used to model a sequence of states (Fig. 7.4a) as they would be generated by a Markov chain (Cristianini and Hahn 2007). Each single position of this sequence is represented by a hidden state, which cannot be directly observed (Fig. 7.4b). However, each state emits symbols according to a multinomial distribution (Fig. 7.4c). This means two random processes are part of the model, one generating the Markov chain of hidden states and another process resulting in the emission of symbols into a sequence which can be observed. In the simplest case, HMMs consist of two different states. For example, for DNA sequences, these states can represent AT-rich or GC-rich segments of the sequence (Fig. 7.4a). The transition probability describes the frequency of switches between two

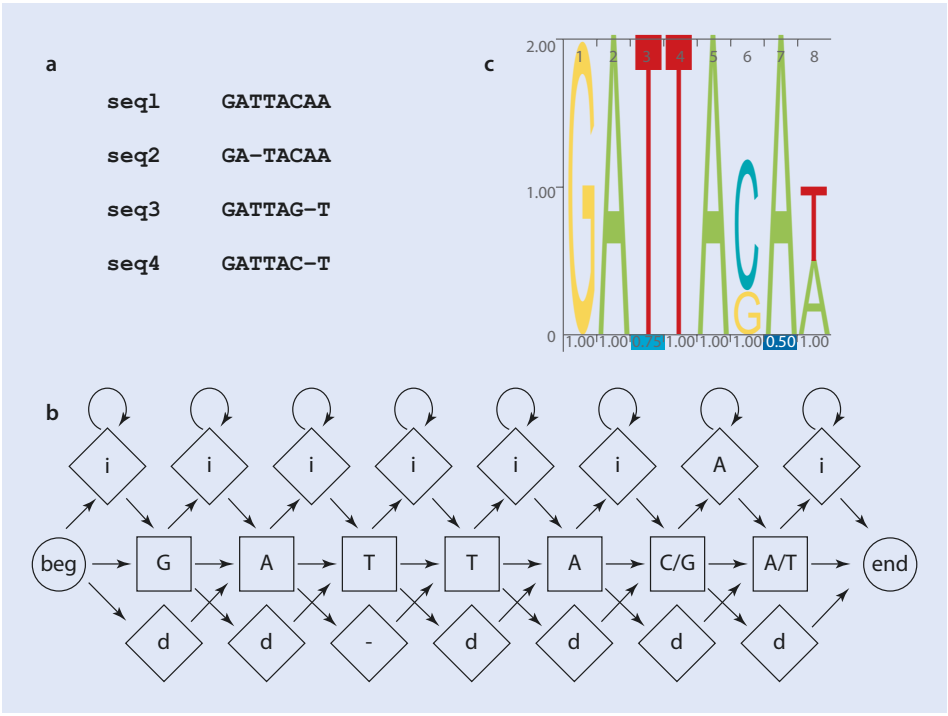


Fig. 7.5 Profile HMMs. **a** Multiple alignment of four sequences. **b** Profile hidden Markov model (pHMM) of the alignment in A. Match states can be found in the *middle row*, insertion states (i) in the *upper row* and deletion states (d) in the *lower row*. Transition probabilities of states are indicated by *arrows*. Only dominant symbols are indicated, even though each symbol is possible for all insertion and match states. Additionally, states for the beginning (beg) and ending (end) are included. Sequences are generated by following the arrows along the states; for insertion states loops are possible, allowing the emission of more than one symbol at this position. **c** Visualization of the information of the pHMM as a logo created by using skylign (Wheeler et al. 2014)

different states, while the emission probability describes the probability that symbols within an observable sequence are produced. For DNA sequences the number of possible symbols is typically four (A, C, G, T), whereas for proteins 20 different symbols representing the different amino acids are standard.

In a phylogenomic framework, HMMs can be used to address three different questions (Eddy 1996): (I) finding similar sequences based on a HMM profile, (II) alignment of sequence-profiles and (III) generation of HMM profiles from multiple alignments. Finding genes with HMMs consists of three steps. Firstly, a multiple sequence alignment for a gene family or cluster of orthologs of interest is generated. In a second step, a profile based on this multiple sequence alignment is generated (Fig. 7.5a). This profile should not only summarize the frequency of a given symbol (nucleotide or amino acid) at any alignment position but also include the probabilities of insertions and deletions (Fig. 7.5b). Match states describe the frequency of nucleotides or amino acids for every sequences position, while insertion or deletion states model the frequencies of indels within the alignment. Each match and each deletion state emit residues with different probabilities. Deletion states only introduce gaps within the alignment. The HMM-based

profile for any given alignment can be visualized as a logo (Fig. 7.5c), in which the probability for the occurrence of a residue at any given alignment position is indicated by the size of its letter (Wheeler et al. 2014). The probabilities for indels are indicated below every position. These logos nicely demonstrate how profile HMMs are used as a blueprint to find sequences matching this pattern. In the last step, the similarity of any given sequence with the model can be assessed and expressed in a score. The higher the score, the better a sequence fits to the HMM. Similar to BLAST analyses, scores can be used to generate e-values for every queried sequence.

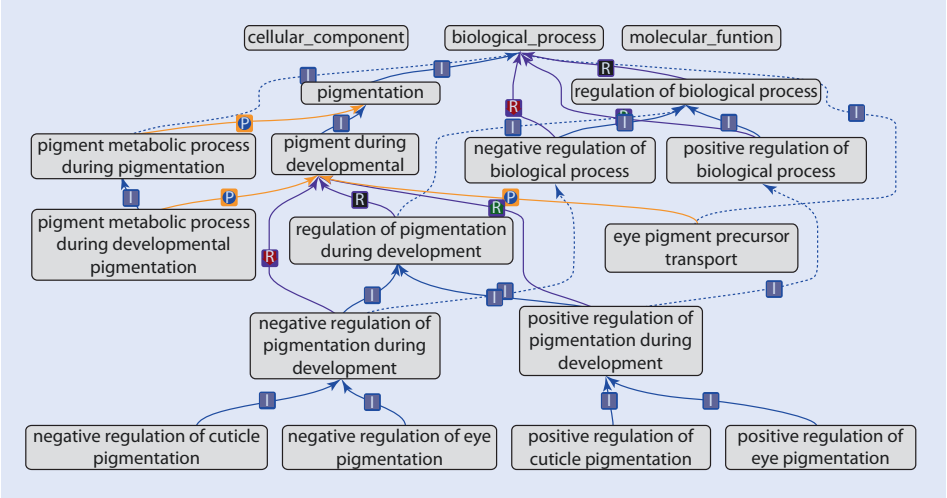
The most widely used software to work with profile HMMs is Sean Eddy's HMMER package, which never has been formally published in a scientific journal, but is freely available under ▶ http://hmmer.wustl.edu. Additionally, the HMMER web server can be accessed (Finn et al. 2011). HMMER can be used to search protein databases by query sequences, as well as to add sequences to an existing multiple alignment. HMMER is also a central to the PFAM database, which hosts a large collection of protein families represented by multiple sequence alignments and their respective hidden Markov profiles (Finn et al. 2016). By querying PFAM, affiliation of a protein sequence to a protein family can be retrieved, as well as the identification of functional domains. As of July 2016, the PFAM database contains more than 16,000 protein families, which are clustered into ~650 clans. A clan includes evolutionary related protein families, as evidenced by domain structure, function or similarities of HMM profiles (Finn et al. 2006).

7.6 Gene Ontology and the Ortholog Conjecture

The term «phylogenomics» was originally coined by Jonathan Eisen to describe the use of phylogenetic approaches to characterize the function of genes based on their evolutionary ancestry (Eisen 1998). Normally, determining the function of a gene or, to be more precise, its gene product (e.g., a protein or RNA) is laborious and difficult to do. Not surprisingly, most gene functions are only well characterized for a handful of model organisms (e.g., *Drosophila melanogaster*, *Caenorhabditis elegans*, *Arabidopsis thaliana*). Whereas sequence data can be nowadays (often easily) generated for every organism we wish to investigate, most of them might be inaccessible for experimental approaches or only with great difficulty. For example, only a small fraction of known bacteria have ever been cultured in the lab (Stewart 2012). Instead of using experimental data, resolving the orthology across genes of interest can be used to derive hypotheses of possible function. At the heart of this idea is the so-called ortholog conjecture. Based on this conjecture, it is hypothesized that orthologs are on average functionally more similar than paralogs (Thomas et al. 2012; Gabaldon and Koonin 2013). Under this premise gene function for experimentally inaccessible organisms is hypothesized using known functions from orthologs of well-investigated organisms. For example, if the function of gene x is characterized for *Drosophila melanogaster*, the same function is assumed for an orthologous gene in other animals, or even non-animals. The idea of the ortholog conjecture goes back to the models of evolution of genes after duplication events (see above), where duplicated genes are retained after neofunctionalization (gain of a new function) or subfunctionalization (partial loss of function). Both processes result into the evolution of paralogs with somehow different functions (Ohno 1970). Whereas the ortholog conjecture has been implicitly used for genome annotation over a decade, it has been rarely explicitly tested (Gabaldon and Koonin 2013). Nehrt et al. (2011) aimed to test the ortholog conjecture

using available functional genomic data from mammals. By using comparative statistical analyses of gene ontology (see below) annotations of different pairs of genes from mouse and humans, they came to the conclusion that paralogs are often functionally more similar than orthologs. This result would basically reject the ortholog conjecture, thereby also casting doubt on the current way of genome annotation. However, the way gene ontologies are used in this analysis has been strongly criticized (Thomas et al. 2012). Instead, subsequent studies based on the analyses of RNA-Seq data from multiple tissues of several species found support for the validity of the ortholog conjecture (Chen and Zhang 2012). Nevertheless, whereas the general trend that orthology predicts function seems reasonable, it must not necessarily be so. Vice versa, shared function must not be a predictor of orthology (Gabaldon and Koonin 2013).

The Gene Ontology (GO) project goes back to a joint initiative of model organism databases (Ashburner et al. 2000). Initially, functional annotation of genes in mouse (*Mus musculus*), fly (*Drosophila melanogaster*) and yeast (*Saccharomyces* spp.) has been done independently. However, it became obvious that this practice will lead to inconsistent ways of annotation, thereby hampering comparative analyses. Consequently, a common database using standardized rules and identical vocabulary has been developed. The GO project comprises three different ontologies to describe functions of gene products related to their cellular components, biological processes and molecular functions. Cellular components are components of a cell, which should be part of a larger object (e.g., mitochondrion, ribosome, etc.). Molecular functions describe activities occurring at a molecular level (e.g., nucleotide binding, virus receptor activity, etc.). Biological processes are a series of events of molecular function (e.g., development, pigmentation). Each of these ontologies represents a structured and controlled vocabulary, which includes terms which show a relationship that can be visualized in a graph (■ Fig. 7.6). Each term belonging to one of three ontologies has a definition, a unique name and an identifier. For example, the term pigmentation is defined as «The accumulation of pigment in an organism, tissue or cell,



■ **Fig. 7.6** Example of the graph structure of Gene Ontology for a set of terms under the biological process node «pigmentation.» Relationships between nodes can, for example, be «*is_a*» (I), «*regulates*» (R) or «*part_of*» (P) (Graph by the Gene Ontology Consortium (CC-BY 4.0))

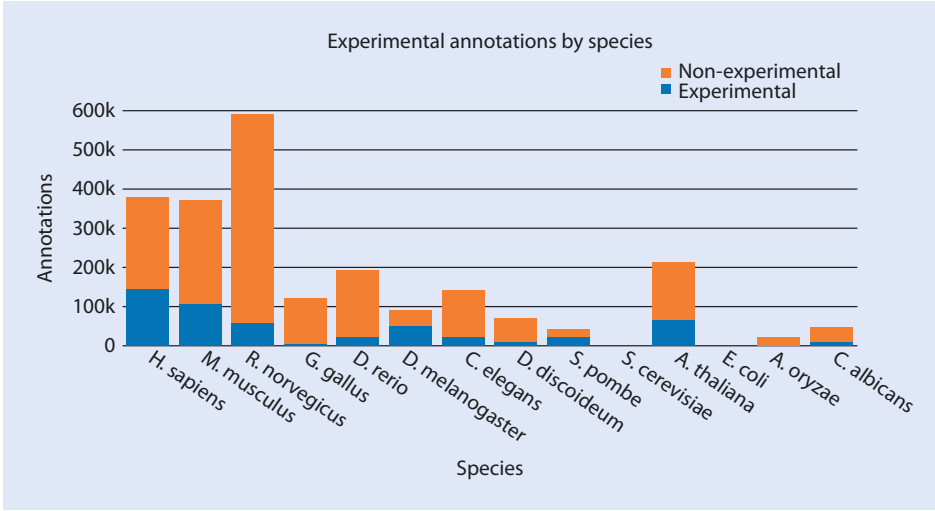


Fig. 7.7 Experimental annotations of model organisms in Gene Ontology (September 2016). Organisms displayed from *left* to *right*: *Homo sapiens*, *Mus musculus*, *Rattus norvegicus*, *Gallus gallus*, *Drosophila melanogaster*, *Caenorhabditis elegans*, *Dictyostelium discoideum*, *Saccharomyces pombe*, *Saccharomyces cerevisiae*, *Arabidopsis thaliana*, *Escherichia coli*, *Aspergillus oryzae*, *Candida albicans*. Plot by the Gene Ontology Consortium (CC-BY 4.0)

either by increased deposition or by increased number of cells.» This term is part of the biological process ontology and is identified as GO: 0043473. Every gene (or transcript) annotated with GO can have several terms in each of the ontologies. For example, the *PAPLA1* gene (phosphatidic acid phospholipase A1) of *Drosophila melanogaster* is annotated with eight different terms belonging to all three ontologies. Currently more than 40,000 terms are in use, and as GO is a dynamic ontology, users are allowed to describe new terms or delete old ones, if necessary (The Gene Ontology Consortium 2015).

Using the GO database (▶ http://geneontology.org/), it is possible to search for all terms in the ontology which are connected to a queried term (e.g., Fig. 7.6), all genes and gene products which are annotated with this term and all annotations using this term. Based on GO-annotated reference organisms, all newly sequenced genomes or transcriptomes can be annotated with GO terms under the assumption of the ortholog conjecture. However, it should be kept in mind that most GO annotations are based on data from a few model organisms and most of them are not validated experimentally (Fig. 7.7). The advantages of ontologies are that big data sets can be computationally analysed across taxa, tissues or developmental stages. Typical questions are, for example, if genes with specific GO terms are enriched in transcriptomes of selected tissues or stages. Several software tools are available for such enrichment analyses (Huang et al. 2009).

7.7 Whole-Genome Duplications

Not only duplications of single genes or segments of genes are frequently occurring but also duplication of the whole genome. Also known as polyploidy, genome duplications can occur due to failure of cell division after mitotic doubling (genomic doubling), failure of

cell division during meiosis (gametic nonreduction) or by polyspermy (Otto and Whitton 2000). However, all these events change the ploidy status of its bearer, which might inhibit the production of viable gametes in sexual species. Polyploidization events seem especially widespread in plants, but are also well-documented for other eukaryote lineages (Lynch 2007). Ancient genome duplications are not easy to detect, as whole-genome duplication (wgd) events are usually followed by genomic instability characterized by massive genome rearrangements and gene loss. For example, in a yeast species (*Saccharomyces cerevisiae*) which underwent wgd, only ~10% of the duplicated genes remained in the genome (Kellis et al. 2004). Mapping wgd events on phylogenies shows an intriguing pattern, as they are found in some cases at the base of evolutionary radiations (Van de Peer et al. 2009). Already Ohno (1970) put forward the idea that two rounds of wgd (2R) occurred in the lineage leading to ray-finned fishes and tetrapods. This idea has been confirmed by the sequencing of complete genomes of many vertebrate species (Van de Peer et al. 2009). Additional rounds of genome duplications are reported for teleosts and other ray-finned fish lineages, including the widely used models zebra fish (*Danio rerio*) and puffer fish (*Takifugu rubripes*) which genomes both show evidence for three wgd events (Meyer and Van de Peer 2005). Likewise, the megadiverse flowering plants (angiosperms) arose from a lineage that underwent several wgd events (De Bodt et al. 2005). It has been proposed that wgd paved the way for evolutionary innovation, as duplicated genes became the chance to evolve new functions. Moreover, wgd might trigger an increased speciation rate, as reciprocal gene loss in separated populations could lead to genetic isolation (Van de Peer et al. 2009; Werth and Windham 1991). Finally, wgd could contribute to a reduced extinction risk due to functional redundancy and mutational robustness (Crow and Wagner 2006). However, some of these hypotheses have been doubted. For example, a relationship between genome duplication and increased rates of lineage diversification in teleost fishes was not supported when inspecting the fossil record (Clarke et al. 2016).

References

Altenhoff AM, Dessimoz C (2009) Phylogenetic and functional assessment of orthologs inference projects and methods. PLoS Comput Biol 5:e1000262

Altenhoff AM, Dessimoz C (2012) Inferring orthology and paralogy. In: Anisimova M (ed) Evolutionary genomics: statistical and computational methods, vol 1. Humana Press, Totowa, pp 259–279.

Altenhoff AM, Škunca N, Glover N, Train C-M, Sueki A, Piližota I, Gori K, Tomiczek B, Müller S, Redestig H, Gonnet GH, Dessimoz C (2015) The OMA orthology database in 2015: function predictions, better plant support, synteny view and other improvements. Nucleic Acids Res 43:D240–D249

Arvestad L, Berglund A-C, Lagergren J, Sennblad B (2003) Bayesian gene/species tree reconciliation and orthology analysis using MCMC. Bioinformatics 19:i7–i15

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G (2000) Gene ontology: tool for the unification of biology. Nat Genet 25:25–29

Brower AVZ, Schawaroch V (1996) Three steps of homology assessment. Cladistics 12:265–272

Chen X, Zhang J (2012) The ortholog conjecture is untestable by the current gene ontology but is supported by RNA sequencing data. PLoS Comput Biol 8:e1002784

Chen F, Mackey AJ, Stoeckert CJ, Roos DS (2006) OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. Nucleic Acids Res 34:D363–D368

Clarke JT, Lloyd GT, Friedman M (2016) Little evidence for enhanced phenotypic evolution in early teleosts relative to their living fossil sister group. Proc Natl Acad Sci U S A 113(41):11531–11536

Cristianini N, Hahn MW (2007) Introduction to computational genomics. A case studies approach. Cambridge University Press, Cambridge

Crow KD, Wagner GP (2006) What is the role of genome duplication in the evolution of complexity and diversity? Mol Biol Evol 23:887–892
De Bodt S, Maere S, Van de Peer Y (2005) Genome duplication and the origin of angiosperms. Trends Ecol Evol 20:591–597
de Pinna MCC (1991) Concepts and tests of hoology in the cladistic paradigm. Cladistics 7:367–394
Doolittle RF (1986) Of URFs and ORFs: a primer on how to analyze derived amino acid sequences. University Science Books, Mill Valley
Drosophila 12 Genomes Consortium (2007) Evolution of genes and genomes on the *Drosophila* phylogeny. Nature 450:203–218
Eddy SR (1996) Hidden Markov models. Curr Opin Struct Biol 6:361–365
Eisen JA (1998) Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis. Genome Res 8:163–167
Emms DM, Kelly S (2015) OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biol 16:157
Finn RD, Mistry J, Schuster-Böckler B, Griffiths-Jones S, Hollich V, Lassmann T, Moxon S, Marshall M, Khanna A, Durbin R, Eddy SR, Sonnhammer ELL, Bateman A (2006) Pfam: clans, web tools and services. Nucleic Acids Res 34:D247–D251
Finn RD, Clements J, Eddy SR (2011) HMMER web server: interactive sequence similarity searching. Nucleic Acids Res 39:W29–W37
Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, Potter SC, Punta M, Qureshi M, Sangrador-Vegas A, Salazar GA, Tate J, Bateman A (2016) The Pfam protein families database: towards a more sustainable future. Nucleic Acids Res 44:D279–D285
Fitch WM (1970) Distinguishing homologous from analogous proteins. Syst Biol 19:99–113
Fitch WM (2000) Homology: a personal view on some of the problems. Trends Genet 16:227–231
Force A, Lynch M, Pickett FB, Amores A, Y-I Y, Postlethwait J (1999) Preservation of duplicate genes by complementary, degenerative mutations. Genetics 151:1531–1545
Gabaldón T (2008) Large-scale assignment of orthology: back to phylogenetics? Genome Biol 9:1–6
Gabaldon T, Koonin EV (2013) Functional and evolutionary implications of gene orthology. Nat Rev Genet 14:360–366
Gerstein MB, Bruce C, Rozowsky JS, Zheng D, Du J, Korbel JO, Emanuelsson O, Zhang ZD, Weissman S, Snyder M (2007) What is a gene, post-ENCODE? History and updated definition. Genome Res 17:669–681
Hahn MW, Han MV, Han S-G (2007) Gene family evolution across 12 *Drosophila* genomes. PLoS Genet 3:e197
Huang DW, Sherman BT, Lempicki RA (2009) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res 37:1–13
Huerta-Cepas J, Szklarczyk D, Forslund K, Cook H, Heller D, Walter MC, Rattei T, Mende DR, Sunagawa S, Kuhn M, Jensen LJ, von Mering C, Bork P (2016) eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. Nucleic Acids Res 44:D286–D293
Innan H, Kondrashov F (2010) The evolution of gene duplications: classifying and distinguishing between models. Nat Rev Genet 11:97–108
Jensen LJ, Julien P, Kuhn M, von Mering C, Muller J, Doerks T, Bork P (2008) eggNOG: automated construction and annotation of orthologous groups of genes. Nucleic Acids Res 36:D250–D254
Johannsen W (1909) Elemente der exakten Erblichkeitslehre. Gustav Fischer Verlag, Jena
Kellis M, Birren BW, Lander ES (2004) Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*. Nature 428:617–624
Koonin EV (2005) Orthologs, paralogs, and evolutionary genomics. Annu Rev Genet 39:309–338
Kristensen DM, Wolf YI, Mushegian AR, Koonin EV (2011) Computational methods for gene orthology inference. Brief Bioinform 12:379–391
Krogh A, Brown M, Mian IS, Sjölander K, Haussler D (1994) Hidden markov models in computational biology. J Mol Biol 235:1501–1531
Kuzniar A, van Ham RCHJ, Pongor S, Leunissen JAM (2008) The quest for orthologs: finding the corresponding gene across genomes. Trends Genet 24:539–551
Lankester ER (1870) On the use of the term homology in modern zoology, and the distinction between homogenetic and homoplastic agreements. Ann Mag Nat Hist 6:34–43
Li L, Stoeckert CJ, Roos DS (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res 13:2178–2189

Lynch M (2007) The origins of genome architecture. Sinauer Assoc, Sunderland
Meyer A, Van de Peer Y (2005) From 2R to 3R: evidence for a fish-specific genome duplication (FSGD). BioEssays 27:937–945
Nehrt NL, Clark WT, Radivojac P, Hahn MW (2011) Testing the ortholog conjecture with comparative functional genomic data from mammals. PLoS Comput Biol 7:e1002073
Ohno S (1970) Evolution by gene duplication. Springer, Berlin
Orgogozo V, Peluffo AE, Morizot B (2016) Chapter 1. The «Mendelian Gene» and the «Molecular Gene»: two relevant concepts of genetic units. In: Virginie O (ed) Current topics in developmental biology, vol 119. Academic Press, p 1–26.
Otto SP, Whitton J (2000) Polyploid incidence and evolution. Annu Rev Genet 34:401–437
Owen R (1843) Lectures on the comparative anatomy and physiology of the invertebrate animals. Longman, Brown/Green/Longmans/London
Park J, Karplus K, Barrett C, Hughey R, Haussler D, Hubbard T, Chothia C (1998) Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods. J Mol Biol 284:1201–1210
Rabiner LR (1989) A tutorial on hidden Markov models and selected applications in speech recognition. Proc IEEE 77:257–286
Rastogi S, Liberles DA (2005) Subfunctionalization of duplicated genes as a transition state to neofunctionalization. BMC Evol Biol 5:28
Remane A (1952) Die Grundlagen des natürlichen Systems, der vergleichenden Anatomie und der Phylogenetik. Akademische Verlagsgesellschaft Geest und Portig, Leipzig
Remm M, Storm CEV, Sonnhammer ELL (2001) Automatic clustering of orthologs and in-paralogs from pairwise species comparisons1. J Mol Biol 314:1041–1052
Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Söding J, Thompson JD, Higgins DG (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal omega. Mol Syst Biol 7:539
Sonnhammer ELL, Koonin EV (2002) Orthology, paralogy and proposed classification for paralog subtypes. Trends Genet 18:619–620
Sonnhammer ELL, Östlund G (2015) InParanoid 8: orthology analysis between 273 proteomes, mostly eukaryotic. Nucleic Acids Res 43:D234–D239
Stewart EJ (2012) Growing unculturable bacteria. J Bacteriol 194:4151–4160
Tatusov RL, Galperin MY, Natale DA, Koonin EV (2000) The COG database: a tool for genome-scale analysis of protein functions and evolution. Nucleic Acids Res 28:33–36
Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, Krylov DM, Mazumder R, Mekhedov SL, Nikolskaya AN, Rao BS, Smirnov S, Sverdlov AV, Vasudevan S, Wolf YI, Yin JJ, Natale DA (2003) The COG database: an updated version includes eukaryotes. BMC Bioinformatics 4:41
The Gene Ontology Consortium (2015) Gene ontology consortium: going forward. Nucleic Acids Res 43:D1049–D1056
Thomas PD, Wood V, Mungall CJ, Lewis SE, Blake JA, on behalf of the Gene Ontology C (2012) On the use of gene ontology annotations to assess functional similarity among orthologs and paralogs: a short report. PLoS Comput Biol 8:e1002386
Van de Peer Y, Maere S, Meyer A (2009) The evolutionary significance of ancient genome duplications. Nat Rev Genet 10:725–732
van der Heijden RT, Snel B, van Noort V, Huynen MA (2007) Orthology prediction at scalable resolution by phylogenetic tree analysis. BMC Bioinformatics 8:83
van Dongen S (2000) Graph clustering by flow simulation. Universiteit Utrecht, Utrecht
Webber C, Ponting CP (2004) Genes and homology. Curr Biol 14:R332–R333
Werth CR, Windham MD (1991) A model for divergent, allopatric speciation of polyploid pteridophytes resulting from silencing of duplicate-dene expression. Am Nat 137:515–526
Wheeler TJ, Clements J, Finn RD (2014) Skylign: a tool for creating informative, interactive logos representing sequence alignments and profile hidden Markov models. BMC Bioinformatics 15:7
Zhang J (2003) Evolution by gene duplication: an update. Trends Ecol Evol 18:292–298

Phylogenetic Analyses

C. Bleidorn, *Phylogenomics*, DOI 10.1007/978-3-319-54064-1_8

- Phylogenetic trees represent the evolutionary relationships of sequences or species (or other taxonomic units) and can be shown as phylograms, cladograms; ultrametric trees, or unrooted.
- Networks or consensus methods can be used to summarize the information of multiple trees.
- Many tree building methods rely on explicit models of sequence evolution; models of nucleotide substitution are nested and can be derived from a general model, whereas amino acid models are broadly classified into empirical and mechanistic models.
- The most widely applied methods of phylogenetic reconstruction are neighbour joining (NJ), maximum parsimony (MP), maximum likelihood (ML) and Bayesian inference (BI).
- Support of branches within a phylogenetic tree can be (among others) measured by bootstrapping or likelihood ratio test approaches.

8.1 Trees

The main aim of phylogenetic systematics is the reconstruction of evolutionary relationships which are represented by a tree. In a phylogenetic tree, the pattern of branches connected by internal nodes (topology) illustrates the relationships of the included terminals (Fig. 8.1). In Fig. 8.1, the terminals represent the taxa A to F. When describing the topology of a tree, it has to be kept in mind that rotation of the axis of internal nodes does not change the topological information (Fig. 8.2). A handy way to describe trees is to refer to clades (Fig. 8.1), which are monophyletic units comprising an ancestor (internal node) and all of its descendants. For example, taxa C and D form a monophyletic group in Fig. 8.1. Moreover, referring to sister lineages (or groups) is a good way to describe trees. For example, taxon C is the sister lineage (or group) of taxon D in Fig. 8.1. Often misused when describing trees is the term «basal» (Krell and Cranston 2004), e.g. when referring in Fig. 8.1 to taxon A as «basal». This is wrong, as basal would imply that taxon A is the ancestral group – which obviously cannot be correct for an extant taxon. However,

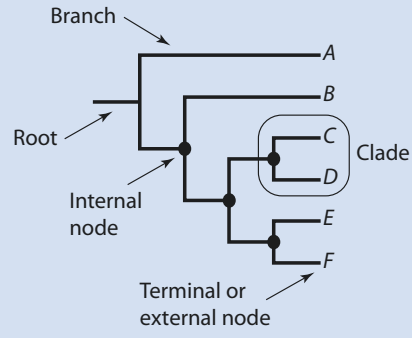


Fig. 8.1 Terms describing the topology of a phylogenetic tree

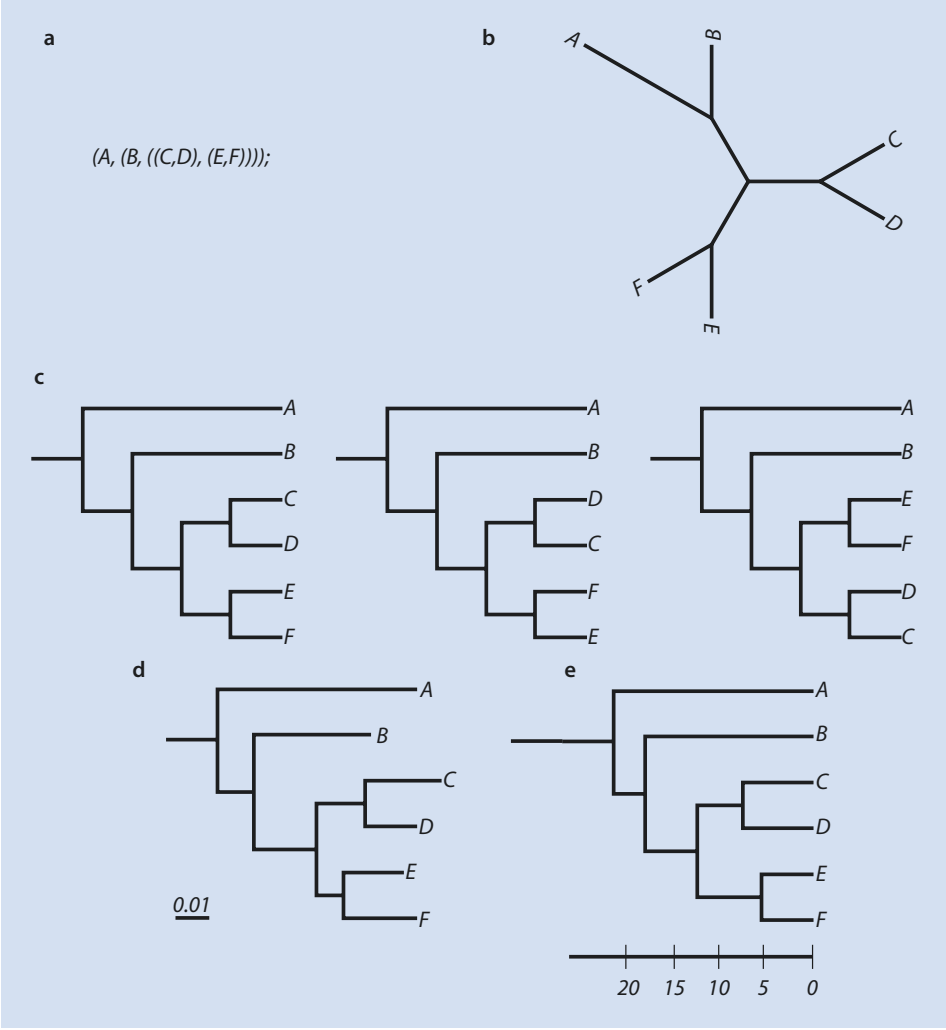


Fig. 8.2 Different ways to represent the same tree topology. **a** Newick format. **b** Unrooted tree. **c** Three different representations of the same topology as a cladogram. **d** Phylogram. **e** Ultrametric tree with time axis

taxon A could be described as «basally branching», as the branch leading to this terminal is closest to the root.

One major problem for phylogenetic analyses is the giant number of possible tree topologies with increasing number of taxa. The number of possible strictly bifurcating unrooted trees (N_u) can be calculated as follows:

$$N_u = (2n-5)! \div 2^{n-3} (n-3)! \qquad \textbf{(8.1)}$$

In this formula, *n* denotes the number of terminals (e.g. taxa, OTUs, sequences) included in the tree. For example, for 4 taxa, the number of unrooted tree topologies is 3, whereas for 10 taxa, there are already 2,027,025 different topologies. The number of possibilities grows exponentially with the number of included terminals. When including 60 terminals, there are more possible tree topologies ($\sim 10^{94}$) than atoms in the universe ($\sim 10^{82}$). This sheer incomprehensible large number of possibilities is a major problem for all phylogenetic methods which include steps analysing all of them.

