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

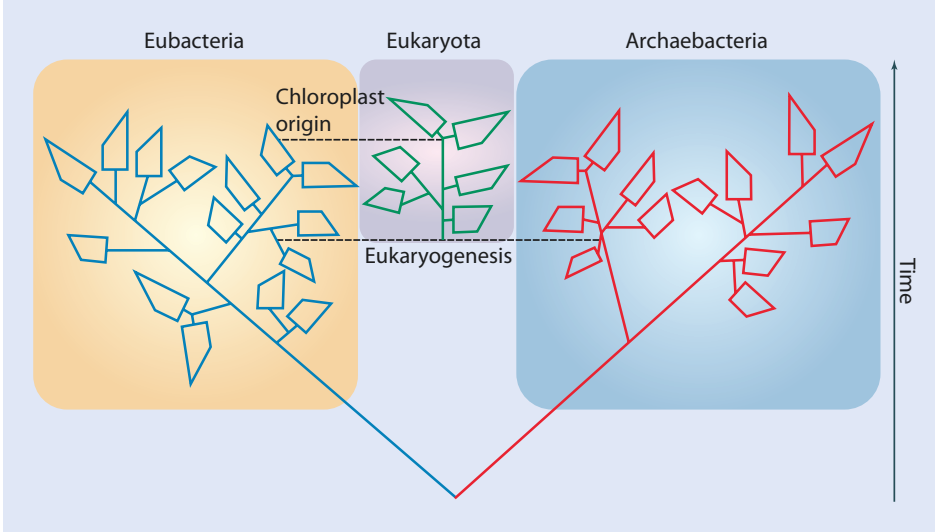

Fig. 1.1 The ring of life hypothesis (Reprinted by permission from Macmillan Publishers Ltd: Nature (McInerney et al. 2014), Copyright 2014)

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