There are different ways to represent the topology of a tree (Fig. 8.2). The standard output format of most phylogenetic software is the Newick format, where nested relationships are shown using brackets (Fig. 8.2a). This format can easily be translated in an unrooted tree (Fig. 8.2b). Unrooted trees can be polarized by choosing outgroups (Nixon and Carpenter 1993). This choice usually depends on prior knowledge, and correct outgroup choice is crucial for every analyses. If outgroups are unknown or not included in the analyses, midpoint rooting can be alternatively used for rooting. In this case, the root is placed at the midpoint of the longest distance between two terminals in a tree. However, midpoint rooting assumes that all included terminals evolve at the same rate and may fail to place the root correctly when this assumption is violated (Hess and De Moraes Russo 2007). Finally, the root of the tree could be inferred as part of the phylogenetic analysis when using nonreversible models of sequence evolution (see below), where different placements of the root affect the outcome of the analysis (Huelsenbeck et al. 2002a). If trees are represented as cladograms (Fig. 8.2c), they only contain topological information. However, if they are represented as phylograms (Fig. 8.2d) or ultrametric trees (Fig. 8.1e), the length of branches carries additional information. In phylograms, branch length is proportional to evolutionary change. A typical measure of branch lengths for molecular data is the average number of substitutions per site in the alignment. The sum of the lengths of the branches linking two terminals (but also internal nodes) in a phylogram is called patristic distance (Fourment and Gibbs 2006). Ultrametric trees are reconstructed under the assumption that the change indicated by the length of branches is proportional to time («molecular clock», ▶ see also Sect. 8.7). Terminals in an ultrametric tree are equidistant from the root, which means that all paths of branches leading from the root to terminals have the same length. Phylogenetic divergence times can be estimated by calibrating ultrametric trees using palaeontological or biogeographic data (Donoghue and Benton 2007; Heads 2005). There are several applications available to visualize trees, and among the most widely used ones are DENDROSCOPE (Huson et al. 2007), ETE (Huerta-Cepas et al. 2010), FIGTREE (▶ http://tree.bio.ed.ac.uk/software/figtree/), ITOL (Letunic and Bork 2016) and TREEVIEW (Page 1996).

Sometimes, it is desirable to summarize the information of two or more topologies in a consensus tree. Several methods for building consensus trees are available (Wilkinson 1994), but only two of them are widely used in phylogenetic systematics: strict consensus and majority-rule consensus (Fig. 8.3). In a strict consensus, only those internal nodes that can be found in all summarized topologies are displayed (Fig. 8.3b); all other nodes are collapsed into multifurcations. Majority-rule consensus trees show those internal nodes which are found in more than half of all summarized topologies (Fig. 8.3c); nodes that do not fulfil this criterion are collapsed. Usually, the frequency how often a node appears is indicated in the tree. Majority-rule consensus trees are widely used to summarize trees from bootstrap analyses and Bayesian inference (▶ see Sect. 8.6). Finally, there is a set of methods that derive a phylogenetic hypothesis (tree topology) from combining the topological information of different source trees. This so-called supertree approach differs from consensus methods, as it does not need an identical set of terminals to combine

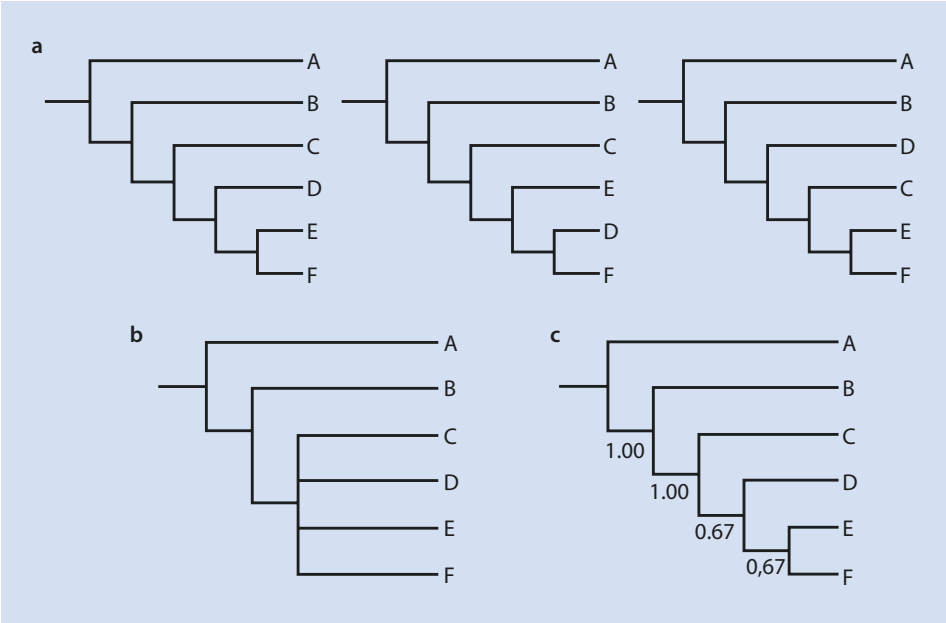


Fig. 8.3 Consensus tree methods. **a** Cladograms of three different topologies. **b** Strict consensus, summarizing those nodes found in all trees. **c** Majority-rule consensus, summarizing those nodes which are found in more than 50% of the trees. Frequency of the occurrence of nodes is given at the nodes

trees. Instead, supertrees can be built from topologies which overlap in its included terminals (Bininda-Emonds 2004). Even though supertrees can be build using a modified version of the strict consensus, by far the most widespread attempt is using a form of matrix representation (Baum 1992; Ragan 1992). Under this approach, all internal nodes of the input trees are coded as characters in a matrix. Each terminal which appears in at least one of the trees will be included in the matrix, and either are coded as present (1) or absent (0) for each character. Finally, distance or parsimony methods (▶ see Sect. 8.5) can be used to reconstruct the supertree.

Every tree is a special kind of a graph. A graph can be broadly defined as a representation of a finite number of nodes connected by branches (edges) to show their relationships (Huson et al. 2010). Trees are connected graphs without cycles, which means there are no reticulations. However, phylogenetic trees might not always be the best way to represent evolutionary relationships (Doolittle and Bapteste 2007). For example, under the presence of hybridization, horizontal gene transfer or recombination reticulate relationships between nodes should be assumed. In this case, networks (Fig. 8.4), which are connected graphs with cycles, are a better way to illustrate evolutionary relationships (Huson and Bryant 2006; Posada and Crandall 2001). Such networks can also be used to visualize conflict within phylogenetic datasets.

8.2 Models of Nucleotide Substitution

Under the assumption of a constant evolutionary rate over time, a linear increase of the number of nucleotide substitutions should be expected after divergence of a pair of sequences. However, as there might be back substitutions, multiple substitutions or

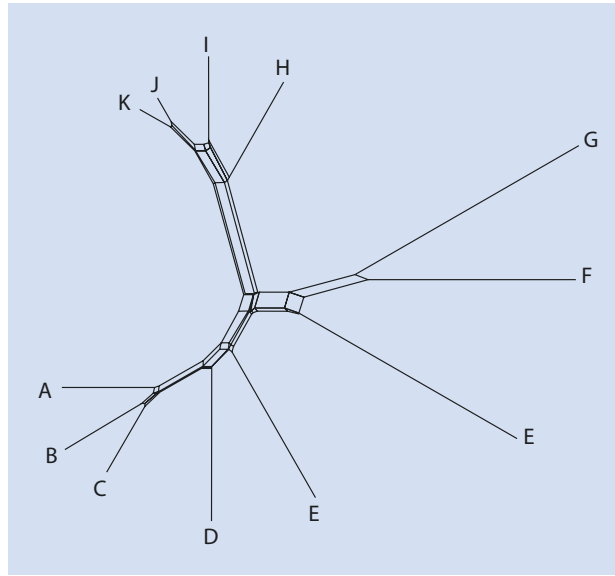


Fig. 8.4 Example of a phylogenetic network

convergent substitutions, comparison of observed distances (p-distances) between pairs of sequences will show a level of saturation after some time of divergence (Page and Holmes 1998). To correct for this saturation, probabilistic models of sequence evolution are used to calculate expected distances. Most methods for phylogenetic reconstruction rely on explicitly formulated models of sequence evolution. Such models are incorporated within distance methods, maximum likelihood and Bayesian inference (▶ see Sect. 8.5). Nucleotide substitution models in use for phylogenetic inference make several assumptions to model substitutions as a stochastic process:

I. For every site of a sequence, it is assumed that the rate of change from one base to another is independent from the history of this site (Markov property).
II. It is assumed that substitution rates are not changing over time (homogeneity).
III. Equilibrium of base frequencies is assumed (stationarity).

Models that fit this description are called time-homogeneous time-continuous stationary Markov models. Substitution rates are summarized by such models in a rate matrix (or Q-matrix), where each entry specifies the probability for any possible nucleotide substitution. Usually, models used in molecular phylogenetics are time reversible, thereby additionally assuming that the rate of change from one base i to another base j is identical to the rate of a change from j to i (j and i can be all possible bases, but must be different bases). The most general model of nucleotide substitutions is the general time reversible (GTR) model (Rodríguez et al. 1990; Tavare 1986), which is summarized by the following Q-matrix:

$$Q = \begin{bmatrix} -\mu(a\pi_C + b\pi_G + c\pi_T) & \mu a\pi_C & \mu b\pi_G & \mu c\pi_T \\ \mu a\pi_A & -\mu(a\pi_A + d\pi_G + e\pi_T) & \mu d\pi_G & \mu e\pi_T \\ \mu b\pi_A & \mu d\pi_C & -\mu(b\pi_A + d\pi_C + f\pi_T) & \mu f\pi_T \\ \mu c\pi_A & \mu e\pi_C & \mu f\pi_G & -\mu(c\pi_A + e\pi_C + f\pi_G) \end{bmatrix} \quad \textbf{(8.2)}$$

Formula 8.2: Q-matrix of the GTR model. Herein, μ is the overall substitution rate, and π_X refers to the different base frequencies (with X either A, C, G or T). The letters a to f represent the frequency of possible substitutions (e.g. a is the frequency of substitutions from A to C or C to A).

The GTR model infers parameters for six different (reversible) substitution types (A–C, A–G, A–T, C–G, C–T, G–T), overall substitution rate and base frequency from the underlying data (a sequence alignment). For the rate of substitution types, one of these parameters is set to 1, whereas all other parameters are relative to the fixed parameter. Thereby in this case, there are five free parameters for the substitution types. There are several other models in use (Fig. 8.5), which can be derived from the GTR model by

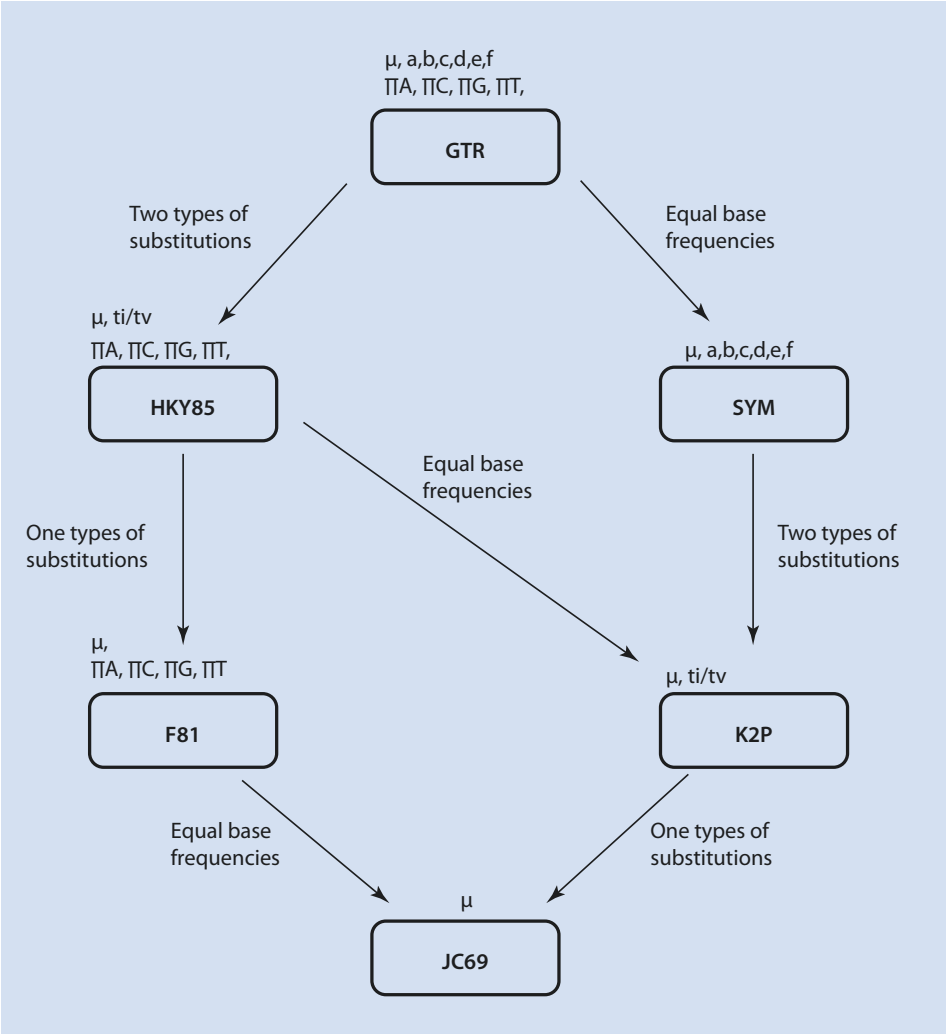


Fig. 8.5 Models of nucleotide substitutions and their interrelationships. Models are nested and can be derived from the general time reversible (GTR) model by restricting parameters. Open parameters are given above the boxes for each model. Herein μ is the overall substitution rate; π_X refers to the different base frequencies (with X either A, C, G or T). The letters a to f represent the frequency of possible substitutions, whereas ti/tv represents the frequency of transitions to transversions. The restricted parameter to transform a more general model to a more restricted one is given at the *arrows*

restricting some of the parameters. For example, in the HKY85 model (Hasegawa et al. 1985), only two types of substitutions (transitions vs. transversions) are distinguished. In the K2P model (Kimura 1980) also, only these two substitution types are distinguished, but additionally equal base frequencies (0.25 for each base) are assumed. The most restricted (and historically oldest) model is the JC69 model (Jukes and Cantor 1969), where all substitution types are assumed to be equally probable and base frequencies are fixed to be equal. A detailed description (including Q-matrices) of models for nucleotide substitutions can be found in Yang (2006) and Page and Holmes (1998).

All discussed models assume that the evolutionary rate is the same for every position of the sequence alignment. However, this assumption has been shown to be unrealistic when working with real data. As such the mutation rate can vary among bases. For example, G and C nucleotides are twice as mutable than A and T nucleotides in most species across the tree of life (Hodgkinson and Eyre-Walker 2011). This is due to the effect that in CpG dinucleotides (a C followed by a G) cytosines are often methylated and thereby prone to deamination, resulting in a T nucleotide (Fryxell and Moon 2005). For most datasets, the rate of fixation of mutations also varies among sites, due to the effect of different selective pressures. Obvious examples are protein-coding genes, where the codon positions evolve under different rates, with the third position usually accumulating substitutions much faster than the other two positions. Likewise, different selective pressures act on different regions of ribosomal RNA genes, and usually conserved and variable regions can be distinguished. By ignoring these variations across sites, the expected distance between a pair of sequence will be underestimated. To include rate heterogeneity across alignment sites, a statistical distribution is used to allow different sites to fall into categories of different substitution rates. Usually, a «gamma model» is used, with several categories of rates approximating a gamma distribution (Yang 1994). The shape of the gamma distribution is defined by the parameter α (Fig. 8.6), which has to be determined to fit the gamma model for a given dataset. Typically, the shape parameter α is rather small (<1) (Yang 1996) resulting in a skewed L-distribution, reflecting that most of the sites show low substitution rates (or are invariable), whereas few sites range in a spectrum from low to high substation rates (Yang 2006). Large values of α would result in a rather bell-shaped distribution where most sites evolve under a similar rate. The continuous gamma distribution can be divided into categories of equal probability. Based on a comparison of several datasets, the use of six to ten rate categories has been suggested as a good approximation of rate heterogeneity (Jia et al. 2014). Models of sequence evolution which incorporate the gamma distribution are marked with a «+ Γ» or «+ G». Inclusion of a gamma distribution is computationally time and memory intensive, which can be a problem for large-scale phylogenomic analyses. Stamatakis (2006) developed a method to approximate rate heterogeneity called «CAT model» which is computationally much faster than inferring the gamma distribution. This approach is implemented in the popular software RAXML (Stamatakis 2014).

Another modification to account for rate heterogeneity is the incorporation of the proportion of invariant sites into models of sequence evolution (Fitch and Margoliash 1967b). Models using this modification are marked with a «+ I». If all alignment sites would change at the same rate, as assumed by all models discussed here, the number of substitutions should follow a Poisson distribution. Real datasets usually do not fit this distribution. However, the exclusion of invariant sites allows a better fit. Models including both modifications (+ I + Γ) assume that a proportion of sites are invariable, while the rates of the remaining sites are gamma distributed (Gu et al. 1995). It is discussed if such kind of models should be used at all, given that the amount of invariable

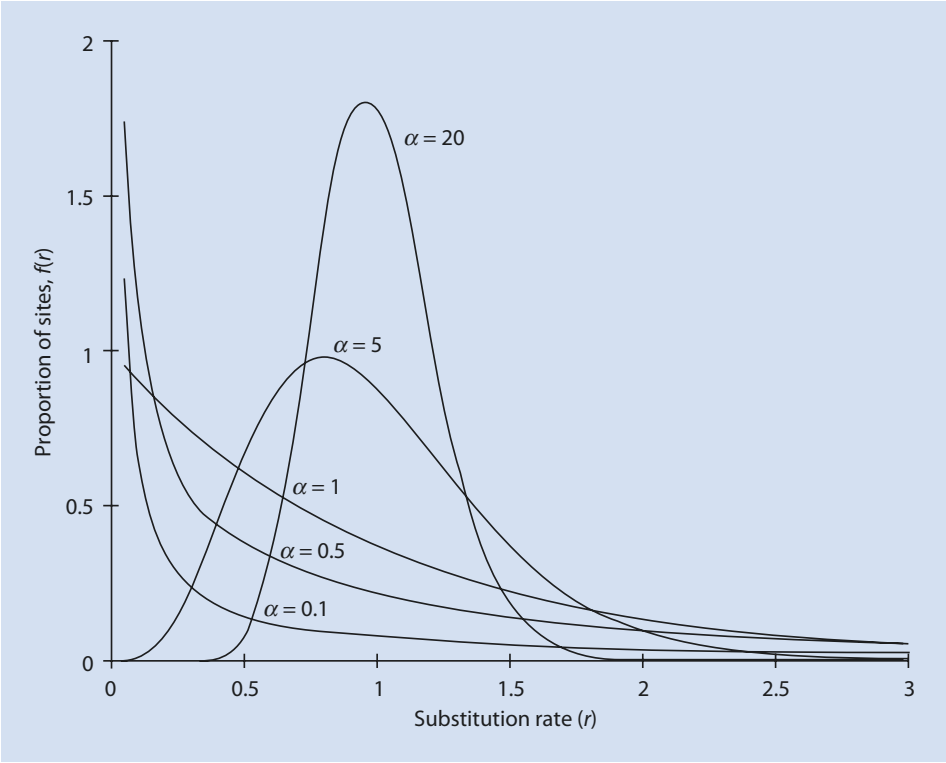


Fig. 8.6 Probability density function of the gamma distribution. The parameter α defines the shape of the distribution. The *x*-axis shows the substitution rate, whereas the *y*-axis represents the proportion of number of sites with this rate (Reprinted from (Yang 1996) with permission from Elsevier)

sites is already included by the gamma distribution and thereby the estimation of parameters for both model modifications is not independent (Sullivan et al. 1999). However, simulation studies and comparisons of real datasets found that models including both parameters (+ I + Γ) often improve phylogenetic analyses (Jia et al. 2014; Kück et al. 2012).

As outlined, two of the main assumptions of Markov models are stationarity and homogeneity of the data. Moreover, the so far discussed models also assume reversibility, meaning that the probability of a substitution from character i to character j equals the reverse case. However, real data often violates one or several of these assumptions (see above). A model not using these assumptions is the general Markov model (GMM), which has been formalized for phylogenetics by Barry and Hartigan (1987). However, this model is in practice too complex, as it requires the reliable estimation for more parameters than usually available for a given dataset. Two more restricted versions of the GMM were developed by Jayaswal et al. (2011), with one model assuming stationarity, while being nonreversible and non-homogeneous (SBH model), and another model assuming stationarity and reversibility, while being non-homogeneous (RBH model). It has been demonstrated for an example dataset that these models have a better fit to the data than standard GTR models or its derivatives (Jayaswal et al. 2011). However, their use was so far restricted to smaller datasets due to the computational complexity.

8.3 Models of Amino Acid Substitutions

Most models for amino acid substitutions can be broadly classified into two classes: empirical models and mechanistic models (Yang 2006). Empirical models are usually derived from a large compilation of sequence alignments. These models are summarized in an amino acid replacement matrix, where each entry is corresponding to the relative rate of replacement of one amino acid by another. The first empirical matrices have been published by Margaret Dayhoff and colleagues (Dayhoff et al. 1972, 1978). The matrices were compiled from available protein alignments of similar sequences which did not differ in more than 15% of their sites. Altogether, 34 protein superfamilies split into 71 alignments have been analysed. For each alignment, a phylogenetic tree using maximum parsimony (▶ see Sect. 8.5) has been created, where internal nodes represent ancestral protein sequences. Mapping all changes on the tree allowed inferring the number of amino acid replacements for all possible pairs. All changes were inferred from sequences with a high identity to reduce the probability of multiple substitutions. As such, all entries of the corresponding matrix are regarded for an evolutionary time interval of 1 amino acid change per 100 amino acid sites (◘ Fig. 8.7). This matrix is known as PAM1 (point accepted mutations) matrix. To derive matrices for sequences separated by a longer time (and experienced more change), the PAM1 matrix can be multiplied by itself. A widely used PAM matrix is the PAM250 matrix (Dayhoff et al. 1978), which has found to be reliable for sequence which differ in up to 80% of their sites.

Jones et al. (1992) published an update of the Dayhoff matrices, based on a much larger database including newly available sequence alignments that fulfilled the original chosen requirements of 85% identity. By using distance methods instead of maximum parsimony, also a slightly different methodology to select pairs of sequences for the final analyses was used. This widely used matrix is known as the JTT model of amino acid substitutions. For both matrices, Dayhoff and JTT, the included phylogenetic analyses methods to count changes along the tree for the generation of the substitution model have been harshly criticized (Whelan and Goldman 2001). As such, it has been noted that both approaches likely underestimate the number of replacements (even for highly similar sequences). Instead, a maximum likelihood (▶ see Sect. 8.5) approach has been proposed, which avoids the outlined problems of the discussed models. This approach is able to use sequences with different degrees of identity and also allows the occurrence of multiple changes (Yang et al. 1998). Using this methodology, Whelan and Goldman (2001) inferred an amino acid replacement matrix based on analysing 182 protein families which is known as the WAG model. A refinement of the WAG matrix using an updated and larger database including nearly 4000 alignments has been published by Le and Gascuel (2008) and is now referred to as the LG model. Later on, Le et al. (2012) introduced the use of different substitution matrices for site with different evolutionary rates. Two sets of matrices called LG4M and LG4X were estimated from a huge number of protein alignments and are used according to different gamma categories (LG4M) or a distribution-free scheme of rate heterogeneity (LG4X). The use of different substitution matrices for different sites (mixture models) has been generally proven to outperform the choice of a single substitution matrix for all sites (Le et al. 2008), but is computationally demanding. All these models are based on sequence data sampled across the tree of life. Several amino acid substitution models have been developed for specific taxa or organelles. These models are solely based on sequence comparisons from the taxon or organelle of interest. Besides others, models are available for mitochondria (MtRev)

		Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
		A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
Ala	A	9857	2	9	10	3	8	17	21	2	6	4	2	6	2	22	35	32	0	2	18
Arg	R	1	9913	1	0	1	10	0	0	10	3	1	19	4	1	4	6	1	8	0	1
Asn	N	4	1	9822	36	0	4	6	6	21	3	1	13	0	1	2	20	9	1	4	1
Asp	D	6	0	42	9859	0	6	53	6	4	1	0	3	0	0	1	5	3	0	0	1
Cys	C	1	1	0	0	9973	0	0	0	1	1	0	0	0	0	1	5	1	0	3	2
Gln	Q	3	9	4	5	0	9876	27	1	23	1	3	6	4	0	6	2	2	0	0	1
Glu	E	10	0	7	56	0	35	9865	4	2	3	1	4	1	0	3	4	2	0	1	2
Gly	G	21	1	12	11	1	3	7	9935	1	0	1	2	1	1	3	21	3	0	0	5
His	H	1	8	18	3	1	20	1	0	9912	0	1	1	0	2	3	1	1	1	4	1
Ile	I	2	2	3	1	2	1	2	0	0	9872	9	2	12	7	0	1	7	0	1	33
Leu	L	3	1	3	0	0	6	1	1	4	22	9947	2	45	13	3	1	3	4	2	15
Lys	K	2	37	25	6	0	12	7	2	2	4	1	9926	20	0	3	8	11	0	1	1
Met	M	1	1	0	0	0	2	0	0	0	5	8	4	9874	1	0	1	2	0	0	4
Phe	F	1	1	1	0	0	0	0	1	2	8	6	0	4	9946	0	2	1	3	28	0
Pro	P	13	5	2	1	1	8	3	2	5	1	2	2	1	1	9926	12	4	0	0	2
Ser	S	28	11	34	7	11	4	6	16	2	2	1	7	4	3	17	9840	38	5	2	2
Thr	T	22	2	13	4	1	3	2	2	1	11	2	8	6	1	5	32	9871	0	2	9
Trp	W	0	2	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	9976	1	0
Tyr	Y	1	0	3	0	3	0	1	0	4	1	1	0	0	21	0	1	1	2	9945	1
Val	V	13	2	1	1	3	2	2	3	3	57	11	1	17	1	3	2	10	0	2	9901

Fig. 8.7 Matrix containing probabilities of amino acid substitutions of proteins undergoing 1% of change (1 accepted point mutation in 100 amino acids, PAM1). All values are multiplied by 10,000. For example, there is a 0.02% probability that the amino acid arginine (*ARG*) will be replaced by alanine (*ALA*)

(Adachi and Hasegawa 1996), chloroplasts (CpRev) (Adachi et al. 2000) or retroviruses (RtRev) (Dimmic et al. 2002). Moreover, specific models have been inferred for mammal (MtMam) (Yang et al. 1998) or arthropod mitochondria (MtArt) (Abascal et al. 2007), to just name a few.

The so far described models have fixed substitution rates, corresponding to values in the matrix estimated from a large database. However, it is also possible to estimate these substation rates directly from the data using the GTR model. Given the high number of parameters to be estimated for a 20 × 20 substitution matrix (208 parameters!), this approach makes only sense for huge datasets.

As mentioned for nucleotide substitution models (see above), these models can again be modified according to a gamma distribution (+ Γ) and by including the amount of invariant sites (+ I). Usually, the amino acid frequencies are specified according to the chosen rate matrix. However, it is also possible estimating the specific amino acid frequencies for the analysed protein alignments (marked with «+ F»).

All so far described models assume homogeneity across sites, and the same underlying model of amino acid substitution applies to all sites (but see LG4M and LG4X). In contrast, the so-called CAT models (which have been unfortunately named the same way as the above-mentioned gamma distribution approximation implemented in RAXML) are site-heterogeneous models, which allow modelling of substitution pattern at different sites of a protein alignment by different substitution matrices (Lartillot and Philippe 2004). CAT models use a Dirichlet distribution (Antoniak 1974) to infer the number of different amino acid matrices with different frequencies, as well as the affiliation of each site to a given matrix (class). Different modifications of the CAT model are available. As such, relative amino acid substitution rates for the matrices used during the analyses can be either fixed (CAT model or CAT-F81 model) or estimated from the data during the analysis (CAT-GTR model). A further modification is the CAT-BP model (Blanquart and Lartillot 2008). This model allows the change of substitution models not only across sites (site heterogeneity) but also across lineages (time heterogeneity). This is facilitated by introducing breakpoints (BP) (Blanquart and Lartillot 2006), which allow switching between substitution matrices. In summary, CAT models infer the number of categories of rate heterogeneity and classify all sites of the alignment accordingly, while each category is modelled using its own relative rate matrix of amino acid substitutions. Besides applying it to amino acid data, CAT-GTR models have been furthermore used for analysing nucleotide data.

Mechanistic models include assumptions about biological processes (e.g. sorting amino acids into classes according to their chemical properties) or are formulated at the codon level. Especially codon models have been proven to outperform empiric models (Miyazawa 2013). However, they come with the computational burden of being extremely time-consuming to calculate, as they have to specify a matrix of 61 × 61 possible codon transformations (stop codons are not included) (Zaheri et al. 2014). Widely used is a simplified version of the codon model proposed by Goldman and Yang (1994). In this model, parameters are estimated for codon pair comparisons. A rate of 0 is applied for codons which differ in two or three positions. Separate rates are estimated for codons differing in only one position. In this case, different rates are estimated for synonymous and non-synonymous transversions, as well as for synonymous and non-synonymous transitions (Ren et al. 2005). Moreover, the frequency of codons can be handled differently for this model, either it assumed that all codons have the same frequency (Fequal) or the frequency is estimated based on a set of nucleotide frequencies (F1 × 4), or estimated from three sets of nucleotide

frequencies for each codon position (F3 × 4) or estimated directly as codon frequency (F61) (Yang 2006). The names in brackets refer to the number of free parameters to be estimated, which range from 1, over 4 and 12, to 61. Further, different codon models have been proposed (Zaheri et al. 2014). Similar to empirical models for amino acid substitutions, an empirical model of codon substitution has been inferred based on more than 17,000 alignments (Schneider et al. 2005). Kosiol et al. (2007) combined the approach described above with knowledge from empirical models regarding amino acid replacement rates based on physicochemical properties. Finally, Zaheri et al. (2014) published a generalized codon model based on a reduced number of parameters. Whereas most codon models are still computationally too intensive for phylogenomic analyses of large datasets, they are frequently used when detecting adaptive molecular evolution in single genes (Yang and Bielawski 2000), e.g. implemented in the software PAML (Yang 2007).

8.4 Model Selection and Data Partitions

8.4.1 Model Selection

The selection of the best fitting model for any given dataset is a crucial step in phylogenetics (Anisimova et al. 2013). Nucleotide substitution models differ in the number of parameters estimated from the dataset – and therefore in the way of realistically describing the data. However, there is a trade-off. More parameters allow a more realistic way of representing the underlying data. But this comes with the danger that too many parameters may over-fit the underlying data (overparametrization), resulting in errors during parameter estimation (Sullivan and Joyce 2005). In contrast, simplified models may not realistically represent the data, which can also mislead phylogenetic reconstruction. In the case of amino acids, empirical models mostly differ regarding the database they were compiled from. Moreover, some models have been specially designed for certain taxa or organellar proteins. Besides this, for all models, the question arises if they should account for rate heterogeneity (+ Γ) and invariant sites (+ I), as well as if the frequency of amino acids should be estimated from the data (+ F). Obviously, methods are needed to select a model that fits the underlying data while trying to avoid overparametrization. The most widely used methods to choose among models are the hierarchical likelihood ratio test (hLRT), the Bayesian information criterion (BIC) and the Akaike information criterion (AIC).

A popular approach using hLRT to select the best fitting nucleotide model has been implemented in a software called MODELTEST (Posada and Crandall 1998), which later on also was updated (JMODELTEST2) in including more models and other selection criteria (Darriba et al. 2012). The basic idea behind the hLRT approach is to calculate the likelihood (► see Sect. 8.5) for a fixed topology (e.g. a simple distance tree of the alignment to be evaluated) given the selected model and compare it with the likelihood for an alternative model:

$$\delta = 2\left(\ln L_1 - \ln L_0\right) \tag{8.3}$$

In this formula, L_1 represents the more complex model (in terms of free parameters) and L_0 the alternative model. The more complex model (which always will result in a better

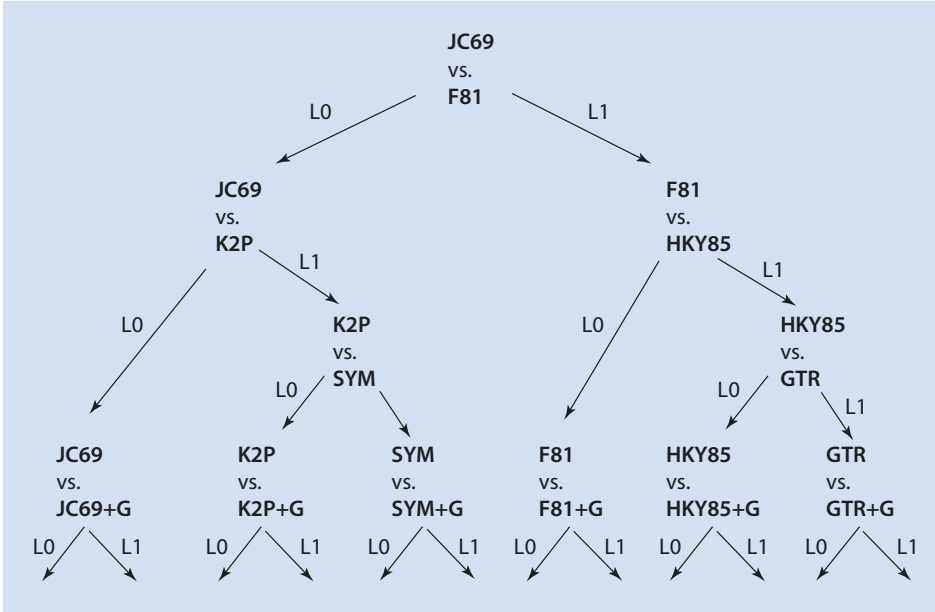


Fig. 8.8 Hierarchical tree for model testing using hLRT as implemented in MODELTEST (Posada and Crandall 1998). The test starts with the comparison of the two least complex models, and progresses along the tree. If the more complex model is chosen, the *arrow* L1 has to be followed. In case of choosing the less complex model, L0 has to be followed. The full tree includes all models and modifications (+I, +G, +I +G) (not shown), and the model is chosen when testing arrives at the bottom of the tree

likelihood value) will be chosen, if the value δ is regarded as significant when evaluated by a χ^2 test statistic, where the degree of freedom equals the difference in the number of free model parameters. For example, in the JC69 model, there is one free parameter (μ) to be estimated, whereas in the HKY85 model, there are five free parameters (μ, ti/tv and three base frequency parameters, whereas the frequency of the fourth will be set to add up to 1). Always two models are compared and can be tested along a tree-like hierarchy (Fig. 8.8). Starting with the comparison of the least complex models (JC69 vs. F81), tests are conducted following the tree hierarchy until a model is selected.

By using the hLRT, two models are compared at a time. In contrast, using the information criteria AIC (Akaike 1973) or BIC allows to simultaneously compare all considered models. Moreover, for hLRT, it is important that the models are nested, which means they can be transformed into each other by restricting or opening parameters, as it is the case for nucleotide substitution models. However, amino acid models do not fulfil this criterion. AIC and BIC are able to compare nested and non-nested models. Like hLRT, both information criteria use likelihood scores calculated under the assumption of the model to be tested, which are then penalized according to the open parameters these models use (Posada and Buckley 2004). The AIC is calculated as:

$$\mathrm{AIC} = -2\log_e L_i + 2K_i \qquad \textbf{(8.4)}$$

The idea behind the AIC is to test the goodness of fit (represented by the likelihood expressed as $\log_e L_i$ in this formula), by also taking into account the variance of the estimated parameters by the model (given as K_i, which represents the number of free parameters estimated by the model). The smaller the AIC, the better is the fit of the model to the data. The BIC is an easy-to-calculate approximation of the Bayes factor (Kass and Wasserman 1995) and is defined as:

$$\mathrm{BIC} = 2\log_e L_i + K_i \log_e n \tag{8.5}$$

In this formula, L_i is the likelihood given the model and the fixed topology, K_i gives the number of free parameters in the model, and n is the length of the alignment (in bp). The BIC measures the relative support of the data for any compared model. Both criteria are implemented in widely used software for the selection of nucleotide (JMODELTEST2) (Darriba et al. 2012) and protein models (PROTTEST3) (Darriba et al. 2011).

8.4.2 Partition Finding

Phylogenomic datasets usually contain hundreds or even thousands of genes (or genetic loci in general). Choosing the same model for the complete dataset is unrealistic, as differences of evolutionary rates across genes or codon positions are to be expected. However, (over-)partitioning by estimating individual models for every gene or locus can easily lead to overparametrization (Li et al. 2008). Consequently, sites or genes that evolve similarly should be merged for model selection. Such an approach is called partitioning, where the dataset is divided into homogenous blocks of sequences which evolve similarly. Consequently, a method for selecting a partition scheme for multigene datasets is needed. Yet, as in the case of possible tree topologies, the number of possible data partitions grows fast with the chosen units. For example, there are >100,000 schemes to partition a dataset of ten genes, and this number grows to more than 100 sextillion possibilities (8,47E + 23) when considering the three different codon positions (30 units) (Li et al. 2008). Lanfear et al. (2012) proposed a heuristic solution to find optimal partition schemes for large datasets, which is computationally manageable and implemented in the software PARTITIONFINDER. As for model testing, a phylogenetic tree is estimated from the data. Given this tree, the best-fit substitution models (as described above) are chosen for the defined units (called subsets, e.g. genes, codon positions). For each subset, the log likelihood (► see Sect. 8.5) is estimated. This allows estimating the likelihood of each analysed partitioning scheme by summing up the likelihood scores of the subsets which are part of this scheme. As the number of potential partitioning schemes gets astronomical even for smaller datasets, a heuristic approach using a greedy algorithm is used to limit the number of analysed schemes. Using information criteria like AIC or BIC, the optimal partitioning scheme is chosen. Nevertheless, this approach is still time intensive for large phylogenomic datasets. Consequently, a faster approach suitable for large to very large datasets has been developed based on a hierarchical clustering approach (Lanfear et al. 2014). With this approach, parameters are first estimated for initial data blocks, which are then combined based on their similarity.

8.5 Inferring Phylogenies

Four widely used methods for phylogenetic reconstruction will be introduced: neighbour joining (NJ), maximum parsimony (MP), maximum likelihood (ML) and Bayesian inference (BI). Several other methods (e.g. UPGMA, minimum evolution) have been proposed, but are basically not in use anymore in modern molecular phylogenetics. Inferring phylogenies based on molecular data can be conducted by either using pairwise distances between sequences (NJ) or based on discrete characters (MP, ML, BI) (Page and Holmes 1998). Usually, with distances, the best tree is reconstructed by clustering, whereas character-based methods apply an optimality criterion to choose the best tree(s) among all possible tree topologies (Yang and Rannala 2012). Historically, the first computer-based analyses of sequence data often relied on distances (Fitch and Margoliash 1967a). However, today, character-based methods are clearly favoured for phylogenetic reconstruction.

8.5.1 Neighbour Joining

Inferring trees by NJ consists of two steps: construction of a matrix of pairwise distances which is used for a subsequent clustering of a tree using the NJ algorithm, which chooses the tree with the smallest sum of branch lengths (Saitou and Nei 1987). Usually, distances between sequences are calculated by considering an evolutionary model (see above). This matrix is clustered into a tree by the NJ algorithm, which uses star decomposition. The algorithm starts with a completely unresolved star tree and successively joins a pair of terminals based on the distance matrix until the tree is fully resolved (Fig. 8.9). Iteratively, terminals are chosen in a way to minimize the total branch length of the tree. After every step, the distance matrix is updated newly, and the recently joined terminals are also

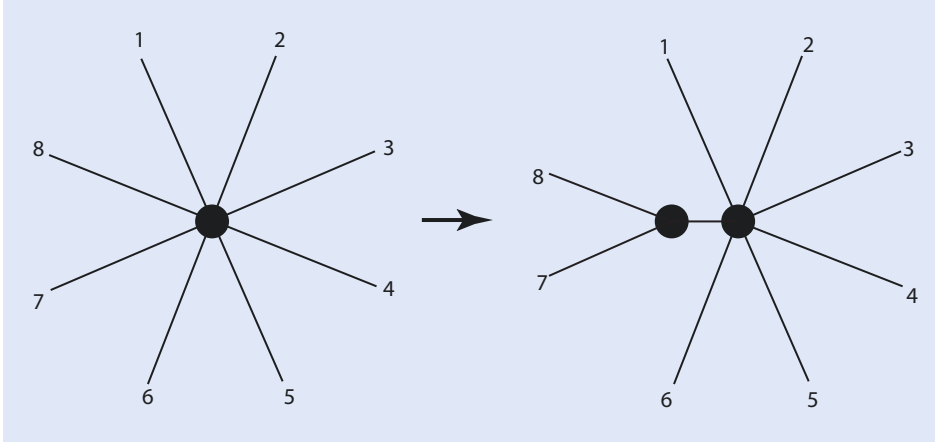


Fig. 8.9 Star decomposition as conducted by the neighbour-joining algorithm. Based on a distance matrix, the two terminals are joined which maximally reduce the total length of the tree, thereby creating a new internal node. After this step, the distance matrix is updated, and the process is repeated until the tree is completely resolved

joined in the matrix as composite terminals. A detailed description of the algorithm is given in Nei and Kumar (2000).

The NJ algorithm is, for example, implemented in the software MEGA7 (Kumar et al. 2016) or PAUP* (Swofford 2003). NJ is computationally superfast, as the time for analysing large datasets can still be measured in (mili)seconds. However, distance methods in general have been shown to be prone to problems with systematic errors and missing data (Brinkmann et al. 2005) and are therefore rarely used for phylogenomic analyses. Nevertheless, this method is often implemented when a quick tree is needed, e.g. guide trees for alignments or starting trees for heuristic searches of character-based methods (see below).

8.5.2 Maximum Parsimony

MP is a phylogenetic inference method using an optimality criterion to decide which trees are the best among all possible trees. As the number of possible trees for larger numbers of analysed sequences is too big to be analysed exhaustively, heuristic methods are used to narrow the space of searched trees (see below). The explicit rational behind MP is the idea that the best hypothesis to explain an observation is the one which requires the fewest assumptions (Steel and Penny 2000). This rational goes back to the medieval Franciscan friar William of Ockham («Ockham's razor») and is now widely used as a scientific method in general. For molecular phylogenetics, MP as method for reconstructing trees was basically introduced by Edwards and Cavalli-Sforza (1963), even though they called it minimum evolution (not to be confused with the distance-based minimum evolution method proposed by Rzhetsky and Nei (1992)!). A couple of years later, Camin and Sokal (1965) also published a parsimony-based reconstruction method, as well as Fortran-based computer programs called CLADON I to III, to carry out the steps of the analysis. Nowadays, there are several different variants of MP in use, which, for example, differ in the way if character transformations are weighted or ordered (Felsenstein 1983). In the following, the so-called Fitch parsimony is explained, where a change between any two character states is possible and all changes count equally (Fitch 1971). For MP analysis, the character states for every single alignment site (character) are mapped on a tree topology while minimizing and counting the assumed changes (steps) (Fig. 8.10). For example, in Fig. 8.10b–d, the different characters are mapped onto the same topology, and the number of transformations (steps) is counted. This MP score is measured across all possible topologies, and the trees with the lowest number of steps are chosen as the most parsimonious trees. Only characters that produce different numbers of steps across topologies are regarded as informative (e.g. Fig. 8.10e–h), whereas all other characters are excluded from the analysis. Informative characters are those which have at least two different character states, which appear at least in two terminals each. The most widely used programs for MP analyses are PAUP* (Swofford 2003) and TNT (Goloboff et al. 2008).

MP is a method which is easy to understand, and due its simplicity, efficient and fast algorithms for analysis are available (Yang and Rannala 2012). However, the lack of an explicit use of evolutionary models is a major drawback for this method. Comparisons of model-based (e.g. ML) and MP inference have been extensively discussed in the literature and especially the journals *Cladistics* and *Systematic Biology* represented a battleground for proponents of these methods in the late 1990s and early 2000s. Most simulation studies show that model-based approaches based on ML inferences (including BI)

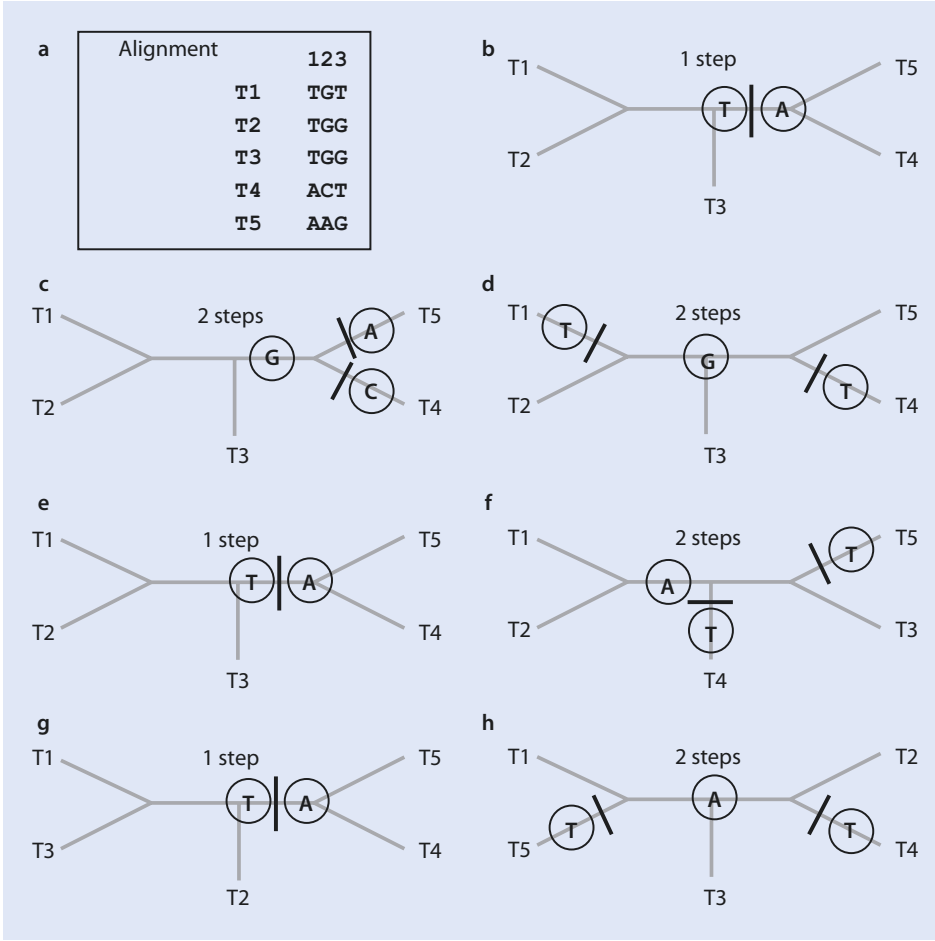


Fig. 8.10 Most parsimonious reconstructions of character change measured in steps. **a** Example alignment. **b**–**d**. Character transformations for alignment positions 1–3 reconstructed on the same unrooted tree. **e**–**h**. Reconstruction of the same alignment position on different unrooted tree topologies, illustrating that the same character can produce different numbers of steps

outperform MP in molecular phylogenetic reconstruction (Felsenstein 2013; Huelsenbeck 1995). However, MP methods are widely used for the phylogenetic reconstruction of absence/presence patterns of genome level characters, e.g. retrotransposons or microRNAs.

8.5.3 Maximum Likelihood

The likelihood function is defined as the probability of the data given the underlying parameters and was originally developed by the statistician R. A. Fisher in the 1920s. In a phylogenetic context, a tree topology represents a model, whereas the branch lengths of this topology and the underlying substitution parameters are parameters of this model (Yang and Rannala 2012). In an ML analyses, the tree topology and its set of branch lengths

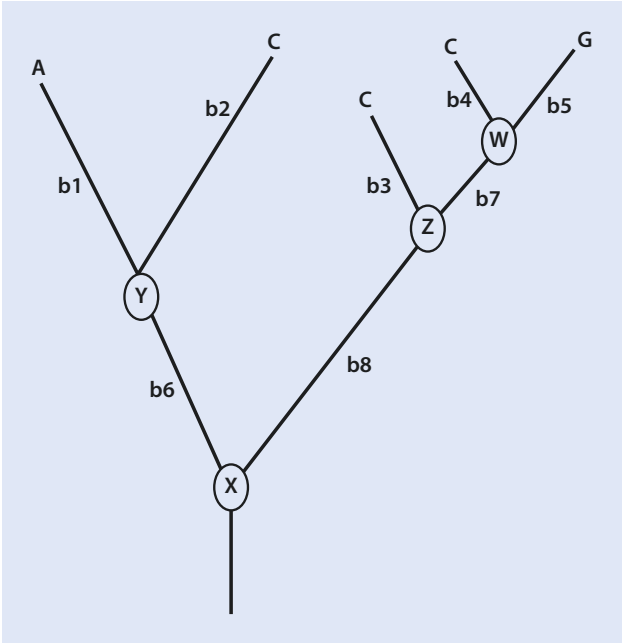


Fig. 8.11 Computing the likelihood for a single alignment site on a fixed tree using Felsenstein's pruning algorithm (see text). The letters at the tips represent character states of the terminals for this topology. Letters in circles represent ancestral nodes; b1–b7 denote branches and its corresponding lengths

are searched for, for which the data (the sequence alignment) most likely evolved as we observe it. As such, ML analyses comprise two steps. First, for a given tree topology, the lengths of individual branches as well as the parameters for an evolutionary model of sequence evolution have to be optimized. The latter part is usually conducted during the model testing procedure, as described above. Second, the most likely topology across all possible topologies has to be found using the likelihood L as an optimality criterion. The calculation of L is very time-consuming; however, Felsenstein (1981) has introduced a ML algorithm (pruning algorithm) for molecular phylogenetics, which has made ML analyses feasible. Using this approach, it is assumed that the evolution at different sites and across lineages is independent. First, the likelihood for a single site for a given topology, given branch lengths (b1–b8 in Fig. 8.11) and a chosen evolutionary model, is calculated (Fig. 8.11). The probability for a single site is the sum of the probabilities of each scenario, overall possible nucleotides that may have existed at the interior nodes (w, x, y, z in Fig. 8.11). This means, computing from the tips to the root, the probability for the presence of every possible character state for each internode has to be calculated, given the underlying substitution model. The algorithm is explained in detail in several textbooks (Felsenstein 2013; Nei and Kumar 2000; Yang 2006). The likelihood L for a given tree for the complete alignment is the product of the site-wise likelihood calculations. As these numbers are very small, usually the negative logarithm of the likelihood is used. The topology which produced the best likelihood value is finally chosen by the optimality criterion.

ML analyses are the state of the art for phylogenomics, and most publications in this field use this approach. In the early 2000s, using ML was still often computationally difficult. However, with improvements of computer technology, availability of high-performance computing clusters (HPC cluster) and especially software, which leverages this development, ML analyses became feasible for even very large datasets. At the

forefront of developing user-friendly software that can also be run on HPC clusters is Alexandros Stamatakis, the developer of the software RAXML (Stamatakis 2014). This program has been well adapted to the environment of HPC clusters, and a related software (EXAML) has been published for phylogenomic analyses on supercomputers (Kozlov et al. 2015). Both programs come with the caveat that for nucleotide analyses only the GTR model (and modifications) can be chosen. Alternative programs for large-scale ML analyses include PHYML (Guindon et al. 2010), FASTTREE (Price et al. 2010) and IQ-TREE (Nguyen et al. 2015). The latter program has also a user-friendly way of finding partitions and models for large datasets implemented.

8.5.4 Heuristic Methods and Genetic Algorithms

Computing likelihoods and optimizing branch lengths for a tree topology is time-consuming. Conducting these operations for all possible topologies to choose the tree that has the highest likelihood is basically impossible for even smaller datasets. Similarly, for MP analyses, it is impossible to investigate all possible topologies for larger datasets. For this reason, heuristic methods which only investigate a fraction of all trees, while at the same time enhancing the chance that this fraction contains the best tree, are used. Typically, a reasonable starting tree is computed to begin the heuristic search. For example, in the widely used software RAXML (Stamatakis 2014), this starting tree is inferred using a MP analysis, but it can also be based on NJ or chosen randomly. Most heuristic methods use rearrangement operations to change this starting tree and to compute new trees for phylogeny inference. Using specific rearrangement rules, different but always feasible numbers are generated based on the starting tree. The most popular heuristic tree rearrangement operations are nearest neighbour interchange (NNI), subtree pruning and regrafting (SPR) and tree-bisection and reconnection (TBR) (Felsenstein 2013). NNI swaps adjacent branches of a tree. Using SPR, a subtree of the tree is removed and regrafted into all possible positions. By TBR, a tree is split into two parts at an interior branch, and all possible connections between branches of these two trees are made. NNI will produce the smallest number of rearranged trees and TBR the largest number. After performing the possible rearrangements, the tree with the best likelihood value is chosen. Using this tree, a new round of rearrangements is performed, and the process is repeated until no better trees are found. Several modified versions of these basic operations exist. All these methods try to limit the tree space in a way that the best tree is still found. However, as there is no guarantee to find the best tree using heuristics, it is strongly recommended to conduct several replicates of the phylogenetic analysis to enhance the chance of finding the best solution.

Alternative ways for heuristic searches of ML analyses are genetic (or evolutionary) algorithms (GA). By using GA, trees represent individuals within a population, whereas the likelihood function is used as a proxy for the fitness of each individual. Fitter trees will produce more offspring trees, which are allowed to mutate over generations (e.g. by using rearrangement operations). A selection step randomly choses rearranged and unchanged trees which will be kept in the next generation. The evolution of trees is monitored over many generations, and after stopping this procedure, the tree with the highest likelihood is chosen. A GA algorithm for ML search is, for example, implemented in METAPIGA (Helaers and Milinkovitch 2010).

8.5.5 Bayesian Inference

Whereas the likelihood describes the probability of observing the data given a hypothesis (and evolutionary model), by using BI, the probability of the hypothesis given the data is described. For BI, prior probabilities and posterior probabilities have to be distinguished. Prior probabilities are assumptions made before the BI analyses. These prior probabilities are then updated according to the analysed data, and posterior probabilities are the result of BI. Using Bayes' theorem (Formula 8.6) in a phylogenetic context, the posterior probability ($f(\theta|X)$ can be calculated by multiplying the prior probability for a tree (and its parameters) ($p(\theta)$) with the likelihood of the observed data (given a tree and its parameters) ($l(X|\theta)$, and as a denominator, a normalizing constant of this product is used ($\int p(\theta) l(X|\theta) d\theta$) (Yang and Rannala 2012).

$$f(\theta \mid X) = \frac{p(\theta) l(X \mid \theta)}{\int p(\theta) l(X \mid \theta) d\theta} \tag{8.6}$$

Obviously, by using this approach in phylogenetics, different assignments of prior probabilities to tree hypotheses would have a huge impact on the posterior probabilities. To circumvent this problem, so-called flat priors are used, where all tree topologies have the same prior probability (Huelsenbeck et al. 2002b). Accordingly, all differences in the posterior probability can be attributed to differences in the likelihood function. However, there is a profound difference how both analyses use parameters of the models of sequence evolution. ML conducts a joint estimation, where the likelihood for all parameters is optimized at once. In this case, the likelihood of one parameter is dependent on the likelihood estimation of every other parameter. In contrast, BI uses a marginal estimation, where the posterior probability of any one parameter is calculated independently of any other parameter. So, even by using flat priors and identical models, ML and BI might infer different phylogenetic trees due to the differences between joint and marginal likelihood estimation (Holder and Lewis 2003).

Solving Bayes' theorem analytically is computationally too intensive. However, an approximation of posterior probabilities by using a Markov chain Monte Carlo (MCMC) approach made BI of phylogenies feasible (Larget and Simon 1999). By using a Markov chain, a series of random variables is generated, and the probability distribution of future states is only dependent on the current state at any point in the chain. For inference of phylogenies, the Markov chain starts with a randomly generated tree including branch lengths. The next step in the chain is to generate a new tree, which is based on the previous tree (e.g. using tree rearrangement heuristics or changing branch length parameters). This is called a proposal. The proposed new tree is accepted given a specific probability based on the Metropolis-Hastings algorithm (Holder and Lewis 2003). Roughly spoken, this means that it will be usually accepted if it exhibits a better likelihood and only sometimes accepted, when it has a worse likelihood. If the proposed tree is accepted, it will become the new current state to propose the next step in the chain. If the newly proposed tree is rejected, the current tree remains, and a new tree has to be proposed for the next step. Running such a Markov chain will quickly generate better trees. However, under specific conditions, there will be no better trees found, and Markov chain will have a «stationary distribution». At this point, all trees (topologies plus branch lengths) sampled are expected

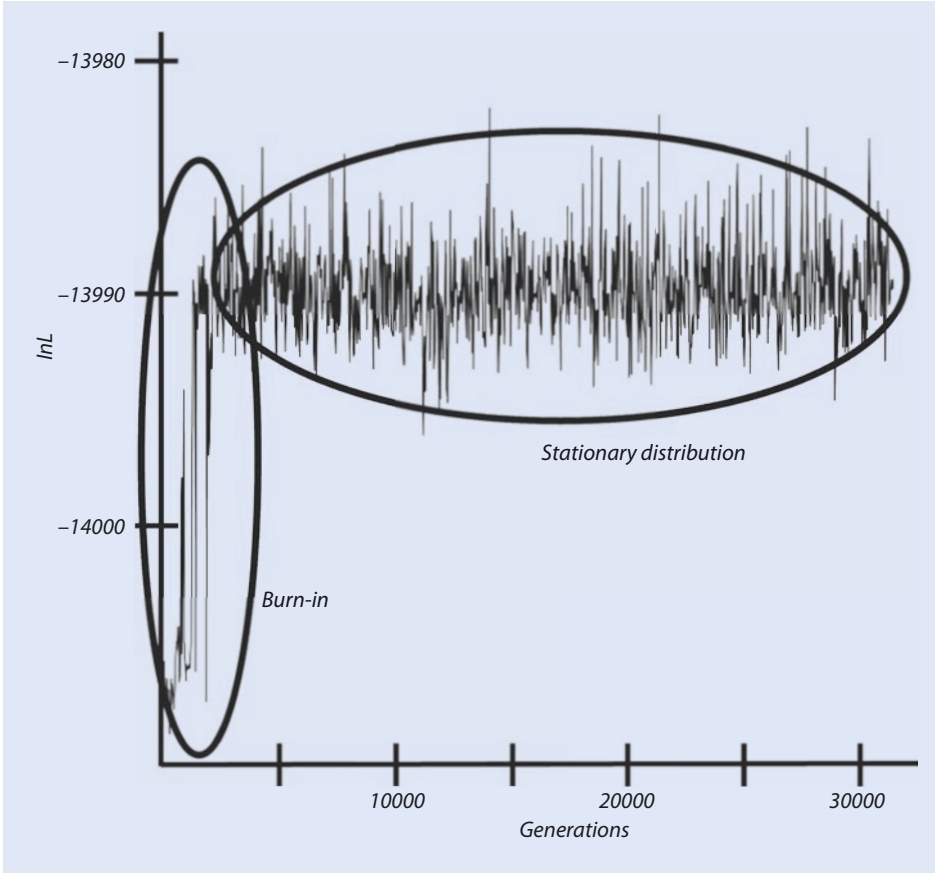


Fig. 8.12 Likelihood scores of a MCMC run plotted against generations. Once stationarity is achieved, trees from this distribution are sampled by discarding all other trees as burn-in. A majority-rule consensus of sampled trees will provide posterior probabilities for every node

to be close to the optimum, and the number of how often a tree has been visited by the chain is interpreted as an approximation of the posterior probability of this tree. By sampling a number of trees from this stationary distribution (all other sampled trees are discarded as *burn-in*) (Fig. 8.12), a majority-rule consensus tree can be generated, where the frequency of each node approximates its posterior probability (Huelsenbeck et al. 2002b). As this approach might be problematic if the chain runs into local optima, usually four Markov chains (and two independent analyses) are run in parallel (Metropolis-coupling, MCMCMC). These chains are differently explorative regarding the tree space (hot chains), and only one chain is used for sampling trees to infer the posterior probability distribution (cold chain). However, the chains are in contact and are allowed to swap their status every *n* generations (Huelsenbeck and Ronquist 2001). A widely used software for BI of phylogenies is MRBAYES (Ronquist et al. 2012). With REVBAYES, a major rewrite of this program has been published (Höhna et al. 2016). Moreover, the program PHYLOBAYES uses BI and has the site-heterogeneous CAT model of sequence evolution integrated (see above) (Lartillot et al. 2009, 2013). The program BEAST uses BI to generate ultrametric trees for molecular clock analyses (Drummond et al. 2012).

Bayesian analyses of phylogenomic datasets have been especially used for molecular clock analyses, where the program BEAST (Drummond et al. 2012) became widely popular. For standard analyses with the aim of retrieving a tree topology with support values (see below), ML seems to be the better alternative, as it is computationally usually faster, whereas the results are often similar. The biggest problem of BI is the question how long chains have to run to become stationary. Several metrics have been published to diagnose stationarity (Nylander et al. 2008), but for large datasets, this becomes difficult. Furthermore, posterior probabilities seem to overestimate the node support (Alfaro et al. 2003; Erixon et al. 2003; Simmons et al. 2004), which usually makes it necessary to either way run a ML analysis with bootstrapping (► see Sect. 8.6).

8.6 Support Measures

Phylogenetic analyses will always result into a tree topology, which raises the major question how much trust can be put into it. Usually, the most interesting is the support for certain interior branches of the tree. One measure, posterior probabilities, has been already introduced in ► Sect. 8.5. The most common measure of support for phylogenies is derived from bootstrap analyses. The bootstrap is a resampling technique commonly used in statistics for estimating the variability of an estimate (Efron 1982). The application of bootstrapping for phylogenetics was introduced by Felsenstein (1985). To conduct bootstrap analyses the original dataset has to be resampled with replacement. As such, so-called pseudoreplicates are generated which contain the same number of alignment sites as the original alignment. Every site in these pseudoreplicates is filled by sites from the original alignment. As this sampling is conducted with replacement, the pseudoreplicates may include some original sites multiple times, where others could be missing. Normally 100 or 1000 pseudoreplicates are generated, which are then analysed as in the original phylogenetic analysis (e.g., with ML, MP or NJ). Alternatively, a «bootstopping» algorithm can estimate the number of necessary replicates (Pattengale et al. 2009). All trees resulting from these analyses are then summarized as a majority-rule consensus tree and the frequencies are given at the nodes. If a branch is found in all replicates, the support is 100%. In statistics, these values are interpreted in the typical fashion that values >95% are statistically significant. This support describes how well a branch is supported by the data, not the probability if a branch is «true». This also implies that the bootstrap basically tests if the dataset is large enough to recover a well-supported solution. Earlier studies dealing with single gene datasets claimed that the bootstrap might be over-conservative, and values >70% can be regarded as significant (Hillis and Bull 1993). However, large datasets as used in phylogenomics seem to inflate highly supported branches, and a 95% threshold of support seems reasonable here. Nevertheless, it should be kept in mind that especially systematic error within the data can lead to significantly supported branches which are wrong. Bootstrap analyses are computationally time intensive, and several approaches which are able to quickly approximate bootstrap values for large datasets have been published, e.g. implemented in IQ-TREE (Minh et al. 2013) and RAXML (Stamatakis et al. 2008). A related resampling method that has been used in phylogenetics is the jackknife, where instead of resampling with replacement, randomly half of the positions are deleted in the pseudoreplicates (Felsenstein 2013).

An alternative way of estimating branch support is based on likelihood ratio tests (► see also Sect. 8.4). In the case of the approximate LRT (aLRT), the idea is based on

comparing internal branches of an inferred tree to the null hypothesis, where the length of this branch is zero (Anisimova and Gascuel 2006). However, for testing purposes, the null hypothesis of a zero branch length is approximated by testing against the putatively incorrect branching. For this, the best topology is compared with the best alternative arrangement around the branch of interest. For any given internal branch, only three topological arrangements are possible in its neighbourhood, which can be easily ordered by their likelihood. The LRT test statistic is calculated as two times the difference in likelihood between the best tree (*L1*) and the best alternative hypothesis (*L2*). The result is compared against a mixed χ^2-distribution. Simulation studies show that the aLRT is much faster and similarly accurate as standard bootstrap approaches, as long as the underlying evolutionary model of the phylogenetic analysis is not strongly violated (Anisimova et al. 2011). A possibility to get a more robust version of the LRT under the presence of model misspecifications is the inclusion of a bootstrapping step in this test. This has been done for the SH-aLRT (Guindon et al. 2010), where a variant of the bootstrap called RELL is used (Kishino et al. 1990). RELL (resampling estimated log likelihoods) is a shortcut to calculate likelihood values for pseudoreplicates. Instead of generating pseudoreplicates of the alignments, site-wise likelihoods of the best tree of the original alignment are bootstrapped. This fast (but maybe inaccurate) method helps to generate a distribution of likelihoods for a large number of pseudoreplicates, without having to perform the time-consuming ML optimization step. The SH-aLRT compares the distribution of the RELL-bootstrapped topologies with a test statistic developed by Shimodaira and Hasegawa (1999). Simulation studies have shown that the SH-aLRT is much more robust for datasets analysed under model violations than the aLRT (Anisimova et al. 2011). LRT for branch support is, for example, implemented in PHYML (Guindon et al. 2010) and IQ-TREE (Nguyen et al. 2015).

8.7 Molecular Clocks

According to the molecular clock hypothesis, which assumes a constant rate of evolution over time, it is possible to date divergence times in phylogenetic trees using the fossil record (Hasegawa et al. 1985). The existence of a molecular evolutionary clock was first hypothesized by Zuckerkandl and Pauling (1965), based on the results of their landmark study which proposed the existence of a uniform rate of evolution among globin genes in different species (Zuckerkandl and Pauling 1962). This result was in line with the finding of Doolittle and Blomback (1964), who found an inverse relationship of species divergence time and difference in protein sequences. However, with the availability of more DNA sequence data, it became obvious that mutation rates can be different across taxa and genes, thereby implying that a strict molecular clock hypothesis is an unrealistic assumption (Kumar 2005). Several methods have been developed to deal with this problem. Sarich and Wilson (1973) and Fitch (1976) proposed a relative-rate test, where the rate of evolution of two (ingroup) sequences is independently compared to an outgroup sequence. By this procedure, it is possible to test if the distance between the two ingroup sequences to its last common ancestor is equal (or not significantly different), as assumed under the molecular clock hypothesis. If a χ^2-test indicates a significant difference in this distance, it is rejected that this pair of sequences evolves according to a molecular clock. With the help of such a test, gene alignments (and included sequences) can be filtered, and only those who fulfil the molecular clock criterion are used for analysis. Relative-rate tests

demonstrated that the assumption of a global molecular clock is unrealistic for most datasets, thereby prohibiting molecular clock analyses for them. However, local molecular clock analyses within a maximum likelihood framework are possible, where some lineages evolve under different evolutionary rates, while other lineages have a constant rate (Yoder and Yang 2000). Sanderson (1997) developed a method (nonparametric rate smoothing), which is based on the assumption that evolutionary rates show autocorrelation over time. This idea goes back to Gillespie (1991), who suggested that substitution rates evolve among lineages and are inherited from ancestors to descendants. Under this assumption, a model can be used which minimizes the change of evolutionary rate between related (ancestor-descendant) lineages while allowing variation across lineages. Nowadays, most widely used are Bayesian approaches which allow the use and comparison of alternative models of substitution changes over time and for different data partitions (Lepage et al. 2007), as, for example, implemented in the software BEAST (Drummond et al. 2012).

An obviously important step for every molecular clock analysis is the calibration of the resulting ultrametric tree. This is usually done by using fossil data, but also biogeographic events can be helpful. Till the end of the 1990s, it was a commonplace to use a single calibration point for molecular clock analyses. Often, a single gene was analysed with the help of one dated internal node, where the substitution rate of the dated lineage was divided by the age of the dated divergence to subsequently transform all genetic distances into absolute time (Renner 2005). Later on, it became standard to use multiple calibration points (if available!), which could be used to cross validate each other (Benton et al. 2009). Moreover, it is possible to assign minimum and maximum ages for any used calibration point. Minimum ages are hard bound, indicating that a certain clade must have at least this age as evidenced by the first appearance in the fossil record. In contrast, maximum ages are more difficult to assign and are thereby soft bound, given as a distribution, which tries to estimate the origin of a species which is always certainly older than its first appearance in the fossil record (Donoghue and Benton 2007). A best practice guide for the justification of fossil calibration has been published by Parham et al. (2012).

The potential and pitfalls of molecular clock analyses are nicely illustrated by several studies dealing with the origin of animals. It always has been a conundrum that animal fossils are either rare or disputed (e.g. the Ediacaran fauna) in the Precambrian (>541 mya) fossil record, whereas basically all major phyla are suddenly found in different Cambrian (541–485 mya) Lagerstaetten (Briggs 2015). This conundrum is known as the «Cambrian explosion». Molecular clocks represent an interesting approach to investigate the timing of animal evolution, and many publications dealing with this topic have been published in the last decades. However, instead of converging to a similar conclusion, most of these studies differ wildly in their results. As such, dates for the emergence of animals range from ~600 mya (Peterson et al. 2004) to ~1300 mya (Hedges et al. 2004). Moreover, often these dates come with a huge error rate, making precise statements difficult (Graur and Martin 2004). These errors are often introduced due to the problems of assigning well-supported calibration dates for such old fossils, questioning the possibility of using molecular clocks for rejecting or supporting hypothesis of early animal evolution in general (dos Reis et al. 2015). However, many examples of dating younger divergences clearly emphasize the power of molecular clock analyses, which have been used for less controversial divergence time estimates of the evolution of, e.g. insects, mammals, humans or plants (dos Reis et al. 2016; Renner 2005). Moreover, molecular clock analyses have been successfully used to analyse virus outbreaks, as in the case of Ebola, HIV or influenza (dos Reis et al. 2016).

References

Abascal F, Posada D, Zardoya R (2007) MtArt: a new model of amino acid replacement for arthropoda. Mol Biol Evol 24:1–5

Adachi J, Hasegawa M (1996) Model of amino acid substitution in proteins encoded by mitochondrial DNA. J Mol Evol 42:459–468

Adachi J, Waddell PJ, Martin W, Hasegawa M (2000) Plastid genome phylogeny and a model of amino acid substitution for proteins encoded by chloroplast DNA. J Mol Evol 50:348–358

Akaike H (1973) Information theory and an extension of the maximum likelihood principle. In: Proceedings of the 2nd international symposium on Information Theory. Budapest, p 267–281

Alfaro ME, Zoller S, Lutzoni F (2003) Bayes or bootstrap? A simulation study comparing the performance of Bayesian Markov chain Monte Carlo sampling and bootstrapping in assessing phylogenetic confidence. Mol Biol Evol 20:255–266

Anisimova M, Gascuel O (2006) Approximate likelihood-ratio test for branches: a fast, accurate, and powerful alternative. Syst Biol 55:539–552

Anisimova M, Gil M, Dufayard J-F, Dessimoz C, Gascuel O (2011) Survey of branch support methods demonstrates accuracy, power, and robustness of fast likelihood-based approximation schemes. Syst Biol 60:685–699

Anisimova M, Liberles DA, Philippe H, Provan J, Pupko T, von Haeseler A (2013) State-of the art methodologies dictate new standards for phylogenetic analysis. BMC Evol Biol 13:161

Antoniak C (1974) Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. Ann Stat 2:1152–1174

Barry D, Hartigan JA (1987) Statistical analysis of hominoid molecular evolution. Stat Sci 2:191–207

Baum BR (1992) Combining trees as a way of combining data sets for phylogenetic inference, and the desirability of combining gene trees. Taxon 41:3–10

Benton MJ, Donoghue PCJ, Asher R (2009) Calibrating and constraining molecular clocks. In: Hedges SB, Kumar S (eds) The timetree of life. Oxford University Press, Oxford, pp 35–86

Bininda-Emonds ORP (2004) The evolution of supertrees. Trends Ecol Evol 19:315–322

Blanquart S, Lartillot N (2006) A Bayesian compound stochastic process for modeling nonstationary and nonhomogeneous sequence evolution. Mol Biol Evol 23:2058–2071

Blanquart S, Lartillot N (2008) A site- and time-heterogeneous model of amino acid replacement. Mol Biol Evol 25:842–858

Briggs DEG (2015) The cambrian explosion. Curr Biol 25:R864–R868

Brinkmann H, van der Giezen M, Zhou Y, de Raucourt GP, Philippe H (2005) An empirical assessment of long-branch attraction artefacts in deep eukaryotic phylogenomics. Syst Biol 54:743–757

Camin JH, Sokal RR (1965) A method for deducing branching sequences in phylogeny. Evolution 19:311–326

Darriba D, Taboada GL, Doallo R, Posada D (2011) ProtTest 3: fast selection of best-fit models of protein evolution. Bioinformatics 27:1164–1165

Darriba D, Taboada GL, Doallo R, Posada D (2012) jModelTest 2: more models, new heuristics and parallel computing. Nat Methods 9:772–772

Dayhoff M, Eck R, Park C (1972) A model of evolutionary change in proteins. In: Dayhoff M (ed) Atlas of protein sequence and structure, vol 5. National Biomedical Research Foundation, Washington, DC, pp 89–99

Dayhoff M, Schwarz R, Orcutt B (1978) A model of evolutionary change in proteins. In: Dayhoff M (ed) Atlas of protein sequence and structure, vol 5, Suppl. 3. National Biomedical Research Foundation, Washington, DC, pp 345–352

Dimmic MW, Rest JS, Mindell DP, Goldstein RA (2002) rtREV: an amino acid substitution matrix for inference of retrovirus and reverse transcriptase phylogeny. J Mol Evol 55:65–73

Donoghue PCJ, Benton MJ (2007) Rocks and clocks: calibrating the tree of life using fossils and molecules. Trends Ecol Evol 22:424–431

Doolittle WF, Bapteste E (2007) Pattern pluralism and the tree of life hypothesis. Proc Natl Acad Sci U S A 104:2043–2049

Doolittle RF, Blomback B (1964) Amino-acid sequence investigations of fibrinopeptides from various mammals: evolutionary implications. Nature 202:147–152

dos Reis M, Thawornwattana Y, Angelis K, Telford Maximilian J, Donoghue Philip CJ, Yang Z (2015) Uncertainty in the timing of origin of animals and the limits of precision in molecular timescales. Curr Biol 25:2939–2950

dos Reis M, Donoghue PCJ, Yang Z (2016) Bayesian molecular clock dating of species divergences in the genomics era. Nat Rev Genet 17:71–80
Drummond AJ, Suchard MA, Xie D, Rambaut A (2012) Bayesian phylogenetics with BEAUti and the BEAST 1.7. Mol Biol Evol 29:1969–1973
Edwards AWF, Cavalli-Sforza LL (1963) The reconstruction of evolution. Heredity 18:553
Efron B (1982) The jackknife, the bootstrap and other resampling plans. CBMS-NSF regional conference series in applied mathematics. Society for Industrial and Applied Mathematics, Philadelphia
Erixon P, Svennblad B, Britton T, Oxelman B (2003) Reliability of Bayesian posterior probabilities and bootstrap frequencies in phylogenetics. Syst Biol 52:665–673
Felsenstein J (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. J Mol Evol 17:368–376
Felsenstein J (1983) Parsimony in systematics: biological and statistical issues. Annu Rev Ecol Evol Syst 14:313–333
Felsenstein J (1985) Confidence limits on phylogenies – an approach using the bootstrap. Evolution 39:783–791
Felsenstein J (2013) Inferring phylogenies. Sinauer Associates, Sunderland
Fitch WM (1971) Toward defining the course of evolution: minimum change for a specific tree topology. Syst Zool 20:406–416
Fitch WM (1976) Molecular evolutionary clocks. In: Ayala FJ (ed) Molecular evolution. Sinauer Associates, Sunderland, pp 160–178
Fitch WM, Margoliash E (1967a) Construction of phylogenetic trees: a method based on mutation distances as estimated from cytochrome c sequences is of general applicability. Science 155:279–284
Fitch WM, Margoliash E (1967b) A method for estimating the number of invariant amino acid coding positions in a gene using cytochrome c as a model case. Biochem Genet 1:65–71
Fourment M, Gibbs MJ (2006) PATRISTIC: a program for calculating patristic distances and graphically comparing the components of genetic change. BMC Evol Biol 6:1
Fryxell KJ, Moon W-J (2005) CpG mutation rates in the human genome are highly dependent on local GC content. Mol Biol Evol 22:650–658
Gillespie J (1991) The causes of molecular evolution. Oxford University Press, New York
Goldman N, Yang Z (1994) A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol Biol Evol 11:725–736
Goloboff PA, Farris JS, Nixon KC (2008) TNT, a free program for phylogenetic analysis. Cladistics 24:774–786
Graur D, Martin W (2004) Reading the entrails of chickens: molecular timescales of evolution and the illusion of precision. Trends Genet 20:80–86
Gu X, Fu YX, Li WH (1995) Maximum likelihood estimation of the heterogeneity of substitution rate among nucleotide sites. Mol Biol Evol 12:546–557
Guindon S, Dufayard J-F, Lefort V, Anisimova M, Hordijk W, Gascuel O (2010) New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol 59:307–321
Hasegawa M, Kishino H, T-a Y (1985) Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J Mol Evol 22:160–174
Heads M (2005) Dating nodes on molecular phylogenies: a critique of molecular biogeography. Cladistics 21:62–78
Hedges SB, Blair JE, Venturi ML, Shoe JL (2004) A molecular timescale of eukaryote evolution and the rise of complex multicellular life. BMC Evol Biol 4:2
Helaers R, Milinkovitch MC (2010) MetaPIGA v2.0: maximum likelihood large phylogeny estimation using the metapopulation genetic algorithm and other stochastic heuristics. BMC Bioinformatics 11:379
Hess PN, De Moraes Russo CA (2007) An empirical test of the midpoint rooting method. Biol J Linn Soc 92:669–674
Hillis DM, Bull JJ (1993) An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. Syst Biol 42:182–192
Hodgkinson A, Eyre-Walker A (2011) Variation in the mutation rate across mammalian genomes. Nat Rev Genet 12:756–766
Höhna S, Landis MJ, Heath TA, Boussau B, Lartillot N, Moore BR, Huelsenbeck JP, Ronquist F (2016) RevBayes: Bayesian phylogenetic inference using graphical models and an interactive model-specification language. Syst Biol 65:726–736

Holder M, Lewis PO (2003) Phylogeny estimation: traditional and Bayesian approaches. Nat Rev Genet 4:275–284
Huelsenbeck JP (1995) Performance of phylogenetic methods in simulation. Syst Biol 44:17–48
Huelsenbeck JP, Ronquist F (2001) MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics 17:754–755
Huelsenbeck JP, Bollback JP, Levine AM (2002a) Inferring the root of a phylogenetic tree. Syst Biol 51:32–43
Huelsenbeck JP, Larget B, Miller RE, Ronquist F (2002b) Potential applications and pitfalls of Bayesian inference of phylogeny. Syst Biol 51:673–688
Huerta-Cepas J, Dopazo J, Gabaldón T (2010) ETE: a python environment for tree exploration. BMC Bioinformatics 11:24
Huson DH, Bryant D (2006) Application of phylogenetic networks in evolutionary studies. Mol Biol Evol 23:254–267
Huson DH, Richter DC, Rausch C, Dezulian T, Franz M, Rupp R (2007) Dendroscope: an interactive viewer for large phylogenetic trees. BMC Bioinformatics 8:460
Huson DH, Rupp R, Scornavacca C (2010) Phylogenetic networks. Concepts, algorithms and applications. Cambridge University Press, Cambridge
Jayaswal V, Jermiin LS, Poladian L, Robinson J (2011) Two stationary nonhomogeneous markov models of nucleotide sequence evolution. Syst Biol 60:74–86
Jia F, Lo N, Ho SYW (2014) The impact of modelling rate heterogeneity among sites on phylogenetic estimates of intraspecific evolutionary rates and timescales. PLoS One 9:e95722
Jones DT, Taylor WR, Thornton JM (1992) The rapid generation of mutation data matrices from protein sequences. Comput Appl Biosci CABIOS 8:275–282
Jukes T, Cantor C (1969) Evolution of protein molecules. In: Munro R (ed) Mammalian protein metabolism. Academic Press, New York, pp 21–132
Kass RE, Wasserman L (1995) A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. J Am Stat Assoc 90:928–934
Kimura M (1980) A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol 16:111–120
Kishino H, Miyata T, Hasegawa M (1990) Maximum likelihood inference of protein phylogeny and the origin of chloroplasts. J Mol Evol 31:151–160
Kosiol C, Holmes I, Goldman N (2007) An empirical codon model for protein sequence evolution. Mol Biol Evol 24:1464–1479
Kozlov AM, Aberer AJ, Stamatakis A (2015) ExaML version 3: a tool for phylogenomic analyses on supercomputers. Bioinformatics 31:2577–2579
Krell F-T, Cranston PS (2004) Which side of the tree is more basal? Syst Entomol 29:279–281
Kück P, Mayer C, Wägele J-W, Misof B (2012) Long branch effects distort maximum likelihood phylogenies in simulations despite selection of the correct model. PLoS One 7:e36593
Kumar S (2005) Molecular clocks: four decades of evolution. Nat Rev Genet 6:654–662
Kumar S, Stecher G, Tamura K (2016) MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. Mol Biol Evol 33:1870–1874
Lanfear R, Calcott B, Ho SYW, Guindon S (2012) PartitionFinder: combined selection of partitioning schemes and substitution models for phylogenetic analyses. Mol Biol Evol 29:1695–1701
Lanfear R, Calcott B, Kainer D, Mayer C, Stamatakis A (2014) Selecting optimal partitioning schemes for phylogenomic datasets. BMC Evol Biol 14:82
Larget B, Simon D (1999) Markov chain Monte Carlo algorithms for the Bayesian analysis of phylogenetic trees. Mol Biol Evol 16:750–759
Lartillot N, Philippe H (2004) A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. Mol Biol Evol 21:1095–1109
Lartillot N, Lepage T, Blanquart S (2009) PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. Bioinformatics 25:2286–2288
Lartillot N, Rodrigue N, Stubbs D, Richer J (2013) PhyloBayes MPI: phylogenetic reconstruction with infinite mixtures of profiles in a parallel environment. Syst Biol 62:611–615
Le SQ, Gascuel O (2008) An improved general amino acid replacement matrix. Mol Biol Evol 25:1307–1320
Le SQ, Lartillot N, Gascuel O (2008) Phylogenetic mixture models for proteins. Philos Trans R Soc Lond Ser B Biol Sci 363:3965–3976

Le SQ, Dang CC, Gascuel O (2012) Modeling protein evolution with several amino acid replacement matrices depending on site rates. Mol Biol Evol 29:2921–2936
Lepage T, Bryant D, Philippe H, Lartillot N (2007) A general comparison of relaxed molecular clock models. Mol Biol Evol 24:2669–2680
Letunic I, Bork P (2016) Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. Nucleic Acids Res 44:W242–W245
Li C, Lu G, Ortí G (2008) Optimal data partitioning and a test case for ray-finned fishes (Actinopterygii) based on ten nuclear loci. Syst Biol 57:519–539
Minh BQ, Nguyen MAT, von Haeseler A (2013) Ultrafast approximation for phylogenetic bootstrap. Mol Biol Evol 30:1188–1195
Miyazawa S (2013) Superiority of a mechanistic codon substitution model even for protein sequences in phylogenetic analysis. BMC Evol Biol 13:257
Nei M, Kumar S (2000) Molecular evolution and phylogenetics. Oxford University Press, New York
Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ (2015) IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol 32:268–274
Nixon KC, Carpenter JM (1993) On outgroups. Cladistics 9:413–426
Nylander JAA, Wilgenbusch JC, Warren DL, Swofford DL (2008) AWTY (are we there yet?): a system for graphical exploration of MCMC convergence in Bayesian phylogenetics. Bioinformatics 24:581–583
Page RDM (1996) Tree view: an application to display phylogenetic trees on personal computers. Compu Appli Biosci CABIOS 12:357–358
Page RD, Holmes E (1998) Molecular evolution: a phylogenetic approach. Blackwell, Osney Mead/Oxford
Parham JF, Donoghue PCJ, Bell CJ, Calway TD, Head JJ, Holroyd PA, Inoue JG, Irmis RB, Joyce WG, Ksepka DT, Patané JSL, Smith ND, Tarver JE, van Tuinen M, Yang Z, Angielczyk KD, Greenwood JM, Hipsley CA, Jacobs L, Makovicky PJ, Müller J, Smith KT, Theodor JM, Warnock RCM (2012) Best practices for justifying fossil calibrations. Syst Biol 61(2):346–359
Pattengale ND, Alipour M, Bininda-Emonds ORP, Moret BME, Stamatakis A (2009) How many bootstrap replicates are necessary? In: Batzoglou S (ed) RECOMB 2009, LNCS 5541. Springer, Berlin/Heidelberg, pp 184–200
Peterson KJ, Lyons JB, Nowak KS, Takacs CM, Wargo MJ, McPeek MA (2004) Estimating metazoan divergence times with a molecular clock. Proc Natl Acad Sci U S A 101:6536–6541
Posada D, Buckley TR (2004) Model selection and model averaging in phylogenetics: advantages of akaike information criterion and bayesian approaches over likelihood ratio tests. Syst Biol 53:793–808
Posada D, Crandall KA (1998) MODELTEST: testing the model of DNA substitution. Bioinformatics 14:817–818
Posada D, Crandall KA (2001) Intraspecific gene genealogies: trees grafting into networks. Trends Ecol Evol 16:37–45
Price MN, Dehal PS, Arkin AP (2010) FastTree 2 – Approximately maximum-likelihood trees for large alignments. PLoS One 5:e9490
Ragan MA (1992) Phylogenetic inference based on matrix representation of trees. Mol Phylogenet Evol 1:53–58
Ren F, Tanaka H, Yang Z (2005) An empirical examination of the utility of codon-substitution models in phylogeny reconstruction. Syst Biol 54:808–818
Renner SS (2005) Relaxed molecular clocks for dating historical plant dispersal events. Trends Plant Sci 10:550–558
Rodríguez F, Oliver JL, Marín A, Medina JR (1990) The general stochastic model of nucleotide substitution. J Theor Biol 142:485–501
Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A, Höhna S, Larget B, Liu L, Suchard MA, Huelsenbeck JP (2012) MrBayes 3.2: efficient bayesian phylogenetic inference and model choice across a large model space. Syst Biol 61:539–542
Rzhetsky A, Nei M (1992) A simple method for estimating and testing minimum-evolution trees. Mol Biol Evol 9:945
Saitou N, Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 4:406–425
Sanderson M (1997) A nonparametric approach to estimating divergence times in the absence of rate constancy. Mol Biol Evol 14:1218
Sarich VM, Wilson AC (1973) Generation time and genomic evolution in primates. Science 179:1144–1147

Schneider A, Cannarozzi GM, Gonnet GH (2005) Empirical codon substitution matrix. BMC Bioinformatics 6:134
Shimodaira H, Hasegawa M (1999) Multiple comparisons of log-likelihoods with applications to phylogenetic inference. Mol Biol Evol 16:1114
Simmons MP, Pickett KM, Miya M (2004) How meaningful are Bayesian support values? Mol Biol Evol 21:188–199
Stamatakis A (2006) Phylogenetic models of rate heterogeneity: a high performance computing perspective. In: Proceedings of the 20th IEEE international parallel & distributed processing symposium (IPDPS2006). IEEE Computer Society Press, Washington, pp 278–286
Stamatakis A (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30:1312–1313
Stamatakis A, Hoover P, Rougemont J (2008) A rapid bootstrap algorithm for the RAxML web servers. Syst Biol 57:758–771
Steel M, Penny D (2000) Parsimony, likelihood, and the role of models in molecular phylogenetics. Mol Biol Evol 17:839–850
Sullivan J, Joyce P (2005) Model selection in phylogenetics. Annu Rev Ecol Syst 36:445–466
Sullivan J, Swofford D, Naylor G (1999) The effect of taxon sampling on estimating rate heterogeneity parameters of maximum-likelihood models. Mol Biol Evol 16:1347
Swofford D (2003) PAUP*: phylogenetic analysis using parsimony (and other methods). Sinauer Associates, Sunderland
Tavare S (1986) Some probabilistic and statistical problems in the analysis of DNA sequences. Lectures Math Life Sci (Amer Math Soc) 17:57–86
Whelan S, Goldman N (2001) A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Mol Biol Evol 18:691–699
Wilkinson M (1994) Common cladistic information and its consensus representation: reduced Adams and reduced cladistic consensus trees and profiles. Syst Biol 43:343–368
Yang Z (1994) Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J Mol Evol 39:306–314
Yang Z (1996) Among-site rate variation and its impact on phylogenetic analyses. Trends Ecol Evol 11:367–372
Yang Z (2006) Computational molecular evolution. Oxford series in ecology and evolution. Oxford University Press, Oxford
Yang Z (2007) PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol 24:1586–1591
Yang Z, Bielawski JP (2000) Statistical methods for detecting molecular adaptation. Trends Ecol Evol 15:496–503
Yang Z, Rannala B (2012) Molecular phylogenetics: principles and practice. Nat Rev Genet 13:303–314
Yang Z, Nielsen R, Hasegawa M (1998) Models of amino acid substitution and applications to mitochondrial protein evolution. Mol Biol Evol 15:1600–1611
Yoder AD, Yang Z (2000) Estimation of primate speciation dates using local molecular clocks. Mol Biol Evol 17:1081–1090
Zaheri M, Dib L, Salamin N (2014) A generalized mechanistic codon model. Mol Biol Evol 31:2528–2541
Zuckerkandl E, Pauling L (1962) Molecular disease, evolution and genetic heterogeneity. In: Kasaha M, Pullman B (eds) Horizons in biochemistry. Academic Press, New Yoek, pp 189–225
Zuckerkandl E, Pauling L (1965) Evolutionary divergence and convergence, in proteins. In: Bryson V, Vogel H (eds) Evolving genes and proteins. Academic Press, New Yoork, pp 441–465

Sources of Error and Incongruence in Phylogenomic Analyses

C. Bleidorn, *Phylogenomics*, DOI 10.1007/978-3-319-54064-1_9

- Phylogenomic analyses can be performed by analysing gene trees separately and using coalescent or supertree analyses or a concatenation of all genes (supermatrix approach).
- Several sources of systematic error may bias phylogenomic studies due to the violation of substitution model assumptions, including problems with compositional heterogeneity, among-lineage rate variation and heterotachy.
- Missing data is usually less problematic for phylogenomic studies, but taxon sampling can be critical.
- Data and taxa should be carefully selected for analysis; highly saturated genes as well as phylogenetically unstable (rogue) taxa should be avoided.
- Discordance of gene trees and species trees is not rare, and potential causes are incongruent lineage sorting, hybridization or horizontal gene transfer.
- Coalescent-based methods are able to reconstruct species tree inference when gene trees are incongruent due to incomplete lineage sorting.

9.1 Incongruence in Phylogenomic Analyses

During the end of the 1990s and the early 2000s, molecular phylogenetic analyses revolutionized phylogenetic systematics. Many results contributed to changing textbook knowledge about the evolutionary relationships of plant and animal systematics and enabled a new picture for the phylogeny of the entire group of eukaryotes (Donoghue and Doyle 2000; Halanych 2004; Adl et al. 2005). Many of these early analyses were based on a single or few genes, leaving many nodes – especially deep in time – unsupported or unresolved. Current practice of phylogenomic analyses can be broadly classified into two different approaches: supermatrix and gene tree-based analyses of hundreds or thousands of genes (Liu et al. 2015). In the case of supermatrix analyses, all gene alignments are concatenated into a single matrix, which is subsequently analysed using the chosen phylogenetic method. In the case of gene tree-based analyses, all genes are analysed separately, and in a second step, the resulting topologies are (subsequently or simultaneously) used to construct a supertree (Bininda-Emonds 2004) or a species tree based on coalescent theory (► see Sect. 9.4). Phylogenomic approaches are able to produce precise estimations of phylogeny; however, this does not mean the result reflects the true evolutionary history (Kumar et al. 2012), as several factors can mislead phylogenetic analyses even when a massive amount of data is available.

The era of phylogenomic analyses to resolve relationships among organisms was basically kick-started in 2003. By analysing 106 different genes to resolve the phylogeny of yeast, Rokas et al. (2003) found incongruence among them, sometimes strongly supporting competing hypotheses (Fig. 9.1). Using a genome-scale approach, the incongruence disappeared when combining all of them. Moreover, it was shown that a concatenation of any 20 out of these 106 genes always recovered the best topology with bootstrap values of at least 95% for each node. Even though details of this study have been criticized to be unrealistic (Gatesy et al. 2007), it clearly supported the idea that phylogenomic approaches could end incongruence in phylogenetics (Gee 2003). Whereas genome-scale approaches for most groups of non-model organisms remained a pipe dream in 2003, the availability of next-generation sequencing (NGS) techniques allowed gathering huge datasets for basically every taxon of interest (Rokas and Abbot 2009).

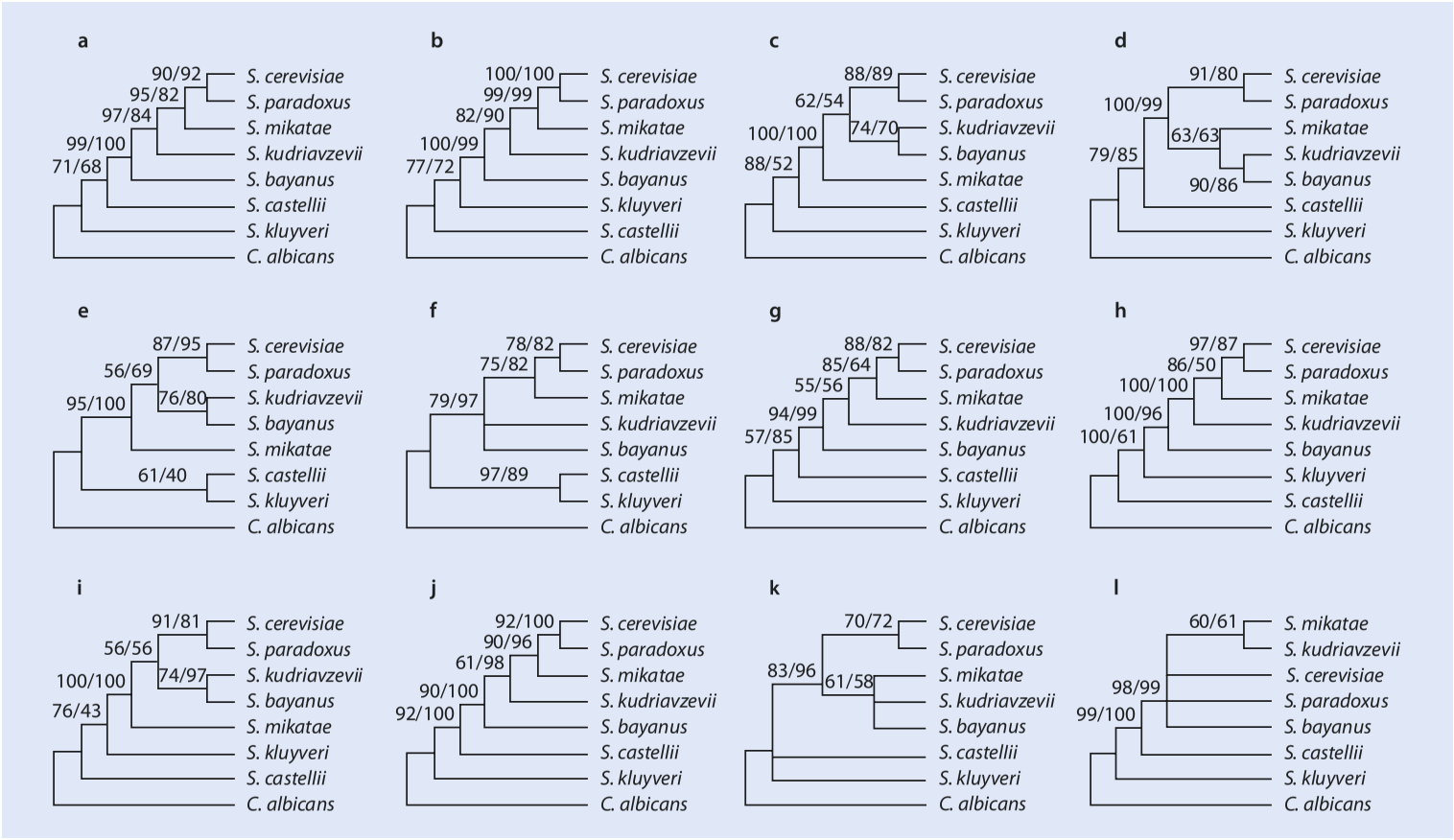


Fig. 9.1 **a–l** Incongruence among gene trees from a phylogenomic analysis of yeast relationships (Reprinted by permission from Macmillan Publishers Ltd.: [*Nature*] Rokas et al. (2003), copyright 2003)

There are several reasons why trees inferred from single genes (i.e. gene trees) might differ with each other (Jeffroy et al. 2006). First, this might be a stochastic error associated with a lack of sufficient phylogenetic signal, which could be overcome by combining more (informative) genes. This approach assumes that combining more genes into a single data matrix should increase the phylogenetic signal-to-noise ratio compared to single genes (de Queiroz and Gatesy 2007). Second, the species tree will be different from a gene tree because of violation of the orthology assumption, incongruent lineage sorting or horizontal gene transfer. There are certain methods detecting such problems and dealing with them in phylogenomic datasets (▶ see Sect. 9.4). Third, systematic errors present in single genes might also lead to artefacts in the phylogenetic reconstruction (▶ see Sect. 9.2). Such systematic errors are usually due to the violation of assumptions of the underlying model for the analyses. Systematic errors can occur because the assumptions of the underlying model are violated, including (I) heterogeneity of the nucleotide/amino acid composition among lineages (compositional signal), (II) variation of the substitution rate among lineages (rate signal) and (III) variation in the substitution rate within nucleotide positions over time (heterotacheous signal). All these patterns are generally not accounted for by the evolutionary model and might negatively impact phylogenetic reconstruction.

Often, high statistical support (e.g. bootstrapping) is taken as a measure that the tree is correct. However, it is important to remember that these measures assess the stability of the obtained relationships to sampling error (White et al. 2007). Bootstrap analyses detect if datasets contain a pattern and how strong this is but are not able to decide whether or not this pattern represents genuine phylogenetic signal. Systematic error can negatively affect phylogenetic inference even with single genes, but it becomes stronger when multiple genes are combined into a supermatrix, simply because the addition of more (biased) genes will increase the support for a biased (wrong) result. As expressed by Jeffroy et al. (2006), phylogenomic analyses, rather than resolving the entire tree of life, might in fact be the beginning of incongruence (▶ see Infobox 9.1 for an example). Furthermore, combined datasets from hundreds of genes often contain large amounts of missing data (Roure et al. 2013), which could additionally influence the analysis (▶ see Sect. 9.3).

Infobox 9.1

Which Taxon Is the Sister Group of All Other Animals?

It was basically written in stone that sponges (Porifera) represent the sister taxon of all other animals, and it was rather discussed if sponges are monophyletic or if different sponge taxa branch off subsequently at the base of the animal tree (Sperling et al. 2007; Philippe et al. 2009). However, some phylogenomic analyses surprisingly started to find that the enigmatic Ctenophora (known as comb jellies or sea gooseberries) could represent the sister taxon of animals (Dunn et al. 2008; Moroz et al. 2014). This placement has important implications regarding how the evolution of several organ systems is understood (◘ Fig. 9.2) (Telford et al. 2016). Under the latter hypothesis, it has to be assumed either that the nervous system, muscles and epithelia evolved twice convergently or that all these characters were already present in the last common ancestor of animals and got lost in sponges. This controversy led to a heated debate about phylogenomics methodology and systematic error and how much trust can be put into phylogenomic analyses of very deep divergences. Proponents of the «Porifera-sister» scenario claimed that the result supporting the «Ctenophora-sister» hypothesis represents an LBA artefact, which might be introduced due to a poor fit of the used evolutionary models with the analysed data, as well as by the out-group choice (Pisani et al. 2015). In contrast, proponents of the «Ctenophora-sister» hypothesis analysed the sensitivity of phylogenomic analyses to model and gene choice (Whelan et al. 2015) and used an

extensive taxon sampling. By analysing possible sources of systematic error, no biases affecting the position of Ctenophora as sister taxon to all other animals were found. Instead, some genes included in previous analyses supporting the «Porifera-sister» hypothesis were identified to introduce conflicting signal, thereby supporting a maybe wrong hypotheses of the placement of Ctenophora. This result is in line with a previous study by Nosenko et al. (2013), who by modifying gene and out-group taxon sampling were able to recover three different but well-supported phylogenies of non-bilaterian animals. This controversy remains still unresolved (Giribet 2016) and shifted to the question which models are suited to analyse datasets with massive substitutional heterogeneity and how to perform phylogenomic analyses for deep phylogenies (Whelan and Halanych 2016).

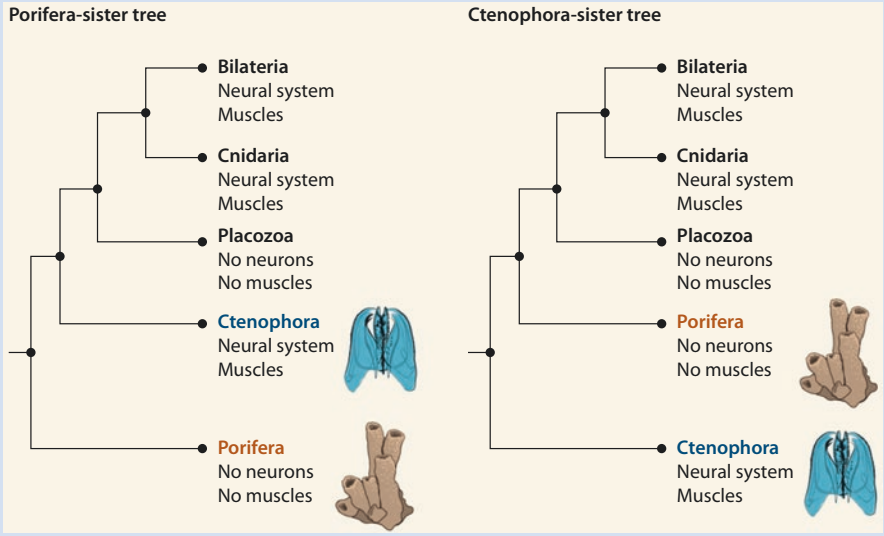


Fig. 9.2 Competing hypotheses regarding which taxon represents the sister group of all other animals and its evolutionary implications (Reprinted by permission from Macmillan Publishers Ltd.: [Nature] (Telford et al. 2016), copyright 2016)

9.2 Systematic Errors

The problem of systematic errors biasing phylogenetic analyses has been recognized early on by Felsenstein (1978). In this paper, he described conditions under which maximum parsimony (MP) inference is misled by the attraction of long branches in a tree irrespective of the true relationships (Fig. 9.3). This phenomenon was termed «long edges attract» by Hendy and Penny (1989), and it is nowadays generally known as long-branch attraction (LBA). Despite maximum likelihood (ML) and Bayesian inference (BI) being more robust than MP to LBA (Philippe et al. 2005b), it was shown that probabilistic phylogenetic reconstruction methods could be also affected by LBA when the assumptions of the underlying model are violated by the data (Huelsenbeck 1995). Many simulation studies have shown that MP is the most sensitive method to the LBA artefact, whereas ML and BI are more robust (Philippe et al. 2005b). Even though LBA is often accounted for when phylogenetic analyses lead to unexpected results, a clear (statistically based) definition of the phenomenon is missing. Some authors defined LBA loosely as a condition where

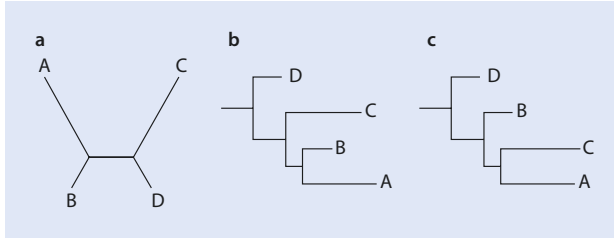


Fig. 9.3 **a** Unrooted four-taxon tree illustrating the classical example of long-branch attraction (LBA), with two long and two short branches, both unrelated. **b** A valid rooted tree of the unrooted topology shown in **a**. **c** Often analyses are misled by LBA, clustering together the long-branched terminals. This rooted topology is a typical artefact occurring in studies with tree A as the underlying true tree

analyses are biased due to a combination of short and long branches (Sanderson et al. 2000; Bergsten 2005), which basically translates to a bias due to variation of the substitution rate across lineages. Parks and Goldman (2014) systematically analysed the placement of long branches using simulation studies and found that also single long branches are difficult to place in a phylogeny, even when using ML. Interestingly they also found that there is no attraction between two long branches, even though they seem to be disproportionally often joined together. This observation has an impact on several approaches which were proposed to detect LBA in real datasets. For example, a common method was to remove one of the long branches from the analysis and to see if the placement of the other long branch remains consistent (Pol and Siddall 2001). However, as also the placement of single long branches is difficult, this might not be a good test. Other approaches to reduce LBA are the exclusion of terminals with very long branches (not an option when they are the taxon of interest) or the exclusion of fast-evolving genes or sites (Bergsten 2005; Pisani 2004; Rivera-Rivera and Montoya-Burgos 2016). Especially classifying all genes (or alignment sites) according to their evolutionary rate and successively removing them from the analysis starting with the fastest class will give a good overview if analyses are biased by the rate signal (Brinkmann et al. 2005). Finally, as LBA is basically a problem of model misspecification, the use of more sophisticated models is recommended. As such, it has been shown that site-heterogeneous CAT models are less affected by LBA due to their ability to better anticipate homoplasy in alignment site patterns (Lartillot et al. 2007), but also ML analyses with carefully selected partitions (and models for each partition) seem to be promising (Whelan and Halanych 2016). In summary, LBA is a very common yet not fully understood phenomenon, and the placement of long branches in phylogenetic analyses remains a difficult task.

Variation in the substitution rate across lineages (rate signal) can lead to the LBA phenomenon (Jeffroy et al. 2006), but this bias can often be handled by using models incorporating rate heterogeneity (Yang 1996). Additionally, the evolutionary rate of an alignment site can vary over time (heterotachy) (Lopez et al. 2002), and this process can also produce LBA (Lockhart and Steel 2005). A specific case of this phenomenon is known as the covarion hypothesis of molecular evolution, which states that substitutions at one alignment site may alter the substitution probability at other sites (Miyamoto and Fitch 1995). Kolaczkowski and Thornton (2004) used a clever simulation scheme to mimic another case of heterotachy. They simulated two sets of sequence alignments using the same topology, but under completely different models of DNA substitutions. By combining these two

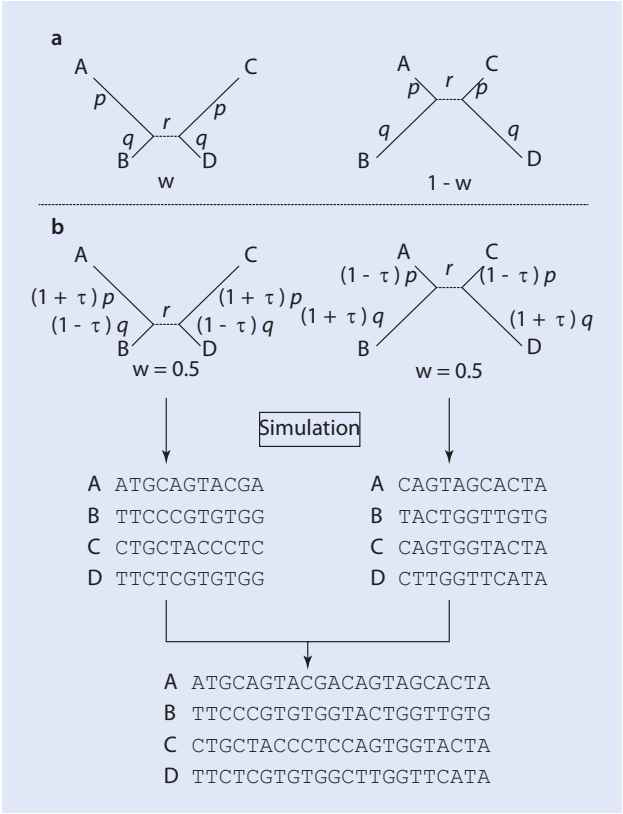


Fig. 9.4 Scheme for the simulation of different levels of heterotachy as used in Kolaczkowski and Thornton (2004). **a** Sequences are simulated under two different sets of branch lengths, including opposing sets of long (p) and short terminal branches. **b** Sequence alignments generated under this simulation scheme can be combined under different weights (w) to simulate different degrees of heterotachy (Figure reprinted from Philippe et al. (2005b))

datasets and giving different weights to the two data partitions, different levels of heterotachy were simulated (Fig. 9.4). Interestingly, these authors found that under higher levels of heterotachy, MP outperforms ML in recovering the correct tree. However, subsequent studies criticized this study for choosing very special and unrealistic parameters for their simulation, as well as for the way how ML analyses were conducted (Philippe et al. 2005b; Spencer et al. 2005). Instead, it could be shown that for realistic simulations of heterotacheous datasets, ML always outperforms MP and should be therefore the preferred method (Philippe et al. 2005b). This phenomenon of heterotachy has been demonstrated to be common in real datasets, where it affects phylogenetic reconstruction (Lopez et al. 2002; Whelan et al. 2011). Some statistical tests for the detection of heterotachy have been proposed (Wu and Susko 2011; Wang et al. 2011). Approaches specifically dealing with heterotachy are the CAT-BP model (Blanquart and Lartillot 2008), as well as a model allowing changing the rate heterogeneity as modelled by the gamma distribution along branches (Bouckaert and Lockhart 2015).

Another systematic error violating model assumptions is compositional bias, which describes significant differences in the nucleotide or amino acid composition across taxa. Most evolutionary models assume that the composition is homogenous across taxa. Several tests for compositional homogeneity are available, including frequency-dependent significance tests, matched-pairs tests or analyses based on Monte Carlo simulations of estimates of the standard deviation of the mean nucleotide or amino acid composition (Steel et al.

1993; Jermiin et al. 2004; Ababneh et al. 2006). With the software SEQVIS, it is possible to visualize compositional heterogeneity in nucleotide alignments (Ho et al. 2006).

A typical example of how compositional bias misleads phylogenetic analyses is that unrelated taxa with convergently evolved elevated GC content might group together, e.g. as demonstrated for drosophilids (Tarrío et al. 2001). Using simulation studies, Jermiin et al. (2004) found that the frequency of successful phylogenetic reconstruction is not only related to the difference in GC content (or base composition) but also to the length of internal branches. Analyses with short internal branches are more easily misled. Compositional bias is also related to rate variation, as especially fast-evolving sites are frequently compositionally biased (Rodríguez-Ezpeleta et al. 2007). Fittingly, third codon positions in protein-coding genes often have a stronger bias in composition, and their removal sometimes increases the accuracy of the phylogenetic analysis. One of the many negative effects of compositional heterogeneity can be the accumulation of convergencies. For example, transitions (replacement of a purine by a purine or pyrimidine by a pyrimidine) are usually more frequently observed than transversions (replacement of a purine by a pyrimidine or reverse), leading to coincident substitutions. It has been shown that recoding all nucleotides to R (purines, A and G) and Y (pyrimidines, C and T) reduces this misleading effect of compositional bias (Phillips and Penny 2003). Recoding can, for example, be conducted with the software BMGE (Criscuolo and Gribaldo 2010), which furthermore is able to identify and exclude characters which contribute to compositional biases based on a matched-pairs test of marginal symmetry. Finally, non-homogeneous nonstationary models that account for variations in the base composition can be used. The model of DNA sequence evolution by Galtier and Gouy (1998), which is implemented in PHYML (Boussau and Gouy 2006), allows varying equilibrium GC contents among lineages and estimation of five parameters: (I) ancestral GC content, (II) location of the root in its branch, (III) transition/transversion ratio, (IV) branch lengths and (V) equilibrium GC contents in each branch. Compositional bias was expected to be more frequent and also misleading on the nucleotide level, as only four different states exist and convergence is to be expected (Hasegawa and Hashimoto 1993; Foster and Hickey 1999). However, compositional bias on the protein level seems also to be frequent and thereby a problem for phylogenetic analyses as well (Lartillot and Philippe 2008; Nesnidal et al. 2010). Kück and Struck (2014) developed a package of scripts to analyse phylogenomic datasets (BACOCA), which can be used to investigate the compositional bias among amino acids. As with nucleotides, recoding of amino acids can reduce the compositional bias. The most commonly used recoding classifies amino acids according to six groups identified by Dayhoff et al. (1978), which tend to replace each other (Susko and Roger 2007). Furthermore, using the CAT-BP model for amino acid data allows lineage-specific compositional shifts across the phylogeny and thus deals with heterogeneous amino acid sequence compositions (Blanquart and Lartillot 2008).

9.3 Missing Data, Phylogenetic Information Content and Taxon Sampling

9.3.1 Missing Data

A typical way to compile a dataset for phylogenomic studies involves the generation of transcriptomes and subsequent selection of putative orthologs for the analyses. Ortholog sets often range from 100 to more than 1000 genes, and it is not unusual that not all genes

are (completely) recovered for all taxa. As such, orthologs are often found incomplete using transcriptome sequencing (Fig. 9.5a). In most cases, missing genes are due to the depth of the sequenced transcriptome or they are just not expressed in the sampled specimen (Roure et al. 2013). Moreover, many genes might have been lost for some taxa during evolution (Fig. 9.5b). Percentages of missing data up to 80% have been reported for phylogenomic studies (Hejnol et al. 2009). The discussion if missing data should be reduced from phylogenetic analyses, e.g. excluding the most incomplete taxa and/or characters, has a long tradition in the literature (Wiens 2003; Wiens and Morrill 2011; Philippe et al. 2004; Wiens 1998). Initially, the question arose if incompletely sampled taxa should be included in phylogenetic analyses of one or few genes or in morphological character matrices. In the latter case, the discussion often centred on fossils, for which it was usually impossible to analyse all characters found in recent taxa. Later the discussion was expanded to genomic datasets, where often substantial amounts of data are missing. Even though some publications addressed missing data as problematic (Lemmon et al. 2009), most studies using real or simulated data could show that the inclusion of incomplete taxa is usually advantageous. One simple reason is that an improved taxon sampling

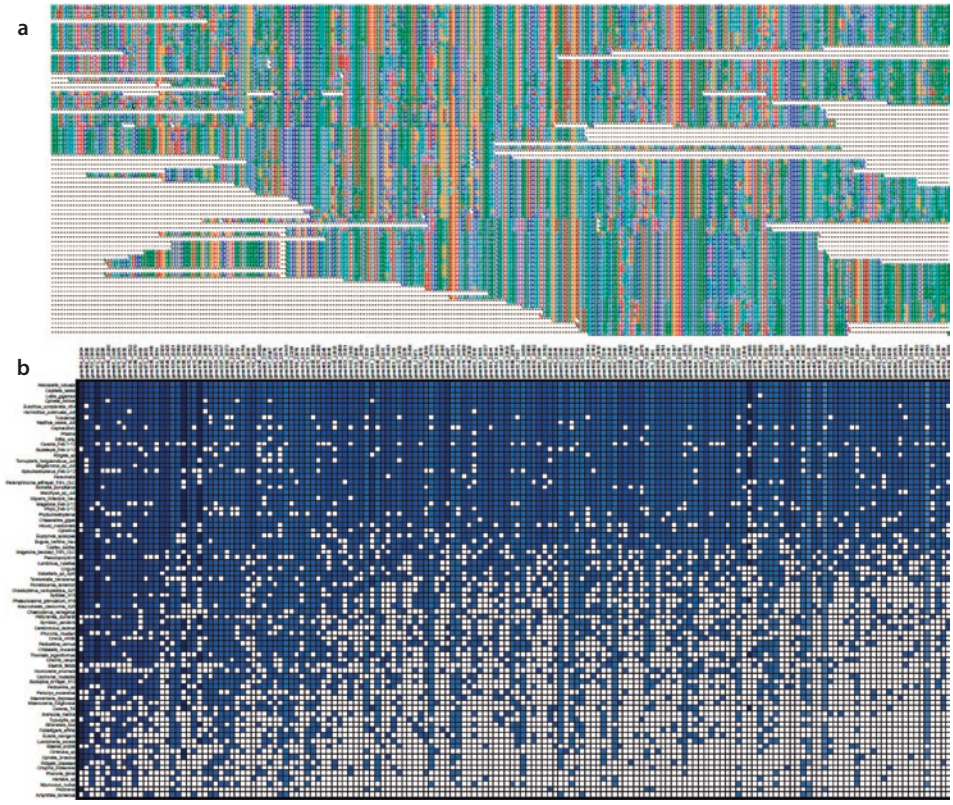


Fig. 9.5 Missing data in phylogenomic analyses. **a** Single gene alignment based on transcriptomic data often includes highly incomplete and partially nonoverlapping gene sequences. **b** The gene coverage (*columns*) is often highly uneven for taxa (*rows*) included in a phylogenomic study. *Blue squares* show presence of genes, *white squares* show absent genes. Matrix based on data from Weigert et al. (2014) constructed with MARE (Misof et al. 2013)

helps to break long branches (Roure et al. 2013). By analysing a large dataset covering diverse eukaryotes, Philippe et al. (2004) could show that 25% of missing data in the original dataset did not negatively impact the analyses. Subsequent random deletion of 50% of the character matrix did not alter the outcome of the analysis, and even when analysing with up to 90% of missing data, similar trees could be obtained. Jiang et al. (2014) found that that adding incomplete data is in particular helpful for resolving poorly supported nodes and showed that missing data does not consistently bias branch lengths. Finally, Hovmöller et al. (2013) have shown that also species tree reconstruction methods relying on coalescent approaches (▶ see Sect. 9.4) are remarkably robust under the presence of up to 50% of missing data. However, if missing data is nonrandomly distributed over the matrix, it may bias analyses, leading to many trees (or subtrees) which are nearly indistinguishable by its likelihood value (Sanderson et al. 2010). A tool for the visualization of the completeness of the supermatrix (◘ Fig. 9.5b), as well as for the exclusion of incompletely sampled genes, is the software MARE (Misof et al. 2013). Using such an approach, differently covered data matrices can be constructed and analysed, and the sensitivity of phylogenomic analyses to missing data can be assessed (Weigert et al. 2014).

9.3.2 More Genes or More Taxa?

Taxon sampling has been profusely discussed in the phylogenetic literature prior to the genomic era. In particular, whether it was better centres the efforts in obtaining more data for a number of taxa or more taxa with relatively fewer data (Rokas and Carroll 2005; Mitchell et al. 2000). This discussion lost power with the (comparatively) cheap price of NGS technologies, which allows the recovery of large amounts of sequences for non-model taxa, and in most cases adding more data is not a bottleneck anymore. The first phylogenomic analyses often relied on a handful of model taxa where complete genomes were available. For example, focussing on animal relationships, these analyses seemed to support the so-called Coelomata hypothesis (arthropods + deuterostomes) and not the widely accepted Ecdysozoa hypothesis (arthropods + nematodes) (Philip et al. 2005). However, these results have been clearly demonstrated to be an artefact related to a limited taxon sampling (Philippe et al. 2005a). The discussion of experimental design has now shifted to which genes and which taxa to include in an analysis (Philippe et al. 2011).

9.3.3 Taxon Sampling

The importance of taxon sampling for phylogenetic analyses is widely acknowledged (Heath et al. 2008; Pollock et al. 2002; Zwickl and Hillis 2002), with only few studies coming to a different conclusion (Rosenberg and Kumar 2001). Rannala et al. (1998) demonstrated in a simulation study that a decrease in taxon sampling leads to an increase in the average branch length of terminals, which could make analyses more susceptible to LBA. This is in line with the finding that the estimation of rate heterogeneity is highly sensitive to taxon sampling (Sullivan et al. 1999). Moreover, estimation of branch lengths becomes also more challenging due to the so-called node density effect under a limited taxon sampling (Hugall and Lee 2007). This effect often leads to an underestimation of branch lengths in sparsely sampled tree regions, because less information is available to infer multiple substitutions, which could have been revealed under the presence of

additional nodes. However, not all included taxa are equally helpful to improve phylogenetic analyses. Certain taxa, so-called rogue taxa, can show a phylogenetically unstable behaviour, characterized by widely different positions in tree topologies estimated from the same dataset (e.g. within bootstrap replicates) (Sanderson and Shaffer 2002). Often, but not always, rogue taxa are characterized by showing large amounts of missing data. Inclusion of such rogue taxa can have a negative impact on support values (especially when using bootstrap), but could also influence tree reconstruction in general (Mariadassou et al. 2012). In fact, Aberer et al. (2013) demonstrated that exclusion of rogue taxa increases the accuracy of phylogenetic analyses. These authors developed an algorithm for the identification and subsequent pruning of rogue taxa, implemented in the software ROGUENAROK. The idea behind the algorithm is to identify taxa, which exclusion results into an increase of support in bootstrap consensus trees. The measure of change in support is called relative bipartition information criterion (RBIC), which is the sum of all support values divided by the maximum support in a fully bifurcating tree of the original dataset. Taxa or combinations of taxa yielding the highest change in RBIC are excluded from the analysis. This analysis can be iteratively repeated until no significant change is observed. Alternatively, the leave stability index (LSI) has been used to identify rogue taxa. The LSI uses the occurrence of taxon triplets in trees from bootstrap analyses (Thorley and Wilkinson 1999). Three different possibilities for the relationship of three taxa (A, B, C) exist in a rooted, bifurcated tree: ((A, B), C), ((A, C), B) and ((B, C), A). The LSI is calculated as the difference of the relative frequency of the most common triplet and the second most common and is averaged over all triplets containing a certain taxon. LSI values of 1 or close to 1 indicate stable taxa, where values closer to 0 indicate instability. A LSI cut-off value can be defined for rogue taxa to be excluded from the analysis. Inference of the LSI is, for example, incorporated in the software PHYUTILITY (Smith and Dunn 2008). A third approach called multiple co-inertia analysis (MCOA) has been explored by de Vienne et al. (2012), which is based on the comparison of pairwise distances between species in all gene tree topologies to identify rogue taxa (described as outlier taxa in this publication).

9.3.4 Gene Sampling

Gene alignments can differ in their missing data, sequence saturation or phylogenetic information content. DNA and protein sequences are regarded as saturated, when sites have undergone multiple substitutions and the number of observed differences no longer reflects «true» evolutionary distances. Slight levels of saturation are corrected by the use of models of sequence evolution, but more saturated sequence alignments can mislead phylogenetic reconstruction. When analysing highly saturated sequences, phylogenetic inference can be driven by sequence composition to a large extent rather than true phylogeny (Xia et al. 2003). DNA sequences are normally more affected by saturation because only four different character states exist compared to the 20 states of amino acids (Philippe et al. 2011). However, saturation can also be problematic at the amino acid level (Van de Peer et al. 2002). A simple method to check for the presence of saturation in nucleotide sequences is by separately plotting the raw numbers of substitutions (p uncorrected distance) of transitions and transversions of all pairwise comparisons of taxa in an alignment against their genetic (usually ML-corrected) distance (Struck et al. 2008). For most protein-coding genes, transitions occur more frequently than transversions and thus are

more likely saturated (Fig. 9.6). Formalized measures of substitution saturation have been introduced by Xia et al. (2003), as implemented in the software DAMBE (Xia 2013), and Struck et al. (2008), as implemented in the BACOCA package of scripts (Kück and Struck 2014). Possible strategies to deal with saturated sequences are use of amino acids, exclusion of the saturated data or recoding (e.g. RY coding or the use of Dayhoff categories for amino acids).

It is important to remember the relationship between sequence saturation and sequence divergence: one gene might be saturated for old divergences but well suited to resolve young divergences, whereas a slower-evolving gene might not be saturated for old divergences but totally uninformative for young ones. The usefulness of a given gene for phylogenetic analyses can be estimated by its phylogenetic informativeness (PI) (Townsend 2007). Briefly summarized, PI estimates the probability that a character resolves a dated four-taxon alignment (more than four taxa can be analysed by providing a consistent topology). Thereby, PI provides an estimate of the amount of phylogenetic signal relative to noise across time (Fig. 9.7). PI can be analysed using the software PHYDESIGN (López-Giráldez and Townsend 2011), which is available online, by providing an alignment, as well as an ultrametric tree as input. Some updates and modifications for the calculation of PI are available in the R package PHYLINFORMR (Dornburg et al. 2016). As an example on how to use PI, in Fig. 9.7, the utility of different classes of phylogenetic markers from percomorph fishes are compared (Gilbert et al. 2015).

A different approach to investigate and visualize phylogenetic information content is based on likelihood mapping (Strimmer and von Haeseler 1997). This method analyses possible four-taxon cases of a given dataset, called quartets. For every quartet, there are three possible fully resolved tree topologies, for which the posterior probability for each of the three possible topologies can be estimated using Bayes' theorem. The three

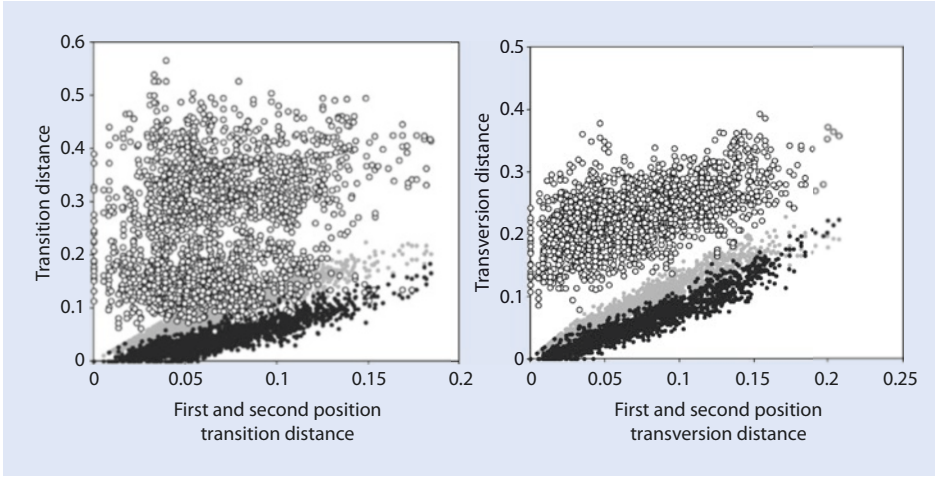


Fig. 9.6 Saturation at different codon positions. Uncorrected pairwise distances are plotted for pairs of taxa, separately for transitions (*left*) and transversions (*right*) and first (*grey*), second (*black*) and third (*white*) codon positions. For unsaturated sequences, the number of substitutions should increase linearly with time (e.g. transversions on first and second positions), whereas for saturated sequences, no increase in the number of substitutions is detected with increasing genetic distance (e.g. transitions on third codon positions) (Reprinted from Dávalos and Perkins (2008), with permission from Elsevier)

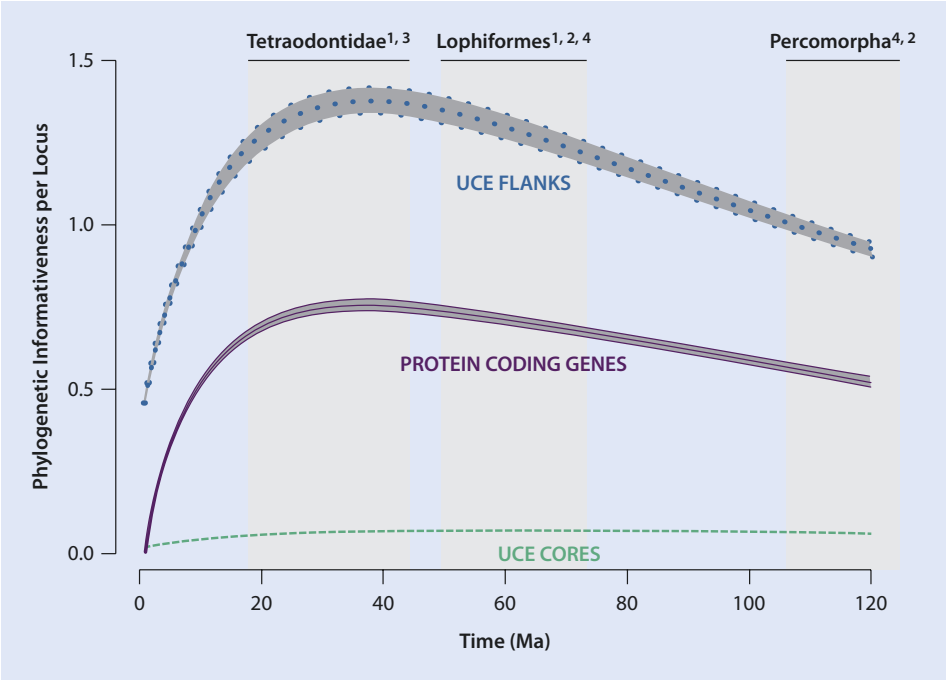


Fig. 9.7 Phylogenetic informativeness and its 95% confidence interval of three different classes of phylogenetic markers from percomorph fishes (UCE core regions, UCE flanking regions, protein-coding genes) plotted against time. Core regions of ultraconserved elements (UCEs) are basically uninformative, whereas flanking regions of UCE show a higher PI than protein-coding genes, with the highest resolution power for divergences between 20 and 40 million years old (Reprinted from (Gilbert et al. 2015), with permission from Elsevier)

posterior probabilities are then used as coordinates to locate a point within a triangular graph where each corner represents one topology. This calculation is repeated for all possible quartets, which are subsequently plotted in the triangle. In the case of an uninformative quartet (starlike evolution), all three probabilities are the same and the point is located in the middle of the triangle. If one tree topology is clearly supported with a probability close to 1, this would point to one of the corners of the triangle (according the supported topology). If two topologies gain similar probability, whereas one topology gets a probability close to 0, the point would be located at one edge of the triangle, between the corners representing the two supported topologies. By analysing all possible quartets of a dataset, the phylogenetic information content can be visualized. The more quartets can be located in one of the corners of the triangle, the higher is the information content of the dataset (Fig. 9.8). Likelihood mapping is implemented in the software TREE-PUZZLE (Schmidt et al. 2002).

Different strategies have been used to select sets of orthologous genes for phylogenetic analyses. Some authors recommend to only include highly informative genes in the analysis (Salichos and Rokas 2013), whereas others suggest that phylogenetic signal can be basically extracted from all ortholog alignments when combined in a supermatrix (Gatesy and Baker 2005). PI represents a possible way to select genes which are suitable for both, supermatrix and coalescent-based methods. Shen et al. (2016) systematically analysed the

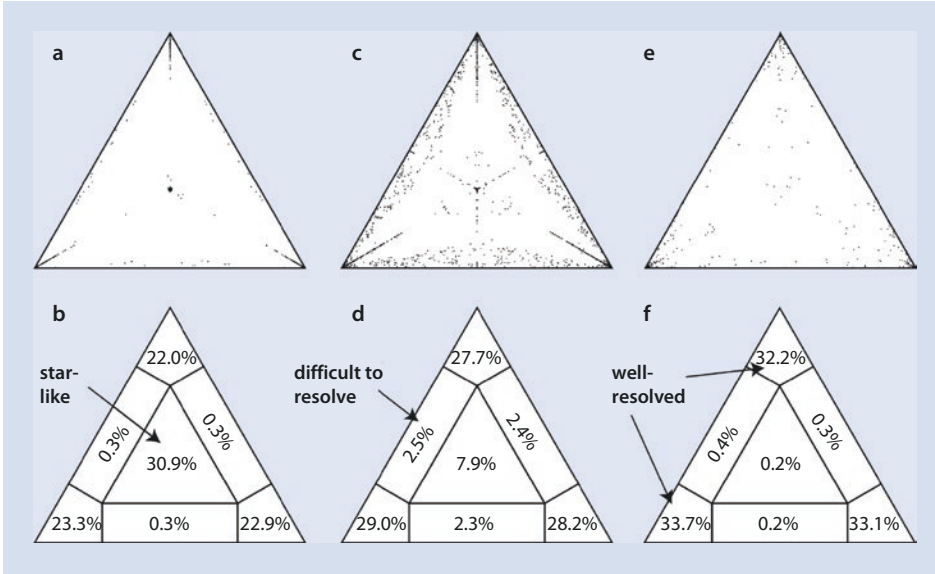


Fig. 9.8 Likelihood mapping using TREE-PUZZLE (Schmidt et al. 2002) for datasets with differences in phylogenetic information content. **a**, **b** In a dataset with low information content, a high percentage (30.9%) of quartets represent starlike evolution. **c**, **d** In this dataset 7.9% of the quartets represent starlike evolution, whereas 2.5% + 2.3% + 2.4% of quartets are in an area where it is difficult to distinguish between two of the three possible tree alternatives. **e**, **f** Most quartets (33.7% + 32.2%, 33.1%) are in well-resolved areas of the tree distribution, indicating high phylogenetic information content. **a**, **c**, **e** show distribution patterns of mapped quartets; **b**, **d**, **f** show occupancies (in percent) for seven areas of interest

association between sequence-based properties, gene function-based properties and gene tree-based properties with phylogenetic information content. The goal was to identify those properties which predict phylogenetic signal of a gene best. Even though most of the investigated properties correlate with each other, a set of properties with the highest relevance could be identified. Interestingly, the most important property to predict phylogenetic signal is gene alignment length, followed by number of parsimony-informative sites and variable sites. This result could be interpreted in favour of binning genes for coalescent analyses (see above), but also for the use of the supermatrix approach, which basically combines all alignments into a highly informative «supergene».

9.4 Incongruence Between Gene Trees and Species Trees

Gene trees may differ from the species tree simply by the stochastic sampling of alleles during speciation events (Degnan and Rosenberg 2009), a phenomenon known as incomplete lineage sorting or deep coalescence (Fig. 9.9). The term «hemiplasy» has been coined to describe incorrect inference of character-state evolution due to genetic polymorphisms which are retained across speciation events (Avise and Robinson 2008; Hahn and Nakhleh 2016). This term should reflect that in this case similarity does not reflect common ancestry, even though the considered character states are homologous (and apomorphic!).

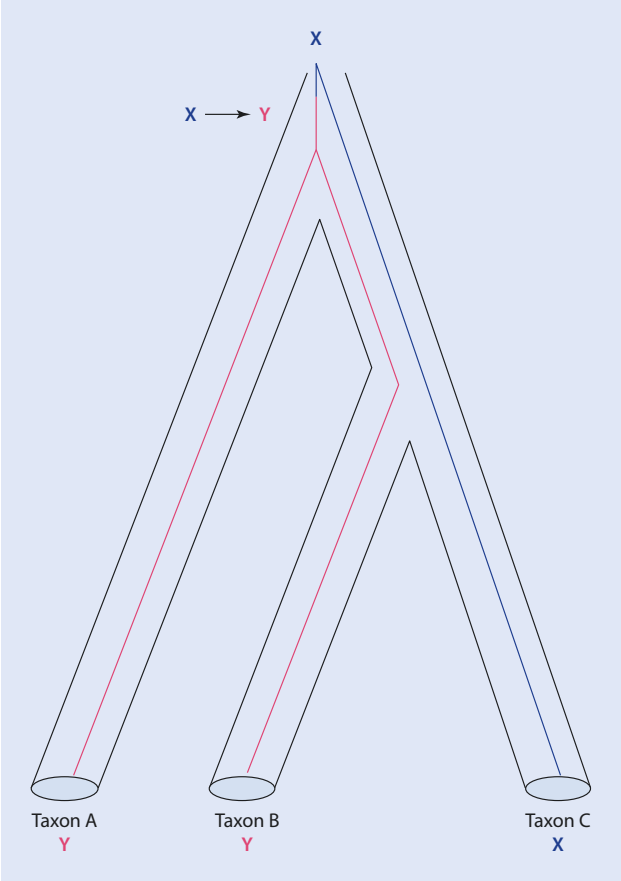


Fig. 9.9 Incomplete lineage sorting can lead to incongruence between gene trees and species trees. The gene tree is drawn in colour inside the species tree (*black*). The last common ancestor of taxa **a**–**c** had two paralogs of a gene (X and Y). Duplicates got lost before the split of the three species, but paralog sorting is incongruent with the species tree

It has been demonstrated that discordance between gene trees and species trees is common, especially in cases where speciation events happened in short time spans, i.e. separated by short branches (Degnan and Rosenberg 2006). A good example of incomplete lineage sorting is represented by the genome-scale analyses of the bird phylogeny, which includes a rapid radiation characterized by many short internal branches. For this phylogeny, not a single gene tree has been found to match the reconstructed species tree (Jarvis et al. 2014). Later on, lineage sorting has been shown to be frequent in the evolutionary history of birds, and a phylogenetic network was used to illustrate their complex history (Fig. 9.10) (Suh et al. 2015).

Several other evolutionary processes can lead to the disagreement between gene trees and species trees, including horizontal gene transfer (HGT), gene duplication and hybridization (Maddison 1997; Knowles and Kubatko 2010). HGT is a process where genes are transferred from one species to another across the phylogeny. Whereas HGT is rather rare in eukaryotes and therefore less problematic for phylogenetic reconstruction, it is common among prokaryotes (Ku and Martin 2016). Gene duplication complicates the inference of orthology (Philippe et al. 2011). Hybridization and introgression are biological processes by which the genetic material of two different species gives rise to hybrids and sometimes new species. Hybridization is most commonly found in plants, but also many examples have been described for animals (Mallet 2007).

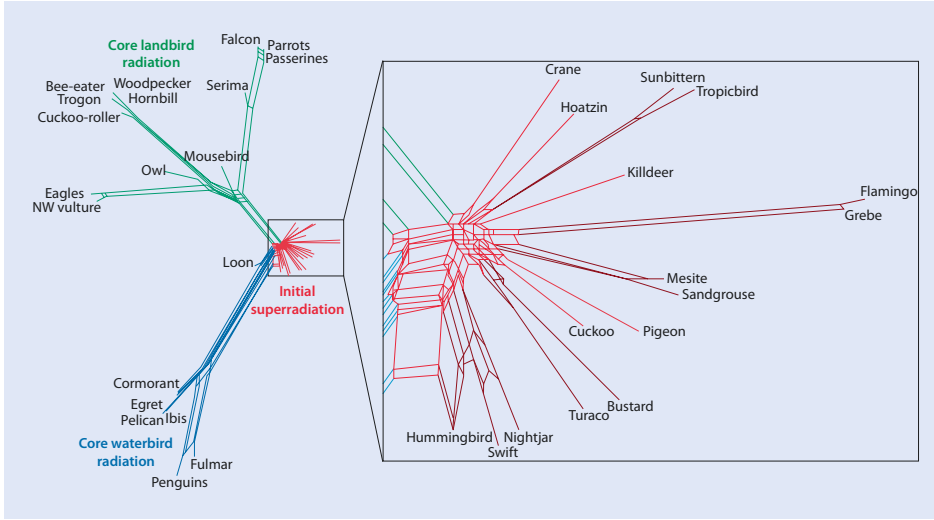


Fig. 9.10 Phylogenetic network analyses of rare genomic change markers reveal a strong discordance of markers, which can be explained by high levels of incomplete lineage sorting (Figure reprinted from Suh et al. (2015))

9

Several phylogenetic methods have been developed to detect and deal with incongruence of gene trees and species trees. In contrast to the supermatrix approach, where genes are concatenated into one single matrix, these methods are usually based on the separate reconstruction of gene trees, which are subsequently (or simultaneously) used to infer the species tree. Most species tree inference methods are rooted within the coalescence theory, a model which has been developed to follow the history of genes (or alleles) back in time. Coalescence models are commonly used in population genetics and are often based on the Wright-Fisher model of genetic drift, assuming nonoverlapping generations, neutral evolution and random joining of populations back in time (Degnan and Rosenberg 2009). The multispecies coalescent (MSC) is used to estimate the probability distribution of gene trees evolving along the branches of a species tree. Each branch of a species tree represents a single population, and lineages of genes entering these populations are traced back through time to a common ancestor at rates given by the model. The coalescence of different gene lineages of the gene trees finally provides the signal for the inference of the overlying species tree (Liu et al. 2015). The MSC has been implemented into ML approaches, e.g. STEM (Kubatko et al. 2009) or MP-EST (Liu et al. 2010), and a Bayesian framework, e.g. BEST (Liu 2008) or BEAST (Drummond et al. 2012). The performance of species tree inference methods is controversially discussed. Gatesy and Springer (2014) criticized that species tree inference is often misled by unreliable gene trees, especially when dealing with phylogenetic analyses at deep timescales. Similar to the idea that the phylogenetic signal-to-noise ratio gets improved by using concatenation of single gene alignments into a supermatrix, statistical binning of genes with a similar signal has been proposed to reduce gene tree estimation errors for species tree inference (Mirarab et al. 2014). Several simulation studies show a superior performance of species tree inference using a Bayesian framework in comparison with other methods, especially in the case when a high probability of gene tree discordance is simulated (Leaché and Rannala 2011). Interestingly, comparison of results from species tree inference and supermatrix methods for real datasets often show rather consistent results (Liu et al. 2015).

For the quantification of incongruence in phylogenomic datasets, Salichos and Rokas (2013) developed a measure called internode certainty (IC). Here, incongruence for a given internal node is measured by calculating the frequency of a bipartition found in the best tree in a given set of gene trees together with the occurrence of conflicting bipartition in these gene trees. Values close to 0 indicate the presence of strong conflict, whereas values close to 1 indicate the absence of conflictive signal. Summing overall ICs will give the tree certainty (TC). The calculation of IC and TC is implemented within the software RAxML (Stamatakis 2014; Kobert et al. 2016).

References

Ababneh F, Jermiin LS, Ma C, Robinson J (2006) Matched-pairs tests of homogeneity with applications to homologous nucleotide sequences. Bioinformatics 22:1225–1231

Aberer AJ, Krompass D, Stamatakis A (2013) Pruning rogue taxa improves phylogenetic accuracy: an efficient algorithm and webservice. Syst Biol 62:162–166

Adl SM, Simpson AGB, Farmer MA, Andersen RA, Anderson OR, Barta JR, Bowser SS, Brugerolle GUY, Fensome RA, Fredericq S, James TY, Karpov S, Kugrens P, Krug J, Lane CE, Lewis LA, Lodge J, Lynn DH, Mann DG, McCourt RM, Mendoza L, Moestrup Ø, Mozley-Standridge SE, Nerad TA, Shearer CA, Smirnov AV, Spiegel FW, Taylor MFJR (2005) The new higher level classification of eukaryotes with emphasis on the taxonomy of protists. J Eukaryot Microbiol 52:399–451

Avise JC, Robinson TJ (2008) Hemiplasy: a new term in the lexicon of phylogenetics. Syst Biol 57:503–507

Bergsten J (2005) A review of long-branch attraction. Cladistics 21:163–193

Bininda-Emonds ORP (2004) The evolution of supertrees. Trends Ecol Evol 19:315–322

Blanquart S, Lartillot N (2008) A site- and time-heterogeneous model of amino acid replacement. Mol Biol Evol 25:842–858

Bouckaert R, Lockhart P (2015) Capturing heterotachy through multi-gamma site models. bioRxiv. doi.org/10.1101/018101

Boussau B, Gouy M (2006) Efficient likelihood computations with nonreversible models of evolution. Syst Biol 55:756–768

Brinkmann H, van der Giezen M, Zhou Y, de Raucourt GP, Philippe H (2005) An empirical assessment of long-branch attraction artefacts in deep eukaryotic phylogenomics. Syst Biol 54:743–757

Criscuolo A, Gribaldo S (2010) BMGE (Block mapping and gathering with entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments. BMC Evol Biol 10:210

Dávalos LM, Perkins SL (2008) Saturation and base composition bias explain phylogenomic conflict in Plasmodium. Genomics 91:433–442

Dayhoff M, Schwarz R, Orcutt B (1978) A model of evolutionary change in proteins. In: Dayhoff M (ed) Atlas of protein sequence and structure, vol 5, Suppl. 3. National Biomedical Research Foundation. Washington, DC, pp 345–352

de Queiroz A, Gatesy J (2007) The supermatrix approach to systematics. Trends Ecol Evol 22:34–41

de Vienne DM, Ollier S, Aguileta G (2012) Phylo-MCOA: a fast and efficient method to detect outlier genes and species in phylogenomics using multiple co-inertia analysis. Mol Biol Evol 29:1587–1598

Degnan JH, Rosenberg NA (2006) Discordance of species trees with their most likely gene trees. PLoS Genet 2:e68

Degnan JH, Rosenberg NA (2009) Gene tree discordance, phylogenetic inference and the multispecies coalescent. Trends Ecol Evol 24:332–340

Donoghue MJ, Doyle JA (2000) Seed plant phylogeny: demise of the anthophyte hypothesis? Curr Biol 10:R106–R109

Dornburg A, Fisk JN, Tamagnan J, Townsend JP (2016) PhyInformR: phylogenetic experimental design and phylogenomic data exploration in R. BMC Evol Biol 16:262

Drummond AJ, Suchard MA, Xie D, Rambaut A (2012) Bayesian Phylogenetics with BEAUti and the BEAST 1.7. Mol Biol Evol 29:1969–1973

Dunn CW, Hejnol A, Matus DQ, Pang K, Browne WE, Smith SA, Seaver E, Rouse GW, Obst M, Edgecombe GD, Sorensen MV, Haddock SHD, Schmidt-Rhaesa A, Okusu A, Kristensen RM, Wheeler WC, Martindale

MQ, Giribet G (2008) Broad phylogenomic sampling improves resolution of the animal tree of life. Nature 452:745–750
Felsenstein J (1978) Cases in which parsimony or compatibility methods will be positively misleading. Syst Zool 27:401–410
Foster PG, Hickey DA (1999) Compositional bias may affect both DNA-based and protein-based phylogenetic reconstructions. J Mol Evol 48:284–290
Galtier N, Gouy M (1998) Inferring pattern and process: maximum-likelihood implementation of a nonhomogeneous model of DNA sequence evolution for phylogenetic analysis. Mol Biol Evol 15:871–879
Gatesy J, Baker RH (2005) Hidden likelihood support in genomic data: can forty-five wrongs make a right? Syst Biol 54:483–492
Gatesy J, DeSalle R, Wahlberg N (2007) How many genes should a systematist sample? Conflicting insights from a phylogenomic matrix characterized by replicated incongruence. Syst Biol 56:355–363
Gatesy J, Springer MS (2014) Phylogenetic analysis at deep timescales: unreliable gene trees, bypassed hidden support, and the coalescence/concatalescence conundrum. Mol Phylogenet Evol 80:231–266
Gee H (2003) Evolution: ending incongruence. Nature 425:782–782
Gilbert PS, Chang J, Pan C, Sobel EM, Sinsheimer JS, Faircloth BC, Alfaro ME (2015) Genome-wide ultraconserved elements exhibit higher phylogenetic informativeness than traditional gene markers in percomorph fishes. Mol Phylogenet Evol 92:140–146
Giribet G (2016) Genomics and the animal tree of life: conflicts and future prospects. Zool Scr 45:14–21
Hahn MW, Nakhleh L (2016) Irrational exuberance for resolved species trees. Evolution 70:7–17
Halanych KM (2004) The new view of animal phylogeny. Annu Rev Ecol Syst 35:229–256
Hasegawa M, Hashimoto T (1993) Ribosomal RNA trees misleading? Nature 361:23–23
Heath TA, Hedtke SM, Hillis DM (2008) Taxon sampling and the accuracy of phylogenetic analyses. J Syst Evol 46:239–257
Hejnol A, Obst M, Stamatakis A, Ott M, Rouse GW, Edgecombe GD, Martinez P, Baguñà J, Bailly X, Jondelius U, Wiens M, Müller WEG, Seaver E, Wheeler WC, Martindale MQ, Giribet G, Dunn CW (2009) Assessing the root of bilaterian animals with scalable phylogenomic methods. Proc R Soc Lond B Biol Sci 276:4261–4270
Hendy MD, Penny D (1989) A framework for the quantitative study of evolutionary trees. Syst Biol 38:297–309
Ho JWK, Adams CE, Lew JB, Matthews TJ, Ng CC, Shahabi-Sirjani A, Tan LH, Zhao Y, Easteal S, Wilson SR, Jermiin LS (2006) SeqVis: visualization of compositional heterogeneity in large alignments of nucleotides. Bioinformatics 22:2162–2163
Hovmöller R, Lacey Knowles L, Kubatko LS (2013) Effects of missing data on species tree estimation under the coalescent. Mol Phylogenet Evol 69:1057–1062
Huelsenbeck JP (1995) Performance of phylogenetic methods in simulation. Syst Biol 44:17–48
Hugall AF, Lee MSY (2007) The likelihood node density effect and consequence for evolutionary studies of molecular rates. Evolution 61:2293–2307
Jarvis ED, Mirarab S, Aberer AJ, Li B, Houde P, Li C, Ho SYW, Faircloth BC, Nabholz B, Howard JT, Suh A, Weber CC, da Fonseca RR, Li J, Zhang F, Li H, Zhou L, Narula N, Liu L, Ganapathy G, Boussau B, Bayzid MS, Zavidovych V, Subramanian S, Gabaldón T, Capella-Gutiérrez S, Huerta-Cepas J, Rekepalli B, Munch K, Schierup M, Lindow B, Warren WC, Ray D, Green RE, Bruford MW, Zhan X, Dixon A, Li S, Li N, Huang Y, Derryberry EP, Bertelsen MF, Sheldon FH, Brumfield RT, Mello CV, Lovell PV, Wirthlin M, Schneider MPC, Prosdocimi F, Samaniego JA, Velazquez AMV, Alfaro-Núñez A, Campos PF, Petersen B, Sicheritz-Ponten T, Pas A, Bailey T, Scofield P, Bunce M, Lambert DM, Zhou Q, Perelman P, Driskell AC, Shapiro B, Xiong Z, Zeng Y, Liu S, Li Z, Liu B, Wu K, Xiao J, Yinqi X, Zheng Q, Zhang Y, Yang H, Wang J, Smeds L, Rheindt FE, Braun M, Fjeldsa J, Orlando L, Barker FK, Jønsson KA, Johnson W, Koepfli K-P, O'Brien S, Haussler D, Ryder OA, Rahbek C, Willerslev E, Graves GR, Glenn TC, McCormack J, Burt D, Ellegren H, Alström P, Edwards SV, Stamatakis A, Mindell DP, Cracraft J, Braun EL, Warnow T, Jun W, Gilbert MTP, Zhang G (2014) Whole-genome analyses resolve early branches in the tree of life of modern birds. Science 346:1320–1331
Jeffroy O, Brinkmann H, Delsuc F, Philippe H (2006) Phylogenomics: the beginning of incongruence? Trends Genet 22:225–231
Jermiin LS, Ho SYW, Ababneh F, Robinson J, Larkum AWD (2004) The biasing effect of compositional heterogeneity on phylogenetic estimates may be underestimated. Syst Biol 53:638–643
Jiang W, Chen S-Y, Wang H, Li D-Z, Wiens JJ (2014) Should genes with missing data be excluded from phylogenetic analyses? Mol Phylogenet Evol 80:308–318

Knowles LL, Kubatko LS (2010) Estimating species trees: an introduction to concepts and models. In: Knowles LL, Kubatko LS (eds) Estimating species trees: practical and theoretical aspects. Wiley-Balckwell, Hoboken, pp 1–14
Kobert K, Salichos L, Rokas A, Stamatakis A (2016) Computing the internode certainty and related measures from partial gene trees. Mol Biol Evol 33:1606–1617
Kolaczkowski B, Thornton JW (2004) Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous. Nature 431:980–984
Ku C, Martin WF (2016) A natural barrier to lateral gene transfer from prokaryotes to eukaryotes revealed from genomes: the 70% rule. BMC Biol 14:89
Kubatko LS, Carstens BC, Knowles LL (2009) STEM: species tree estimation using maximum likelihood for gene trees under coalescence. Bioinformatics 25:971–973
Kück P, Struck TH (2014) BaCoCa—a heuristic software tool for the parallel assessment of sequence biases in hundreds of gene and taxon partitions. Mol Phylogenet Evol 70:94–98
Kumar S, Filipski AJ, Battistuzzi FU, Kosakovsky Pond SL, Tamura K (2012) Statistics and Truth in Phylogenomics. Mol Biol Evol 29:457–472
Lartillot N, Brinkmann H, Philippe H (2007) Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model. BMC Evol Biol 7:S4
Lartillot N, Philippe H (2008) Improvement of molecular phylogenetic inference and the phylogeny of Bilateria. Philos Trans R Soc Lond Ser B Biol Sci 363:1463–1472
Leaché AD, Rannala B (2011) The accuracy of species tree estimation under simulation: a comparison of methods. Syst Biol 60:126–137
Lemmon AR, Brown JM, Stanger-Hall K, Lemmon EM (2009) The effect of ambiguous data on phylogenetic estimates obtained by maximum likelihood and bayesian inference. Syst Biol 58:130–145
Liu L (2008) BEST: bayesian estimation of species trees under the coalescent model. Bioinformatics 24:2542–2543
Liu L, Xi Z, Wu S, Davis CC, Edwards SV (2015) Estimating phylogenetic trees from genome-scale data. Ann N Y Acad Sci 1360:36–53
Liu L, Yu L, Edwards SV (2010) A maximum pseudo-likelihood approach for estimating species trees under the coalescent model. BMC Evol Biol 10:302
Lockhart P, Steel M (2005) A tale of two processes. Syst Biol 54:948–951
López-Giráldez F, Townsend JP (2011) PhyDesign: an online application for profiling phylogenetic informativeness. BMC Evol Biol 11:152
Lopez P, Casane D, Philippe H (2002) Heterotachy, an important process of protein evolution. Mol Biol Evol 19:1–7
Maddison WP (1997) Gene trees in species trees. Syst Biol 46:523–536
Mallet J (2007) Hybrid speciation. Nature 446:279–283
Mariadassou M, Bar-Hen A, Kishino H (2012) Taxon influence index: assessing taxon-induced incongruities in phylogenetic inference. Syst Biol 61:337–345
Mirarab S, Bayzid MS, Boussau B, Warnow T (2014) Statistical binning enables an accurate coalescent-based estimation of the avian tree. Science 346 1250463.
Misof B, Meyer B, von Reumont BM, Kück P, Misof K, Meusemann K (2013) Selecting informative subsets of sparse supermatrices increases the chance to find correct trees. BMC Bioinformatics 14:348
Mitchell A, Mitter C, Regier JC (2000) More taxa or more characters revisited: combining data from nuclear protein-encoding genes for phylogenetic analyses of noctuoidea (Insecta: lepidoptera). Syst Biol 49:202–224
Miyamoto MM, Fitch WM (1995) Testing the covarion hypothesis of molecular evolution. Mol Biol Evol 12:503–513
Moroz LL, Kocot KM, Citarella MR, Dosung S, Norekian TP, Povolotskaya IS, Grigorenko AP, Dailey C, Berezikov E, Buckley KM, Ptitsyn A, Reshetov D, Mukherjee K, Moroz TP, Bobkova Y, Yu F, Kapitonov VV, Jurka J, Bobkov YV, Swore JJ, Girardo DO, Fodor A, Gusev F, Sanford R, Bruders R, Kittler E, Mills CE, Rast JP, Derelle R, Solovyev VV, Kondrashov FA, Swalla BJ, Sweedler JV, Rogaev EI, Halanych KM, Kohn AB (2014) The ctenophore genome and the evolutionary origins of neural systems. Nature 510: 109–114
Nesnidal MP, Helmkampf M, Bruchhaus I, Hausdorf B (2010) Compositional heterogeneity and phylogenomic inference of metazoan relationships. Mol Biol Evol 27:2095–2104
Nosenko T, Schreiber F, Adamska M, Adamski M, Eitel M, Hammel J, Maldonado M, Müller WEG, Nickel M, Schierwater B, Vacelet J, Wiens M, Wörheide G (2013) Deep metazoan phylogeny: when different genes tell different stories. Mol Phylogenet Evol 67:223–233

Parks SL, Goldman N (2014) Maximum likelihood inference of small trees in the presence of long branches. Syst Biol 63:798–811
Philip GK, Creevey CJ, McInerney JO (2005) The opisthokonta and the ecdysozoa may not be clades: stronger support for the grouping of plant and animal than for animal and fungi and stronger support for the coelomata than ecdysozoa. Mol Biol Evol 22:1175–1184
Philippe H, Brinkmann H, 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
Philippe H, Derelle R, Lopez P, Pick K, Borchiellini C, Boury-Esnault N, Vacelet J, Renard E, Houliston E, Quéinnec E, Da Silva C, Wincker P, Le Guyader H, Leys S, Jackson DJ, Schreiber F, Erpenbeck D, Morgenstern B, Wörheide G, Manuel M (2009) Phylogenomics revives traditional views on deep animal relationships. Curr Biol 19:706–712
Philippe H, Lartillot N, Brinkmann H (2005a) Multigene analyses of bilaterian animals corroborate the monophyly of ecdysozoa, lophotrochozoa, and protostomia. Mol Biol Evol 22:1246–1253
Philippe H, Snell EA, Bapteste E, Lopez P, Holland PWH, Casane D (2004) Phylogenomics of eukaryotes: impact of missing data on large alignments. Mol Biol Evol 21:1740–1752
Philippe H, Zhou Y, Brinkmann H, Rodrigue N, Delsuc F (2005b) Heterotachy and long-branch attraction in phylogenetics. BMC Evol Biol 5:50
Phillips MJ, Penny D (2003) The root of the mammalian tree inferred from whole mitochondrial genomes. Mol Phylogenet Evol 28:171–185
Pisani D (2004) Identifying and removing fast-evolving sites using compatibility analysis: an example from the arthropoda. Syst Biol 53:978–989
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 U S A 112:15402–15407
Pol D, Siddall ME (2001) Biases in maximum likelihood and parsimony: a simulation approach to a 10-taxon case. Cladistics 17:266–281
Pollock DD, Zwickl DJ, McGuire JA, Hillis DM (2002) Increased taxon sampling is advantageous for phylogenetic inference. Syst Biol 51:664–671
Rannala B, Huelsenbeck JP, Yang Z, Nielsen R (1998) Taxon sampling and the accuracy of large phylogenies. Syst Biol 47:702–710
Rivera-Rivera CJ, Montoya-Burgos JI (2016) LS^3: a method for improving phylogenomic inferences when evolutionary rates are heterogeneous among taxa. Mol Biol Evol 33:1625–1634
Rodríguez-Ezpeleta N, Brinkmann H, Roure B, Lartillot N, Lang BF, Philippe H (2007) Detecting and overcoming systematic errors in genome-scale phylogenies. Syst Biol 56:389–399
Rokas A, Abbot P (2009) Harnessing genomics for evolutionary insights. Trends Ecol Evol 24:192–200
Rokas A, Carroll SB (2005) More genes or more taxa? The relative contribution of gene number and taxon number to phylogenetic accuracy. Mol Biol Evol 22:1337–1344
Rokas A, Williams B, King N, Caroll S (2003) Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature 425:798–804
Rosenberg MS, Kumar S (2001) Incomplete taxon sampling is not a problem for phylogenetic inference. Proc Natl Acad Sci U S A 98:10751–10756
Roure B, Baurain D, Philippe H (2013) Impact of missing data on phylogenies inferred from empirical phylogenomic data sets. Mol Biol Evol 30:197–214
Salichos L, Rokas A (2013) Inferring ancient divergences requires genes with strong phylogenetic signals. Nature 497:327–331
Sanderson MJ, McMahon MM, Steel M (2010) Phylogenomics with incomplete taxon coverage: the limits to inference. BMC Evol Biol 10:155
Sanderson MJ, Shaffer HB (2002) Troubleshooting molecular phylogenetic analyses. Annu Rev Ecol Syst 33:49–72
Sanderson MJ, Wojciechowski MF, Hu J-M, Khan TS, Brady SG (2000) Error, bias, and long-branch attraction in data for two chloroplast photosystem genes in seed plants. Mol Biol Evol 17:782–797
Schmidt HA, Strimmer K, Vingron M, von Haeseler A (2002) TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. Bioinformatics 18:502–504
Shen X-X, Salichos L, Rokas A (2016) A genome-scale investigation of how sequence, function, and tree-based gene properties influence phylogenetic inference. Genome Biol Evol 8:2565–2580
Smith SA, Dunn CW (2008) Phyutility: a phyloinformatics tool for trees, alignments and molecular data. Bioinformatics 24:715–716

Spencer M, Susko E, Roger AJ (2005) Likelihood, parsimony, and heterogeneous evolution. Mol Biol Evol 22:1161–1164
Sperling EA, Pisani D, Peterson KJ (2007) Poriferan paraphyly and its implications for Precambrian palaeobiology. Geol Soc Lond Spec Publ 286:355–368
Stamatakis A (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30:1312–1313
Steel MA, Lockhart PJ, Penny D (1993) Confidence in evolutionary trees from biological sequence data. Nature 364:440–442
Strimmer K, von Haeseler A (1997) Likelihood-mapping: a simple method to visualize phylogenetic content of a sequence alignment. Proc Natl Acad Sci U S A 94:6815–6819
Struck TH, Nesnidal MP, Purschke G, Halanych KM (2008) Detecting possibly saturated positions in 18S and 28S sequences and their influence on phylogenetic reconstruction of Annelida (Lophotrochozoa). Mol Phylogenet Evol 48:628–645
Suh A, Smeds L, Ellegren H (2015) The dynamics of incomplete lineage sorting across the ancient adaptive radiation of neoavian birds. PLoS Biol 13:e1002224
Sullivan J, Swofford D, Naylor G (1999) The effect of taxon sampling on estimating rate heterogeneity parameters of maximum-likelihood models. Mol Biol Evol 16:1347
Susko E, Roger AJ (2007) On reduced amino acid alphabets for phylogenetic inference. Mol Biol Evol 24:2139–2150
Tarrío R, Rodríguez-Trelles F, Ayala FJ (2001) Shared nucleotide composition biases among species and their impact on phylogenetic reconstructions of the drosophilidae. Mol Biol Evol 18:1464–1473
Telford MJ, Moroz LL, Halanych KM (2016) Evolution: a sisterly dispute. Nature 529:286–287
Thorley JL, Wilkinson M (1999) Testing the phylogenetic stability of early tetrapods. J Theor Biol 200:343–344
Townsend JP (2007) Profiling phylogenetic informativeness. Syst Biol 56:222–231
Van de Peer Y, Frickey T, Taylor JS, Meyer A (2002) Dealing with saturation at the amino acid level: a case study based on anciently duplicated zebrafish genes. Gene 295:205–211
Wang H-C, Susko E, Roger AJ (2011) Fast statistical tests for detecting heterotachy in protein evolution. Mol Biol Evol 28:2305–2315
Weigert A, Helm C, Meyer M, Nickel B, Arendt D, Hausdorf B, Santos SR, Halanych KM, Purschke G, Bleidorn C, Struck TH (2014) Illuminating the base of the annelid tree using transcriptomics. Mol Biol Evol 31:1391–1401
Whelan NV, Halanych KM (2016) Who let the CAT out of the bag? Accurately dealing with substitutional heterogeneity in phylogenomic analyses. Syst Biol 52:696–704
Whelan NV, Kocot KM, Moroz LL, Halanych KM (2015) Error, signal, and the placement of Ctenophora sister to all other animals. Proc Natl Acad Sci U S A 112:5773–5778
Whelan S, Blackburne BP, Spencer M (2011) Phylogenetic substitution models for detecting heterotachy during plastid evolution. Mol Biol Evol 28:449–458
White W, Hills S, Gaddam R, Holland B, Penny D (2007) Treeness triangles: visualizing the loss of phylogenetic signal. Mol Biol Evol 24:2029–2039
Wiens JJ (1998) Does adding characters with missing data increase or decrease phylogenetic accuracy? Syst Biol 47:625–640
Wiens JJ (2003) Missing data, incomplete taxa, and phylogenetic accuracy. Syst Biol 52:528–538
Wiens JJ, Morrill MC (2011) Missing data in phylogenetic analysis: reconciling results from simulations and empirical data. Syst Biol 60:719–731
Wu J, Susko E (2011) A test for heterotachy using multiple pairs of sequences. Mol Biol Evol 28:1661–1673
Xia X (2013) DAMBE5: A comprehensive software package for data analysis in molecular biology and evolution. Mol Biol Evol 30:1720–1728
Xia X, Xie Z, Salemi M, Chen L, Wang Y (2003) An index of substitution saturation and its application. Mol Phylogenet Evol 26:1–7
Yang Z (1996) Among-site rate variation and its impact on phylogenetic analyses. Trends Ecol Evol 11:367–372
Zwickl DJ, Hillis DM (2002) Increased taxon sampling greatly reduces phylogenetic error. Syst Biol 51:588–598

Rare Genomic Changes

C. Bleidorn, *Phylogenomics*, DOI 10.1007/978-3-319-54064-1_10

- Several different marker systems employing genome-level character have been used to find additional support for phylogenetic hypothesis.
- Genome-level characters include absence/presence patterns of mobile elements, microRNAs and introns, as well as gene order rearrangement and changes in the genetic code.
- Retrotransposon integrations spread by a copy-and-paste mechanism through the genome and are close to a perfect phylogenetic marker for shallow phylogenies (divergences of <65 mya).
- Absence/presence of microRNAs can be used to resolve deep phylogenies, but frequent convergent loss makes analyses difficult.
- Several mechanisms (inversion, transposition, tandem duplication random loss, translocation, fusion, fission) can result in the rearrangement of gene order.
- Maximum parsimony variants can be used to analyse absence/presence matrices to reconstruct phylogenetic trees.

10.1 The Perfect Phylogenetic Marker

The ideal phylogenetic marker is a character that, after it has evolved, will not be lost again, and homology can unambiguously be assigned across taxa due to its conservation. DNA or amino acid sequences are far from being perfect markers, and many problems can arise in phylogenetic analyses (Jeffroy et al. 2006). As an alternative, genome-level characters became popular to complement existing phylogenetic analyses and to test hypotheses with an independent set of characters (Rokas and Holland 2000). Possible phylogenetic markers are integrations of mobile elements, absence/presence (a/p) of microRNAs or introns, gene order rearrangements or changes in the genetic code. A big difference between these kinds of markers in comparison to analysing sequence data is how they are expected to change over time. For sequence data usually a clocklike change is assumed with the expectation that over time the numbers of changes accumulate linearly, even though the pace of change might be different in different lineages. In contrast, genome-level characters are expected to change non-clocklike in a saltatory way (Boore 2006). This makes analysing rare genomic change data tricky, as evolutionary models are more difficult to apply. However, if the changes are indeed rare, the presence of such characters might be an additional strong support for the monophyly of its bearers, which could be especially interesting for clades that are difficult to resolve by sequence data alone. For example, molecular systematic analyses based on a single or few genes consistently recovered a monophyletic group including crustacean and insect taxa (Pancrustacea) (Friedrich and Tautz 1995). This result was controversial, as it contradicted the former textbook knowledge which united insects with myriapods (Tracheata). A single translocation of a tRNA in the rather conserved arthropod mitochondrial genome, which was only found in analysed insects and crustaceans, gave additional strong support for the Pancrustacea hypothesis (Boore et al. 1998), which is now generally accepted. Another famous example is the analysis of the presence of some mobile elements (SINEs) in specific positions in the genome, which supported the monophyly of whales, ruminants and hippopotamuses (Shimamura et al. 1997). These promising results spurned the search for rare genomic changes to resolve difficult phylogenetic questions, but also led to the question how to analyse these markers and how to weigh their support.

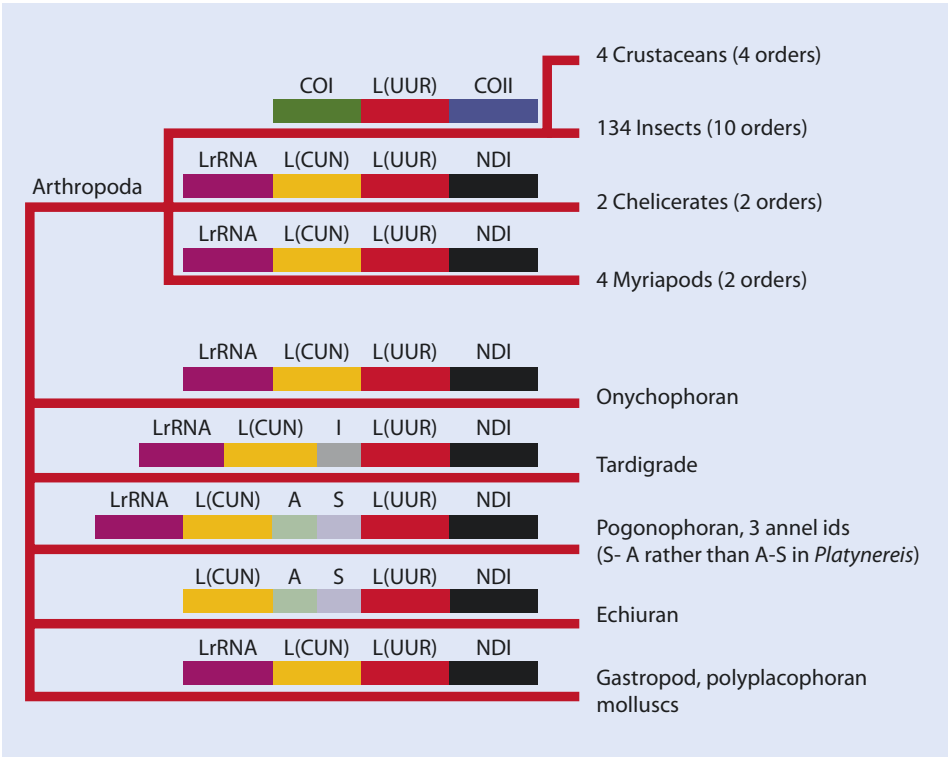


Fig. 10.1 Arthropod relationships deduced from rare genomic changes. Insects and crustaceans are united by a single tRNA (trnL) translocation within the mitochondrial genome, which is found to be (syn) apomorphic by outgroup comparison (Reprinted by permission from Macmillan Publishers Ltd.: Nature (Boore et al. 1995), copyright 1998)

Analysing genome-level characters is in the most cases based on matrices with a/p character states for the investigated taxa. This matrix can then be analysed with maximum parsimony (MP) or other approaches. As such, the analysis of genome-level characters is similar to that of morphological data. If a tree is known, characters and their states can be mapped on the phylogeny to distinguish plesiomorphic and apomorphic character states (Hennig 1965). This distinction goes back to the work of the German entomologist Willi Hennig and is seen as the foundation of the cladistic method and brought important changes of how to address phylogenetic systematics in general (Richter and Meier 1994). The plesiomorphic character state is the ancestral state present in a taxon and retained from its ancestor. For example, in Fig. 10.1 the position of the trnL(UUR) between the genes trnL(CUN) and ND1 represents a plesiomorphy, as supported by the tree and outgroup comparison. In contrast, apomorphic character states are derived states. Only these characters can be used to support the monophyly of a group of taxa. Additionally, autapomorphies (apomorphic character states found in a single lineage) and synapomorphies (apomorphic character states supporting the monophyly of a group of taxa) can be distinguished. For example, in Fig. 10.1 the position of the trnL(UUR) between the genes COI and COII is interpreted as a synapomorphy for a clade uniting insects and crustaceans. It is important to keep in mind that these terms are relative, related to where in the

phylogenetic tree they are applied. Whereas the character state for the position of trnL(UUR) is apomorphic for Pancrustacea, it is the plesiomorphic state if we would apply it to describe the same character state within the insect phylogenetic tree. In most cases plesiomorphy or apomorphy can only be assigned after (a posteriori) the phylogenetic analysis, as the characters (and the direction of their evolution) are polarized by using the resulting phylogeny. However, the presence state of some genome-level characters (e.g. SINE insertions, see below) is quasi used a priori as the apomorphic character state, as absence is unlikely the derived state (Shedlock and Okada 2000).

10.2 Mobile Elements

Retrotransposons are mobile elements that have the ability to integrate into the genome at a new site within their cell of origin (Kazazian 2004). In contrast to DNA transposons, which use a cut-and-paste mechanism of copying, retrotransposons use a copy-and-paste mechanism to integrate at new sites. This is achieved by transcription of the retrotransposons into RNA, which are then reverse transcribed and reintegrated into the genome, thereby duplicating the element. Two major classes of retrotransposons are recognized: LTR retrotransposons, which either contain long terminal repeats (LTR) at both ends, or non-LTR retrotransposons (or LINEs) which lack LTRs and possess a polyadenylate sequence at their 3'termini (Kazazian 2004). Unlike LTR retrotransposons that generate uniform target site duplications and require the presence of their terminal repeats for integration, non-LTR element copies are often truncated at their 5'ends (Malik and Eickbush 1998). Short interspersed nuclear elements (SINEs) are mobile elements that originated from the accidental retrotransposition of small RNA polymerase III transcripts, such as 7SL RNAs, tRNAs or 5S RNA. Therefore SINEs always feature an internal RNA polymerase III promoter at their 5′ end for their transcription. SINEs are nonautonomous, and to be replicated, they completely rely on the machinery of the cell and the activity of autonomous retrotransposons, such as LINEs (Kramerov and Vassetzky 2005). Some SINEs are known to occur in huge copy numbers in their «host» genome, as, for example, the primate specific *Alu* SINE family, which makes up around 11% of the human genome (Deininger 2011).

The copy-and-paste mechanism makes retrotransposons an almost perfect phylogenetic marker. They are regarded as being nearly homoplasy-free, as convergent integrations at the exact same genomic positions are highly unlikely (Ray et al. 2006), even though some few examples of apparent homoplasy exists (Han et al. 2011). And they are basically polarized characters, such that the absence of a retrotransposon at a given locus is usually the ancestral state (Ray et al. 2006). The caveat is that these markers are only suited to resolve relatively young divergences (50 mya and younger), as otherwise the homology between integrated sequences is difficult to detect as mutations are accumulated over time (Shedlock and Okada 2000). Retrotransposons have been successfully used to address population diversity in plants (Kalendar et al. 2011) or the phylogeny of birds (Suh et al. 2011) or mammals (Kriegs et al. 2006), but there are no examples to use them to infer deeper phylogenies.

Retrotransposon activity in the genome of their host varies over evolutionary timescales. Different groups of retrotransposons may have different (and also overlapping) times of activities before they get inactive. Activity of retrotransposons can be triggered by mutations within inactive sequences (e.g. acquisition of a new promoter), but also due

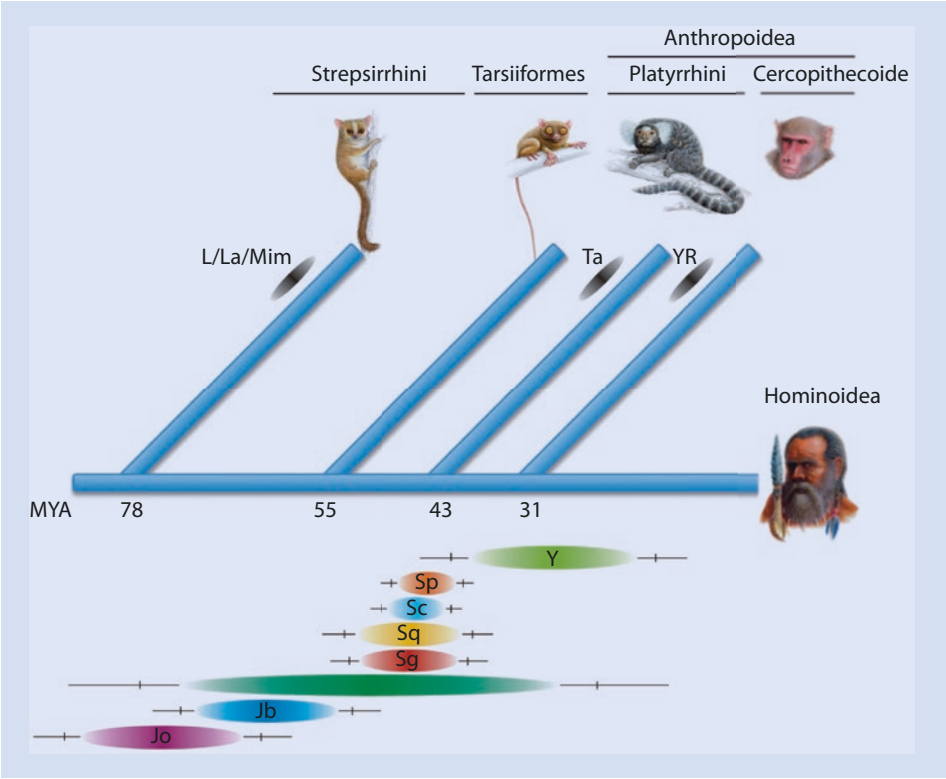


Fig. 10.2 Activity patterns of different groups of SINEs (*Jo, Jb, Sx, Sg, Sc, Sp, Y*) along primate evolution as modelled by the TinT method (Reprinted from Churakov et al. (2010))

to horizontal gene transfer into a new host (Huang et al. 2012). Inactivity or death of a group of retrotransposons occurs when the last active copy loses its activity due to mutation. Groups of retrotransposons can be classified by its sequence similarity, e.g. using the software REPEATMASKER (Tarailo-Graovac and Chen 2009), and some methods exist to trace their activity over time (Kriegs et al. 2007; Giordano et al. 2007). The «Transposition in Transposition» (TinT) method is based on the idea that evolutionary younger actively transposed elements are able to insert into older elements, but the opposite is not possible. Within a probabilistic framework, information of the occurrence of nested retrotransposon insertion patterns is used to model the timing of element activity (Churakov et al. 2010). Absolute timescales of the relative chronological order can be inferred by mapping these activity patterns on a dated phylogeny (Fig. 10.2), which also highlights that retrotransposons are often only informative for a short window of the evolutionary timescale. For example, analyzing *Jb* SINE patterns would not be informative to investigate ape (Hominoidea) evolution, as its activity window predates the origin of this clade (Fig. 10.2).

For phylogenetic analyses, a/p of retrotransposons is scored for each homologous integration within a character matrix. Especially integrations of retrotransposons within intron regions are suitable for analysis, as due to the conserved nature of the adjacent exons these genomic regions are easier to orthologize. Besides the orthology of the genomic region, homology of the retrotransposons must be carefully considered, which is

complicated due to random mutational decay over time. It is not unusual to manually inspect every single alignment of orthologous genes to verify the homology of retrotransposons (Suh et al. 2015). There are two strategies for the phylogenetic analysis of retrotransposon data. First, a/p matrices can be analysed directly using MP for the inference of a tree, e.g. Kaiser et al. (2007). However, reflecting their activity sometimes these markers are only informative for certain windows of evolutionary time (see above), and several parts of the tree remain unresolved. Alternatively, presence of shared retrotransposon integrations is mapped onto a tree topology and congruence with a/p patterns can be used to favour one of several competing hypotheses or to give additional support for the monophyly of selected groups in a tree. For example, Kriegs et al. (2006) analysed retrotransposon integrations across mammals and mapped their data on existing trees. A statistic framework for evaluating support from retrotransposons has been proposed by Waddell et al. (2001). Their likelihood-based test statistics show that at least five unambiguous markers (five retrotransposon integrations supporting a given clade, with no other integration in conflict) are required for a certain node to gain significant p-values.

Even though retrotransposon markers usually show only very low levels of homoplasy from convergent integration at the exactly same site, several examples of conflicting nodes have been found when addressing the phylogeny of fast radiations. For example, Nishihara et al. (2009) investigated retrotransposon integrations of placental mammals and found nearly the same number of loci (21–25 loci) supporting three different hypotheses. Similarly, Suh et al. (2015) investigated the radiation of birds based on thousands of carefully selected retrotransposons and found that a third of these are supporting conflicting hypotheses. In both analyses the conflicting retrotransposons map to parts of the phylogeny which are characterized by short internodes. Consequently, the conflict within this retrotransposon data is not interpreted as convergence due to parallel integrations, but as a persistence of ancestral polymorphisms, a phenomenon known as incomplete lineage sorting (ILS). The affected regions of the tree have been found as notoriously difficult to reconstruct, even in the light of massive datasets. High amounts of ILS in combination with short internodes point to a nearly simultaneous divergence of deep lineages within mammals and birds, which might be unresolvable into a bifurcating tree. Instead, a phylogenetic network illustrating the conflict within this part of the tree seems to be a better representation of the phylogenetic relationships of these groups (Suh et al. 2015; Hallström and Janke 2010). Not surprisingly, high amounts of ILS based on retrotransposon data has been also reported for Lake Tanganyika cichlids (Takahashi et al. 2001), the posterchild for adaptive radiations.

Mobile elements in general have been firstly described in plants (McClintock 1950), a discovery which later was honoured with the Nobel Prize in physiology or medicine in 1983 for Barbara McClintock. And even though different classes of mobile elements are extremely abundant in plant genomes, most studies exploiting these elements as phylogenetic markers are from vertebrate animals. Kalendar et al. (2011) summarized the use of mobile elements as markers in plant phylogeny and evolution, with most studies addressing the population level. Yaakov et al. (2012) used so-called miniature inverted-repeat transposable elements (MITEs) to investigate wheat biodiversity and evolution. MITEs are small mobile elements of up to a few hundred base pairs in size, flanked by tandemly inverted repeats (Wicker et al. 2007). Similarly to SINEs, they are nonautonomous and can occur in high copy numbers. MITEs were first discovered in plants (Wessler et al. 1995), but are also abundantly found in many eukaryotic genomes, including humans (Morgan 1995). The study by Yaakov et al. (2012) analysed a/p matrices of MITE-polymorphism

across wheat accessions and found this marker system phylogenetically informative. Similar positive results were reported for the analyses of barley biodiversity using MITEs (Lyons et al. 2008).

10.3 MicroRNAs

MicroRNAs are short non-coding RNAs involved in the regulation of gene expression. They are found in plants and animals, but seem to have evolved convergently in these lineages (Shabalina and Koonin 2008). Moreover, several viruses have been identified harbouring microRNA sequences, with most examples stemming from herpesviruses (Skalsky and Cullen 2010). Premature microRNAs form hairpin-like secondary structures (Fig. 10.3). This stem-loop precursor is further processed into double-stranded RNA of approximately 22 bp (Kim 2005). Nucleotides 2–7 of the mature microRNA 5′-end are called seeds and play an important role in mRNA-target recognition (Liu et al. 2008). Often more than a hundred targets can be recognized by a single microRNA, and miRNAs complementarily bind to target mRNAs, where they repress translation and/or induce mRNA degradation (Lewis et al. 2003). Mature microRNAs were shown to be highly conserved across animal taxa (Sempere et al. 2006), and several hundred distinct microRNA families have been reported for Metazoa (Kozomara and Griffiths-Jones 2011).

Interestingly, it has been found that microRNA families are continuously emerging and expanding through animal evolution (Hertel et al. 2006); however, once evolved microRNAs were regarded to be rarely lost secondarily (Sempere et al. 2006). The expression of many microRNAs is known to be tissue specific (Clark et al. 2010; Christodoulou et al. 2010), and additionally, the disparity of microRNAs of a given animal taxon can often be linked to its morphological complexity (e.g. number of cell types) (Heimberg et al. 2008; Sempere et al. 2006; Kosik 2009). Given these properties, the potential of microRNAs as a powerful phylogenetic marker system is obvious and they were used in several phylogenetic analyses (Heimberg et al. 2010; Sperling et al. 2011; Rota-Stabelli et al. 2011; Sperling et al. 2009; Campbell et al. 2011; Helm et al. 2012).

As in the case of retrotransposons, microRNAs are coded as a/p in a matrix and can be analysed directly (e.g. using MP) or by mapping onto an existing phylogenetic tree. The advantage of microRNAs over retrotransposons is that they have shown to be

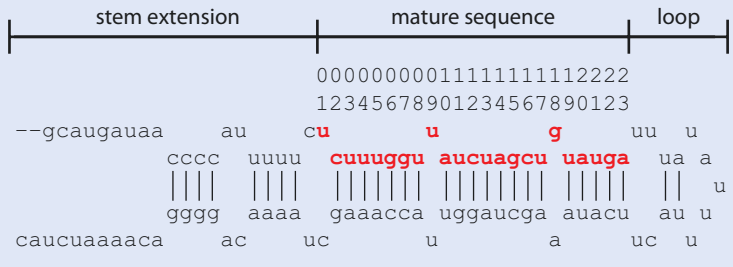


Fig. 10.3 Features of premature microRNA secondary structure exemplified by *mir-9* from the annelid *Capitella teleta* as found in miRBase (Kozomara and Griffiths-Jones 2014) accession MI0010052. The mature sequence is indicated by red bases. Positions 2–8 of the mature sequence are also known as seed and play an important role in mRNA-target recognition

phylogenetically informative over deep evolutionary timescales (Tarver et al. 2013). Even though mature microRNAs are represented by very small sequences (~22 bp), they remain remarkably conserved over time. However, in contrast to analyses of retrotransposon data, phylogenetic analyses of microRNAs seem to be more problematic. Based on the presence of a single or few microRNAs, Acoela are supported as a deuterostome in-group taxon (Philippe et al. 2011), Mandibulata as monophyletic (Rota-Stabelli et al. 2011) and Tardigrada as sister group of Onychophora + Arthropoda (Campbell et al. 2011). However, these results might be biased due to highly heterogeneous rates of microRNA gain and loss, as well as sampling error (Thomson et al. 2014). For example, there is evidence that microRNAs get lost due to loss of their function. For example, *mir-10* is a phylogenetically conserved microRNA present in most bilaterian lineages regulating a subset of *hox* genes (Pearson et al. 2005). Interestingly, this microRNA is directly located within the *hox* cluster and has been convergently lost in lineages with a disintegrated *hox* cluster, such as nematodes and tunicates (Tanzer et al. 2005). Major loss of microRNAs is also reported for tunicates (Fu et al. 2008), where at least 11 families of bilaterian microRNAs are missing. An analysis of chordate phylogeny also revealed several losses of microRNA families in different lineages (Heimberg et al. 2010). Frequent gain and loss on a short evolutionary timescale has further been demonstrated for *Drosophila* (Nozawa et al. 2010). Moreover, it remains a practical problem that the absence of microRNAs can be only safely concluded with available complete genome sequences. Instead of MP analyses or mapping, Thomson et al. (2014) explored the performance of microRNA a/p matrices under different evolutionary models. Their re-analyses casted doubt on the results of several published phylogenetic studies, and they conclude that the potential of microRNA data to resolve the (animal) tree of life has been overstated.

10.4 Introns

In eukaryotes genes are interrupted by spliceosomal introns which are removed from transcripts prior to their translation (Jeffares et al. 2006). The absolute number of introns within a genome and the number of introns within a gene are highly variable. However, intron positions of most introns are conserved across eukaryotes, and variation in intron numbers is explained by either intron gain or loss (Rogozin et al. 2003). Possible sources for the generation of new genomic introns are DNA transposons (Huff et al. 2016). Generally, intron gain seems to occur less frequently; however, increased intron gain and decreased intron loss was observed in evolutionarily conserved genes (Carmel et al. 2007). Moreover, intron loss is regarded to be nearly irreversible (Roy and Gilbert 2005a), as intron gain at exactly the same site happens only rarely (Sverdlov et al. 2005). Therefore, shared intron positions should indicate homology (Roy 2016), and analysing a/p patterns of intron positions has been proposed as a useful phylogenetic marker of deep divergences (Rokas and Holland 2000). For example, introns have been used to analyse relationships of deep divergences within Metazoa (Roy and Gilbert 2005b) or ray-finned fishes (Venkatesh et al. 1999).

A straightforward way to analyse a/p data of introns is based on MP. In its general form, MP gives all character transformations the same probability. However, as intron loss is thought to be nearly irreversible, this assumption might be violated. To circumvent this problem, the use of a special form of MP called Dollo parsimony can be used to analyse such datasets and is also applicable for microRNA data. Dollo's Law states that complex

characters cannot be «re-evolved» once they got lost, and instead alternative ways lead to convergent solutions (which should be detectable as such) (Dollo 1893). This century-old idea has to be treated carefully given the actual knowledge of the genetic and developmental bases of complex morphological characters (Hall 2003), but might be a fitting description for what we know about the evolution of microRNAs and introns. Dollo parsimony was introduced by Farris (1977). For analysis, characters are polarized a priori, and the presence of the complex character state is coded as 1, whereas the absent, likely ancestral state is coded by 0. Using this algorithm, only one change from 0 to 1 is allowed during the analysis, whereas as many reversions from 1 to 0 as necessary to explain the observed data are possible. By applying Dollo parsimony for intron a/p matrices, the number of (parallel) intron gains is minimized, whereas losses can be frequent. As the rates of intron loss can vary dramatically across taxa (Jeffares et al. 2006), Zheng et al. (2007) introduced a modified Dollo parsimony algorithm that uses different weights for the cost of an intron loss in different branches. Alternatively, explicit phylogenetic models have been developed to analyse large matrices of intron a/p data across species. In this case, based on different tree topologies (hypotheses), different expectations regarding ratios of intron gain and lost are formulated and compared with the data (Roy and Gilbert 2005b). Both types of analyses implicitly assume constant rates of intron loss and violations of this assumption might lead to long-branch attraction (Irimia and Roy 2008). Especially the frequent occurrence of multiple independent losses of the same intron in distantly related species is problematic for phylogenetic analyses and questions the usefulness of this marker system in general (Krzywinski and Besansky 2002; Kiontke et al. 2004).

An approach to limit the impact of convergent intron gains or losses is the analysis of near intron pairs (NIPs) (Krauss et al. 2008). Such NIPs include two intron positions in an alignment of orthologous genes that are separated by a small number of nucleotides. It is known that exons smaller than ~50 bp are only rarely found, which could be related to problems of splicing such small sequences (Irimia and Roy 2008). Therefore, introns found at nearby positions are unlikely to have coexisted, and given that multiple gains at exactly the same site are very rare, this data can be used to infer a phylogenetic tree. Krauss et al. (2008) proofed that this method is in principle useful for phylogenetic reconstruction. Lehmann et al. (2013) used NIPs to infer metazoan phylogeny and found them to clearly outperform Dollo parsimony analyses based on all introns. However, as the number of suitable characters is strongly reduced by this approach, parts of the tree which correspond to taxa or time periods with low levels of intron gain are difficult to resolve.

10.5 Gene Order

The order of genes in the genome has been extensively used as a phylogenetic marker (Boore 2006; Sankoff et al. 1992). The first use of gene order to infer evolutionary relationships goes back to Sturtevant and Dobzhansky (1936), who analysed inversions located in a chromosome to study the evolution of some drosophilids. Most studies using gene order as phylogenetic markers are based on organellar genomes, as in the case of plant chloroplasts (Downie and Palmer 1992; Cosner et al. 2004) or animal mitochondrial genomes (Boore and Brown 1998; Bleidorn et al. 2007). For example, animal mitochondrial genomes are usually circular molecules that harbour around 37 genes. Every gene can be either transcribed from the plus or the minus strand, and several mechanisms have been described how gene order can be rearranged (Boore 1999). Due to its small size, many

animal mitochondrial genomes have been sequenced already with the Sanger technique. Using next-generation sequencing techniques, complete animal mitochondria can now be reconstructed fast and easily from shallowly sequenced whole genome shotgun libraries, an approach which is known as genome skimming (Richter et al. 2015). Similarly, complete chloroplast genome have been reconstructed using this approach (Malé et al. 2014). Not surprisingly, many different mitochondrial gene orders are observed (e.g. Fig. 10.4), and the possibility of convergent changes resulting in the same order is rather low (Dowton et al. 2002), even though some examples are known (Shao and Barker 2003).

Several types of rearrangements are defined based on the comparison of closely related species with different gene orders of unichromosomal genomes: inversions (Fig. 10.5a), transpositions (Fig. 10.5b), inverse transpositions (transpositions where the re-inserted fragment is inverted) (Fig. 10.5c) and tandem duplications followed by random loss (TDRL) of one of the gene copies (Fig. 10.5d) (Bernt et al. 2013). A web-based application called CREX (Bernt et al. 2007) is available, which based on common intervals finds parsimonious scenarios for the rearrangement of a pair of gene orders. More complicated are cases where more than one chromosome exists, as, for example, in most eukaryotic

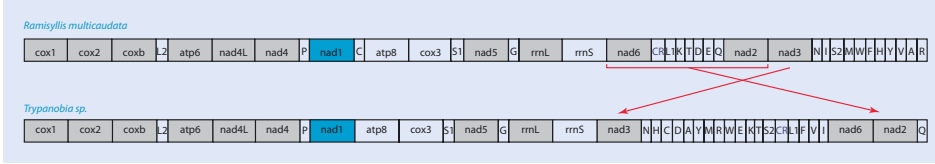


Fig. 10.4 Comparison of the mitochondrial gene order of two closely related annelids (Syllidae). All genes are transcribed from the same strand. Changes are indicated by *arrows*. Mitochondrial protein coding and ribosomal genes are abbreviated with 3–4 letters, tRNA genes are given in the one-letter code (Reprinted from Aguado et al. (2015))

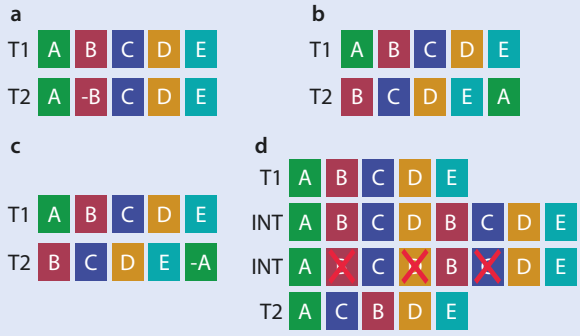
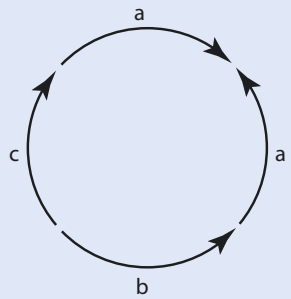


Fig. 10.5 Different types of rearrangements hypothesized for animal mitochondrial genomes. Examples of gene order of five genes (A-E) in two taxa (T1 and T2). Genes on the minus strand are indicated with a minus sign (–). **a** Inversion of a single gene (B to –B). **b** Transposition of a single gene (A). **c** Inverse transposition of a single gene (A to –A). **d** Tandem duplication random loss (TDRL) scenario. Change of the gene order from T1 to T2 is shown with intermediate states (INT), where three genes (B, C, D) are tandemly duplicated and one copy of each gene got lost. Sometimes different scenarios (e.g. transposition and TDRL) are equally likely to explain changes in the gene order

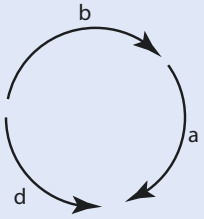
nuclear genomes. In this case, translocations, fusions and fissions are additional possible scenarios (Hu et al. 2014). A translocation describes the break of a chromosome, with one part attaching to another chromosome. A fusion joins two chromosomes, while a fission breaks a single chromosome into two parts. Analysing gene orders is further complicated by the deletion and duplication of genes, and in the latter case even whole genome duplications are not unusual.

Different methods have been proposed to analyse gene order data. Genes can be coded on different strands of a double-stranded DNA molecule, thereby allowing four different types of adjacency (Moret et al. 2013) of two genes following each other: (a^t, b^t), (a^h, b^t), (a^t, b^h) and (a^h, b^h). In this examples, *a* and *b* denote different genes, whereas ^{h} (head) and ^{t} (tail) refer to their orientation to each other. For example, when the two genes *b* and *a* are described on the plus strand one after another, their orientation would be (b^h, a^t) (Fig. 10.6a, genome 2). The easiest way to compare the order of two genomes is to estimate the number of breakpoints (Blanchette et al. 1997). If two genes *a* and *b* are adjacent in taxon 1 but not in taxon 2, they determine a breakpoint. The number of breakpoints between two unichromosomal genomes represents the most general measure of gene order distance, as it requires no assumptions about the mechanisms of gene order evolution (e.g. differences between inversions and transpositions). Further distance measures are the inversion distance and the double-cut-and-join (DCJ) distance. The inversion distance equals the minimum number of inversions to transform one unichromosomal gene order into another one, given the same gene content and absence of duplications (Hannenhalli and Pevzner 1999). The (DCJ) distance is a model that accounts for most events altering gene order, such as inversions, translocations, fusions and fissions

a



Genome 1



Genome 2

b

	adjacency information						content information			
	$\{a^h, a^h\}$	$\{a^t, b^h\}$	$\{a^t, c^h\}$	$\{b^t, c^t\}$	$\{a^h, d^h\}$	$\{b^t, d^t\}$	a	b	c	d
Genome 1	1	1	1	1	0	0	1	1	1	0
Genome 2	0	1	0	0	1	1	1	1	0	1

Fig. 10.6 Two circular example genomes coded for gene order analysis. **a** Genome 1 and genome 2, showing the order of genes (a–d). *Arrows* show the orientation of the gene. **b** Coding of the two genomes into a matrix with adjacency and content information. The orientation of genes is given as head to head (e.g. a^h, a^h), or tail to head (e.g. a^t, b^h), and so on. Presence of an adjacency is coded as 1, absence as 0. Similarly, gene content information is coded as absent (1) or present (0) (Reprinted from Lin et al. (2012a))

(Yancopoulos et al. 2005). Calculations of different pairwise distances between genomes can be conducted with the software UNIMOG (Hilker et al. 2012). Whereas most distance measures were developed for unichromosomal genomes with the exact same gene content, alternatives are available for mutichromosomal genomes and/or when gene duplications or deletions occurred (Moret et al. 2013). Instead of using distance measures, gene order can be also coded into a matrix, where each observed combination of adjacent genes (and their orientation) represents a character, and the absence or presence of the adjacency of these genes is coded for each genome (Fig. 10.6). Phylogenetic analyses of gene order data can be either conducted using distance-based methods such as neighbour joining or by analysing encoded gene order matrices using an optimality criterion such as MP or maximum likelihood (ML) (Moret et al. 2013). Several programs specifically for the phylogenetic analysis of gene order data have been published, e.g. GRAPPA (Moret et al. 2001), MGR (Bourque and Pevzner 2002), MLGO (Hu et al. 2014) or TIBA (Lin et al. 2012b). Lin et al. (2012a) developed a likelihood approach where gene order and content are coded into a matrix (see Fig. 10.6) and transition probabilities between character states are estimated from this matrix, which can then be used for ML analyses, e.g. by using the program RAXML (Stamatakis 2014). Matrices based on pairwise distances derived from gene order can be further analysed using neighbour joining, e.g. as implemented in MEGA (Kumar et al. 2016).

10.6 Changes in the Genetic Code

After the structure of the DNA double helix was discovered in 1953, it took more than a decade to completely decipher its code (Cobb 2015). Based on this code, nucleotide triplets are translated into amino acids. When discovered, it was surprising that the code was highly degenerated, as most of the 20 amino acids were represented by more than one triplet. Initially, it was considered that the genetic code is truly universal and not evolvable, meaning that the pattern of degeneracy could represent a «frozen accident» (Crick 1968). However, after the discovery that human nuclear and mitochondrial genes use different codes (Barrell et al. 1979), it became obvious that the code is indeed evolvable. Later on, many exceptions from the standard genetic code have been described, with most of them found in mitochondrial genomes (Knight et al. 2001). Different models have been proposed how codons can be reassigned. Based on the codon-capture model it is hypothesized that a codon first disappears from the coding sequences of the genome, resulting into loss of function of this specific codon. In case it reappears due to nucleotide substitutions in any coding sequence, a reassignment to a new tRNA is possible (Osawa and Jukes 1989). Alternatively, a codon might be translated ambiguously, and one of its variant becomes fixed (Schultz and Yarus 1994). There are examples available for both models, which may just represent differences in the timing if reassigned codons appear after or before the loss of the old codon (Sengupta et al. 2007).

Given that changes in the genetic code are rare events, they bear the potential to be used as a phylogenetic marker. By comparatively analysing mitochondrial genomes, Castresana et al. (1998) found support for the monophyly of a group uniting enteropneusts and echinoderms based on predicted changes of the genetic code. Similarly, Telford et al. (2000) used a change in the genetic code as further support for the monophyly of the flatworm taxon Rhabditophora, whereas Keeling and Doolittle (1997) used such data to

evaluate different hypotheses regarding the phylogenetic position of the taxon *Girardia* within diplomonads.

As the loss of a codon triplet sequence in the coding part of the genome is an important step towards codon reassignment, it comes without surprise that most code changes have been reported from the rather small animal mitochondrial genomes. Abascal et al. (2012) screened more than 300 arthropod mitochondrial genomes and found that ~20% do not bear the codon AGG. Interestingly, in nearly half of the investigated species, this codon is translated into lysine, whereas the other half shows a translation into leucine. When mapping these changes onto an arthropod phylogeny, it became clear that a reassignment of this codon occurred frequently within this group, exhibiting high levels of convergence, thereby diminishing its usefulness as a phylogenetic character. The same authors also published a software called GENDECODER which can be used to automatically scan genomes for the presence of reassigned codons (Abascal et al. 2006). This method is based on the idea that if the appearance of a particular codon in an investigated species is linked to an alignment position for which a specific amino acid is conserved in a set of reference species, the same translation is assumed for the query. Alternatively, the software FACIL uses hidden Markov models to predict genetic codes by comparison with a reference database (Dutilh et al. 2011).

References

Abascal F, Posada D, Zardoya R (2012) The evolution of the mitochondrial genetic code in arthropods revisited. Mitochondr DNA 23:84–91

Abascal F, Zardoya R, Posada D (2006) GenDecoder: genetic code prediction for metazoan mitochondria. Nucleic Acids Res 34:W389–W393

Aguado MT, Glasby CJ, Schroeder PC, Weigert A, Bleidorn C (2015) The making of a branching annelid: an analysis of complete mitochondrial genome and ribosomal data of *Ramisyllis multicaudata*. Sci Rep 5:12072

Barrell BG, Bankier AT, Drouin J (1979) A different genetic code in human mitochondria. Nature 282:189–194

Bernt M, Braband A, Schierwater B, Stadler PF (2013) Genetic aspects of mitochondrial genome evolution. Mol Phylogenet Evol 69:328–338

Bernt M, Merkle D, Ramsch K, Fritzsch G, Perseke M, Bernhard D, Schlegel M, Stadler PF, Middendorf M (2007) CREx: inferring genomic rearrangements based on common intervals. Bioinformatics 23:2957–2958

Blanchette M, Bourque G, Sankoff D (1997) Breakpoint phylogenies. Genome Inform Ser Workshop Genome Inform 8:25–34

Bleidorn C, Eeckhaut I, Podsiadlowski L, Schult N, McHugh D, Halanych KM, Milinkovitch MC, Tiedemann R (2007) Mitochondrial genome and nuclear sequence data support Myzostomida as part of the annelid radiation. Mol Biol Evol 24:1690–1701

Boore JL (1999) Animal mitochondrial genomes. Nucleic Acids Res 27:1767–1780

Boore JL (2006) The use of genome-level characters for phylogenetic reconstruction. Trends Ecol Evol 21:439–446

Boore JL, Brown WM (1998) Big trees from little genomes: mitochondrial gene order as a phylogenetic tool. Curr Opin Genet Dev 8:668–674

Boore JL, Collins T, Stanton D, Daehler L, Brown WM (1995) Deducing the pattern of arthropod phylogeny from mitochondrial DNA rearrangements. Nature 376:163–165

Boore JL, Lavrov DV, Brown WM (1998) Gene translocation links insects and crustaceans. Nature 392:667–668

Bourque G, Pevzner PA (2002) Genome-scale evolution: reconstructing gene orders in the ancestral species. Genome Res 12:26–36

Campbell LI, Rota-Stabelli O, Edgecombe GD, Marchioro T, Longhorn SJ, Telford MJ, Philippe H, Rebecchi L, Peterson KJ, Pisani D (2011) MicroRNAs and phylogenomics resolve the relationships of Tardigrada and suggest that velvet worms are the sister group of Arthropoda. Proc Natl Acad Sci U S A 108:15920–15924

Carmel L, Rogozin IB, Wolf YI, Koonin EV (2007) Evolutionarily conserved genes preferentially accumulate introns. Genome Res 17:1045–1050

Castresana J, Feldmaier-Fuchs G, S-i Y, Satoh N, Pääbo S (1998) The mitochondrial genome of the hemichordate *Balanoglossus carnosus* and the evolution of deuterostome mitochondria. Genetics 150:1115–1123

Christodoulou F, Raible F, Tomer R, Simakov O, Trachana K, Klaus S, Snyman H, Hannon GJ, Bork P, Arendt D (2010) Ancient animal microRNAs and the evolution of tissue identity. Nature 463:1084–1088

Churakov G, Grundmann N, Kuritzin A, Brosius J, Makałowski W, Schmitz J (2010) A novel web-based TinT application and the chronology of the Primate Alu retroposon activity. BMC Evol Biol 10:376

Clark AM, Goldstein LD, Tevlin M, Tavare S, Shaham S, Miska EA (2010) The microRNA miR-124 controls gene expression in the sensory nervous system of *Caenorhabditis elegans*. Nucleic Acids Res 38:3780–3793

Cobb M (2015) Life's greatest secret: the race to crack the genetic code. Basic Books, London

Cosner ME, Raubeson LA, Jansen RK (2004) Chloroplast DNA rearrangements in Campanulaceae: phylogenetic utility of highly rearranged genomes. BMC Evol Biol 4:27

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

Deininger P (2011) Alu elements: know the SINEs. Genome Biol 12:236

Dollo L (1893) Les lois de l' evolution. Bull Belg Soc Geol, Palaeontol Hydrol 8:164–166

Downie SR, Palmer JD (1992) Use of chloroplast DNA rearrangements in reconstructing plant phylogeny. In: Soltis PS, Soltis DE, Doyle JJ (eds) Molecular systematics of plants. Springer US, Boston, pp 14–35

Dowton M, Castro LR, Austin AD (2002) Mitochondrial gene rearrangements as phylogenetic characters in the invertebrates: the examination of genome 'morphology'. Invertebr Syst 16:345–356

Dutilh BE, Jurgelenaite R, Szklarczyk R, van Hijum SAFT, Harhangi HR, Schmid M, de Wild B, Françoijs KJ, Stunnenberg HG, Strous M, Jetten MSM, Op den Camp HJM, Huynen MA (2011) FACIL: fast and accurate genetic code inference and logo. Bioinformatics 27:1929–1933

Farris JS (1977) Phylogenetic analysis under Dollo's law. Syst Zool 26:77–88

Friedrich M, Tautz D (1995) Ribosomal DNA phylogeny of the major extant arthropod classes and the evolution of myriapods. Nature 376:165–167

Fu XH, Adamski M, Thompson EM (2008) Altered miRNA repertoire in the simplified chordate, *Oikopleura dioica*. Mol Biol Evol 25:1067–1080

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:e137

Hall BK (2003) Descent with modification: the unity underlying homology and homoplasy as seen through an analysis of development and evolution. Biol Rev 78:409–433

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

Han K-L, Braun EL, Kimball RT, Reddy S, Bowie RCK, Braun MJ, Chojnowski JL, Hackett SJ, Harshman J, Huddleston CJ, Marks BD, Miglia KJ, Moore WS, Sheldon FH, Steadman DW, Witt CC, Yuri T (2011) Are transposable element insertions homoplasy free?: An examination using the Avian tree of life. Syst Biol 60:375–386

Hannenhalli S, Pevzner PA (1999) Transforming cabbage into turnip: polynomial algorithm for sorting signed permutations by reversals. J ACM 46:1–27

Heimberg AM, Cowper-Sallari R, Semon M, Donoghue PCJ, Peterson KJ (2010) microRNAs reveal the interrelationships of hagfish, lampreys, and gnathostomes and the nature of the ancestral vertebrate. Proc Natl Acad Sci U S A 107:19379–19383

Heimberg AM, Sempere LF, Moy VN, Donoghue PCJ, Peterson KJ (2008) MicroRNAs and the advent of vertebrate morphological complexity. Proc Natl Acad Sci U S A 105:2946–2950

Helm C, Bernhart SH, Siederdissen CHZ, Nickel B, Bleidorn C (2012) Deep sequencing of small RNAs confirms an annelid affinity of Myzostomida. Mol Phylogenet Evol 64:198–203

Hennig W (1965) Phylogenetic systematics. Annu Rev Entomol 10:97–116

Hertel J, Lindemeyer M, Missal K, Fried C, Tanzer A, Flamm C, Hofacker IL, Stadler PF, Students Bioinformatics Computer L (2006) The expansion of the metazoan microRNA repertoire. BMC Genomics 7:25

Hilker R, Sickinger C, Pedersen CNS, Stoye J (2012) UniMoG—a unifying framework for genomic distance calculation and sorting based on DCJ. Bioinformatics 28:2509–2511
Hu F, Lin Y, Tang J (2014) MLGO: phylogeny reconstruction and ancestral inference from gene-order data. BMC Bioinformatics 15:354
Huang CRL, Burns KH, Boeke JD (2012) Active transposition in genomes. Annu Rev Genet 46:651–675
Huff JT, Zilberman D, Roy SW (2016) Mechanism for DNA transposons to generate introns on genomic scales. Nature 538:533–536
Irimia M, Roy SW (2008) Spliceosomal introns as tools for genomic and evolutionary analysis. Nucleic Acids Res 36:1703–1712
Jeffares DC, Mourier T, Penny D (2006) The biology of intron gain and loss. Trends Genet 22:16–22
Jeffroy O, Brinkmann H, Delsuc F, Philippe H (2006) Phylogenomics: the beginning of incongruence? Trends Genet 22:225–231
Kaiser VB, van Tuinen M, Ellegren H (2007) Insertion events of CR1 retrotransposable elements elucidate the phylogenetic branching order in Galliform Birds. Mol Biol Evol 24:338–347
Kalendar R, Flavell AJ, Ellis THN, Sjakste T, Moisy C, Schulman AH (2011) Analysis of plant diversity with retrotransposon-based molecular markers. Heredity 106:520–530
Kazazian HH (2004) Mobile elements: drivers of genome evolution. Science 303:1626–1632
Keeling PJ, Doolittle WF (1997) Widespread and ancient distribution of a noncanonical genetic code in diplomonads. Mol Biol Evol 14:895–901
Kim VN (2005) MicroRNA biogenesis: coordinated cropping and dicing. Nat Rev Mol Cell Biol 6:376–385
Kiontke K, Gavin NP, Raynes Y, Roehrig C, Piano F, Fitch DHA (2004) *Caenorhabditis* phylogeny predicts convergence of hermaphroditism and extensive intron loss. Proc Natl Acad Sci U S A 101: 9003–9008
Knight RD, Freeland SJ, Landweber LF (2001) Rewiring the keyboard: evolvability of the genetic code. Nat Rev Genet 2:49–58
Kosik KS (2009) OPINION MicroRNAs tell an evo-devo story. Nat Rev Neurosci 10:754–759
Kozomara A, Griffiths-Jones S (2011) miRBase: integrating microRNA annotation and deep-sequencing data. Nucleic Acids Res 39:D152–D157
Kozomara A, Griffiths-Jones S (2014) miRBase: annotating high confidence microRNAs using deep sequencing data. Nucleic Acids Res 42:D68–D73
Kramerov DA, Vassetzky NS (2005) Short retroposons in Eukaryotic genomes. In: International review of cytology, vol 247. Academic Press, New York, pp 165–221
Krauss V, Thümmler C, Georgi F, Lehmann J, Stadler PF, Eisenhardt C (2008) Near intron positions are reliable phylogenetic markers: an application to holometabolous insects. Mol Biol Evol 25:821–830
Kriegs JO, Churakov G, Kiefmann M, Jordan U, Brosius J, Schmitz J (2006) Retroposed elements as archives for the evolutionary history of placental mammals. PLoS Biol 4:e91
Kriegs JO, Matzke A, Churakov G, Kuritzin A, Mayr G, Brosius J, Schmitz J (2007) Waves of genomic hitchhikers shed light on the evolution of gamebirds (Aves: Galliformes). BMC Evol Biol 7:190
Krzywinski J, Besansky NJ (2002) Frequent intron loss in the white gene: a cautionary tale for phylogeneticists. Mol Biol Evol 19:362–366
Kumar S, Stecher G, Tamura K (2016) MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. Mol Biol Evol 33:1870–1874
Lehmann J, Stadler PF, Krauss V (2013) Near intron pairs and the metazoan tree. Mol Phylogenet Evol 66:811–823
Lewis BP, Shih IH, Jones-Rhoades MW, Bartel DP, Burge CB (2003) Prediction of mammalian microRNA targets. Cell 115:787–798
Lin Y, Fei H, Tang J, Moret BME 2012a Maximum likelihood phylogenetic reconstruction from high-resolution whole-genome data and a tree of 68 eukaryotes. In: Proceedings of the 18th pacific symposium on Biocomputing (PSB'13), Singapore, World Scientific, pp 285–296
Lin Y, Rajan V, Moret BME (2012b) TIBA: a tool for phylogeny inference from rearrangement data with bootstrap analysis. Bioinformatics 28:3324–3325
Liu N, Okamura K, Tyler DM, Phillips MD, Chung WJ, Lai EC (2008) The evolution and functional diversification of animal microRNA genes. Cell Res 18:985–996
Lyons M, Cardle L, Rostoks N, Waugh R, Flavell AJ (2008) Isolation, analysis and marker utility of novel miniature inverted repeat transposable elements from the barley genome. Mol Gen Genomics 280:275–285

Malé P-JG, Bardon L, Besnard G, Coissac E, Delsuc F, Engel J, Lhuillier E, Scotti-Saintagne C, Tinaut A, Chave J (2014) Genome skimming by shotgun sequencing helps resolve the phylogeny of a pantropical tree family. Mol Ecol Resour 14:966–975

Malik HS, Eickbush TH (1998) The RTE class of non-LTR retrotransposons is widely distributed in animals and is the origin of many SINEs. Mol Biol Evol 15:1123–1134

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

Moret BME, Lin Y, Tang J (2013) Rearrangements in phylogenetic inference: compare, model, or encode? In: Chauve C, El-Mabrouk N, Tannier E (eds) Models and algorithms for genome evolution. Springer London, London, pp 147–171

Moret BME, Wyman S, Bader D, Warnow T, Yan M A 2001 new implementation and detailed study of breakpoint analysis. In: Proceedings of the 6th pacific symposim on Biocomputing (PSB'01), Singapore, World Scientific, pp 583–594

Morgan GT (1995) Identification in the human genome of mobile elements spread by DNA-mediated transposition. J Mol Biol 254:1–5

Nishihara H, Maruyama S, Okada N (2009) Retroposon analysis and recent geological data suggest near-simultaneous divergence of the three superorders of mammals. Proc Natl Acad Sci U S A 106:5235–5240

Nozawa M, Miura S, Nei M (2010) Origins and evolution of microRNA genes in *Drosophila* species. Genome Biol Evol 2:180–189

Osawa S, Jukes TH (1989) Codon reassignment (codon capture) in evolution. J Mol Evol 28:271–278

Pearson JC, Lemons D, McGinnis W (2005) Modulating Hox gene functions during animal body patterning. Nat Rev Genet 6:893–904

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 470:255–258

Ray DA, Xing J, Salem AH, Batzer MA (2006) SINEs of a nearly perfect character. Syst Biol 55:928–935

Richter S, Meier R (1994) The development of phylogenetic concepts in Hennig's early theoretical publications (1947–1966). Syst Biol 43:212–221

Richter S, Schwarz F, Hering L, Böggemann M, Bleidorn C (2015) The utility of genome skimming for phylogenomic analyses as demonstrated for glycerid relationships (Annelida, Glyceridae). Genome Biol Evol 7:3443–3462

Rogozin IB, Wolf YI, Sorokin AV, Mirkin BG, Koonin EV (2003) Remarkable interkingdom conservation of intron positions and massive, lineage-specific intron loss and gain in eukaryotic evolution. Curr Biol 13:1512–1517

Rokas A, Holland PWH (2000) Rare genomic changes as a tool for phylogenetics. Trends Ecol Evol 15:454–459

Rota-Stabelli OR-SO, Campbell L, Brinkmann H, Edgecombe GD, Longhorn SJ, Peterson KJ, Pisani D, Philippe H, Telford MJ (2011) A congruent solution to arthropod phylogeny: phylogenomics, microRNAs and morphology support monophyletic Mandibulata. Proc R Soc Lond B Biol Sci 278:298–306

Roy SW (2016) How common is parallel intron gain? Rapid evolution versus independent creation in recently created introns in daphnia. Mol Biol Evol 33:1902–1906

Roy SW, Gilbert W (2005a) Rates of intron loss and gain: implications for early eukaryotic evolution. Proc Natl Acad Sci U S A 102:5773–5778

Roy SW, Gilbert W (2005b) Resolution of a deep animal divergence by the pattern of intron conservation. Proc Natl Acad Sci U S A 102:4403–4408

Sankoff D, Leduc G, Antoine N, Paquin B, Lang BF, Cedergren R (1992) Gene order comparisons for phylogenetic inference: evolution of the mitochondrial genome. Proc Natl Acad Sci U S A 89:6575–6579

Schultz DW, Yarus M (1994) Transfer RNA mutation and the malleability of the genetic code. J Mol Biol 235:1377–1380

Sempere LF, Cole CN, McPeek MA, Peterson KJ (2006) The phylogenetic distribution of metazoan microRNAs: Insights into evolutionary complexity and constraint. J Exp Zool Part B 306B:575–588

Sengupta S, Yang X, Higgs PG (2007) The mechanisms of codon reassignments in mitochondrial genetic codes. J Mol Evol 64:662–688

Shabalina SA, Koonin EV (2008) Origins and evolution of eukaryotic RNA interference. Trends Ecol Evol 23:578–587

Shao R, Barker SC (2003) The highly rearranged mitochondrial genome of the plague thrips, Thrips imaginis (Insecta: Thysanoptera): convergence of two novel gene boundaries and an extraordinary arrangement of rRNA Genes. Mol Biol Evol 20:362–370

Shedlock AM, Okada N (2000) SINE insertions: powerful tools for molecular systematics. BioEssays 22:148–160

Shimamura M, Yasue H, Ohshima K, Abe H, Kato H, Kishiro T, Goto M, Munechika I, Okada N (1997) Molecular evidence from retroposons that whales form a clade within even-toed ungulates. Nature 388:666–670

Skalsky RL, Cullen BR (2010) Viruses, microRNAs, and host interactions. Annu Rev Microbiol 64:123–141

Sperling EA, Pisani D, Peterson KJ (2011) Molecular paleobiological insights into the origin of the Brachiopoda. Evol Dev 13:290–303

Sperling EA, Vinther J, Moy VN, Wheeler BM, Semon M, Briggs DEG, Peterson KJ (2009) MicroRNAs resolve an apparent conflict between annelid systematics and their fossil record. Proc R Soc Lond B Biol Sci 276:4315–4322

Stamatakis A (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30:1312–1313

Sturtevant AH, Dobzhansky T (1936) Inversions in the third chromosome of wild races of *Drosophila pseudoobscura*, and their use in the study of the history of the species. Proc Natl Acad Sci U S A 22:448–450

Suh A, Paus M, Kiefmann M, Churakov G, Franke FA, Brosius J, Kriegs JO, Schmitz J (2011) Mesozoic retroposons reveal parrots as the closest living relatives of passerine birds. Nat Commun 2:443

Suh A, Smeds L, Ellegren H (2015) The dynamics of incomplete lineage sorting across the ancient adaptive radiation of neoavian birds. PLoS Biol 13:e1002224

Sverdlov AV, Rogozin IB, Babenko VN, Koonin EV (2005) Conservation versus parallel gains in intron evolution. Nucleic Acids Res 33:1741–1748

Takahashi K, Terai Y, Nishida M, Okada N (2001) Phylogenetic relationships and ancient incomplete lineage sorting among cichlid fishes in Lake Tanganyika as revealed by analysis of the insertion of retroposons. Mol Biol Evol 18:2057–2066

Tanzer A, Amemiya CT, Kim CB, Stadler PF (2005) Evolution of microRNAs located within Hox gene clusters. J Exp Zool Part B 304B:75–85

Tarailo-Graovac M, Chen N (2009) Using repeatMasker to identify repetitive elements in genomic sequences. Current Protocols in Bioinformatics Chapter 4, Unit 4, p 108

Tarver JE, Sperling EA, Nailor A, Heimberg AM, Robinson JM, King BL, Pisani D, Donoghue PCJ, Peterson KJ (2013) miRNAs: Small genes with big potential in metazoan phylogenetics. Mol Biol Evol 30:2369–2382

Telford MJ, Herniou EA, Russell RB, Littlewood DTJ (2000) Changes in mitochondrial genetic codes as phylogenetic characters: two examples from the flatworms. Proc Natl Acad Sci U S A 97:11359–11364

Thomson RC, Plachetzki DC, Mahler DL, Moore BR (2014) A critical appraisal of the use of microRNA data in phylogenetics. Proc Natl Acad Sci U S A 111:E3659–E3668

Venkatesh B, Ning Y, Brenner S (1999) Late changes in spliceosomal introns define clades in vertebrate evolution. Proc Natl Acad Sci U S A 96:10267–10271

Waddell PJ, Kishino H, Ota R (2001) A phylogenetic foundation for comparative mammalian genomics. Genome Inform 12:141–154

Wessler SR, Bureau TE, White SE (1995) LTR-retrotransposons and MITEs: important players in the evolution of plant genomes. Curr Opin Genet Dev 5:814–821

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

Yaakov B, Ceylan E, Domb K, Kashkush K (2012) Marker utility of miniature inverted-repeat transposable elements for wheat biodiversity and evolution. Theor Appl Genet 124:1365–1373

Yancopoulos S, Attie O, Friedberg R (2005) Efficient sorting of genomic permutations by translocation, inversion and block interchange. Bioinformatics 21:3340–3346

Zheng J, Rogozin IB, Koonin EV, Przytycka TM (2007) Support for the Coelomata Clade of Animals from a Rigorous Analysis of the Pattern of Intron Conservation. Mol Biol Evol 24:2583–2592

Service Part

C. Bleidorn, *Phylogenomics*, DOI 10.1007/978-3-319-54064-1

Glossary

Absence/presence coding (a/p coding) Coding of a matrix where for all characters only the information if absent (0) or present (1) are coded as character states.

Adaptor Short, chemically synthesized, double-stranded DNA molecule which can be linked to the ends of other DNA molecules.

Alignment Hypothesis of positional homologies between the nucleotides and amino acids of a pair or multiple sequences. Global alignments assume positional homology across all positions of the aligned sequences. Local alignments optimize positional for fragments (substrings) of two sequences.

Annotation Prediction and description of genes (and other genetic loci) within a genome.

Apomorphy Character state showing the derived condition.

Arminia Bielefeld German football club with a more than 100-year tradition.

Assembly Process of constructing contiguous sequences (contigs) from sequence reads.

Base calling Transformation of raw sequencing data (e.g. pictures) into a sequence read.

Base composition Frequency of each nucleotide in a sequence.

Bayes factor Ratio of probabilities of two models that is used to evaluate the relative support of one model in relation to another in comparison with Bayesian models.

Bayesian inference Likelihood-based method of phylogenetic reconstruction rooted in a Bayesian framework. Posterior probability of a tree is calculated by multiplying its prior probability with the likelihood of the observed data while using a normalizing constant of this product as denominator. Markov chain Monte Carlo methods are used to approximate the posterior probability.

Bifurcation Two branches which are connected by an internal node.

Biotinylation Attachment of a biotin label to a DNA or RNA molecule.

BLAST Algorithm that uses a heuristic approach to find sequences similar to a query sequence within a database.

Blunt end End of a double-stranded DNA molecular with nucleotides in both strands terminating at the same position.

Bootstrap Statistical method based on resampling and analysis of pseudoreplicates to measure the confidence of nodes (internal branches) in a phylogenetic tree.

Character A feature that can be compared across organisms, e.g. positions in an alignment, retrotransposon integrations.

Character matrix Representation of characters and their according character states for a group of terminals. Terminals are represented by rows, characters by columns.

Character state Conditions of a character that can be observed (and coded in a matrix) across compared organisms, e.g. the four nucleotides (A,C,G,T) in case if position in the alignment is the character.

Codon Nucleotide triplet coding for a single amino acid.

complementary DNA (cDNA). Double-stranded DNA synthesized from single-stranded RNA.

Compositional bias Bias in the base compositions across terminals in an alignment.

Convergence Independent gain of a character or character state in at least two different lineages.

Chromatin Macromolecular complex of DNA and proteins (histones) in eukaryotic chromosomes, mainly involved in DNA packaging and protection, but also in the control of gene expression.

Clade Monophyletic group in a phylogenetic tree, thereby representing at least two terminals (which share a common ancestor).

Cladogram. Phylogenetic topology where the branch length does not contain extra information.

Coalescence Model to follow the history of genes (or alleles) back in time.

Contig Contiguous sequence derived from assembling overlapping sequence reads.

Coverage Number of times a sequence position within a contig is covered by sequence reads.

c-value DNA content of a haploid cell in picogram (pg). A c-value of 1 pg equals ~978 Mb.

Dirichlet distribution Family of continuous multivariate probability distributions which are often used as prior distributions in a Bayesian framework.

DNA barcode Standardized DNA region for species identification.

Endosymbiont A symbiont which lives within the body or cells of another organism.

Euler path Path in a mathematical graph which goes over every edge (branch) exactly once.

Evolutionary distance Observed distance for a pair of sequences which has been corrected by a substitution model.

Fastq Widely used file format for next-generation sequencing reads, encoding identity and quality of each nucleotide.

Fluorophore Fluorescent label attached to a nucleotide which can be detected by a laser.

Gap, gap position. Placeholder inserted in sequences within an alignment to mark indel events.

Gene Ontology (GO) The GO project comprises three different ontologies (controlled vocabularies) to describe functions of gene products related to their cellular components, biological processes and molecular functions.

Hamiltonian path Path in a mathematical graph that goes over every node exactly once.

Heterotachy Variation of substitution rates within lineages over time.

Heuristic Approach that does not guarantee finding the optimal solution, but which is faster, speeding up the process.

Homology A character is regarded as homologous when it has been inherited from a common ancestor. Different types of homology are considered in molecular systematics. Positional homology describes the homology of alignment sites, where the site itself represents the character and the nucleotides (or amino acids) at this site are the possible character states. With orthology, paralogy and xenology, different types of homology of genetic loci are distinguished.

Hidden Markov model (HMM) HMMs are probabilistic models which model a system under the assumption that it can be represented by a Markov process with hidden (unobserved) states. HMMs are used to find sequence similarities or to generate alignments.

Homopolymer Series of consecutive identical nucleotides within a DNA molecule.

Illumina Company distributing sequencers. The name Illumina is often synonymously used for the type of sequencing their machines are using, namely, reversible terminator sequencing.

Indel A type of mutation where an insertion or deletion event involves a small number of base pairs.

Introgression Gene flow from one species into the gene pool of another species.

Invariant site Alignment site which shows the same character state for all terminals.

Ion Torrent Company distributing sequencers. The name Illumina is often synonymously used for the type of sequencing their machines are using, namely, ion semiconductor sequencing.

***k*-mer** All possible *k*-1 overlapping substrings, of a given length *k*, of a sequence.

Long-branch attraction (LBA) Condition where a phylogenetic analysis is biased due to a combination of short and long branches.

Mapping Alignment of sequence reads to a reference.

Majority-rule consensus Consensus representation of a set of trees, where all internal nodes are displayed which can be found in more than half of the summarized topologies. Frequencies of how often an internal node is found are usually indicated.

Mate pairs Sequenced ends of DNA fragments which are usually separated by a defined size.

Markov process Stochastic process where changes of states are only depending on the current state and given transition rules.

Maximum likelihood (ML) Likelihood-based optimality criterion to find the best tree in a phylogenetic analysis through the computation of probabilities of character evolution given an explicit evolutionary model.

Maximum parsimony (MP) Optimality criterion which selects the phylogenetic tree(s) minimizing the total number of character state changes.

Monophyletic group Group containing a (hypothetical) ancestor and all of its descendants.

Monte Carlo method Computational algorithm using repeated random sampling.

Multifurcation More than two branches connected by an internal node. Also known as polytomy.

Neighbor joining (NJ) Algorithm, which chooses the tree with the smallest sum of branch lengths given a matrix of pairwise distances.

Neofunctionalization Process where one copy of a gene acquires a new function after a gene duplication event.

Network Phylogenetic networks are graph-like representations of a topology. In contrast to bifurcating trees, networks allow reticulations.

Newick format File format for the representation of phylogenetic trees.

Nonfunctionalization Fate when one copy loses all functional ability after a gene duplication event. Also known as pseudogenization.

Operational taxonomic unit (OTU) Terminal used in phylogenetic analyses.

Ortholog conjecture Hypothesis that orthologs are on average functionally more similar than paralogs.

Orthology Pairs of homologous genes which have emerged through a speciation event are called orthologs.

Outgroup Taxon or taxa used to determine the root of a phylogenetic tree.

Overlap-layout-consensus (OLC) Assembly strategy where reads are aligned pairwise and according overlap information is stored within a mathematical graph. Based on the Hamilton path, relative order (layout) of reads and consensus of contigs is achieved.

PacBio Company distributing sequencers. The name Illumina is often synonymously used for the type of sequencing their machines are using, namely, single molecule real-time (SMRT) sequencing.

Paired-end Sequencing of a molecule from both ends.

Paralogy Pairs of homologous genes which have emerged through a gene duplication event are called paralogs. Inparalogs are paralogs that arose by duplication after the speciation event separating the lineages which are compared; outparalogs are those paralogs where the duplication event happened before the speciation event.

p-distance Observed distance between a pair of sequences, calculated by counting the differences divided by the alignment length.

PFAM Database containing a large collection of protein families represented by multiple sequence alignments and their respective hidden Markov profiles.

Phred Quality score for the description of accuracy of sequencing reads.

Phylogeny Evolutionary history of genes or organisms.

Phylogram Phylogenetic topology where the branch length contains information about evolutionary change.

Plesiomorphy Character state showing the ancestral condition.

Poisson distribution Discrete probability distribution expressing a given number of events in a fixed interval.

Polarity Direction of character change.

Polyadenylation Addition of a series of A's to the 3′ end of a eukaryotic mRNA.

RADseq Restriction site-associated DNA sequencing, a method used to sequence a reduced, but consistent representation of the genome.

Rate heterogeneity Difference in substitution rate across alignment positions.

Read (or sequence read) Output of a sequencing platform given as DNA (or RNA) sequence.

Restriction enzyme Enzyme that cuts DNA at a specific site.

Retrotransposon Mobile elements that have the ability to integrate into the genome at a new site within their cell of origin.

RNA-Seq High-throughput random sequencing of cDNA fragments using next-generation sequencing technology.

Root Point of a topology where it is hypothetically connected to the remaining tree of life. Rooted trees are used to polarize character evolution.

Scaffold Ordering of contigs based on additional information (e.g. mate pairs, optical mapping).

Sequencing library Collection of DNA or RNA linked with an adaptor, which can subsequently be used for sequencing.

Splicing Removal of introns from the primary transcript in eukaryotic genes.

Sticky end End of a double-stranded DNA molecular where nucleotides in the tow strands are terminating in different positions, thereby creating an extension, which can be targeted for ligation.

Strict consensus Consensus representation of a set of trees, where only internal nodes are displayed which can be found in all summarized topologies.

Subfunctionalization Persistence of partial ancestral functions in different copies of a gene after a duplication event.

Substitution models Statistical methods use substitution models to describe how sequences change over time. Substitution models are available for nucleotide and amino acid sequences.

Supertree Phylogenetic tree which is built according to certain rules (methods) based on a set of tree topologies which overlap in their taxon sampling.

Taxon A taxon (plural: taxa) is a group of organisms that is given a formal taxonomic name.

Taxon sampling Collection of taxa included into a phylogenetic study.

Terminal Entity included in a phylogenetic analysis which will be resolved at the tips of the tree, e.g. taxa, individuals and genes.

Topology The topology of a tree describes the branching pattern of a phylogeny.

Transcript RNA copy of a gene.

Transposition Movement of a genetic element from one site to another site in a DNA molecule.

Ultrametric tree Phylogenetic tree where the path lengths from root to tip are equidistant.

Unitig High-confidence contigs.

Ultraconserved elements (UCEs) Highly conserved segments of animal, plant or fungal genomes which are not functionally transcribed. UCEs have been successfully exploited as a phylogenetic marker.

Whole-genome shotgun (wgs) sequencing Genome sequencing strategy where sheared DNA is sequenced and assembled bioinformatically afterwards.

Xenology Pairs of homologous genes where its common evolutionary history involves horizontal gene transfer of at least one of these genes are called xenologs.

Index

F

G

H

I

J

K

L

M

N

O

P

Q

R

S

T

U

W

X

Z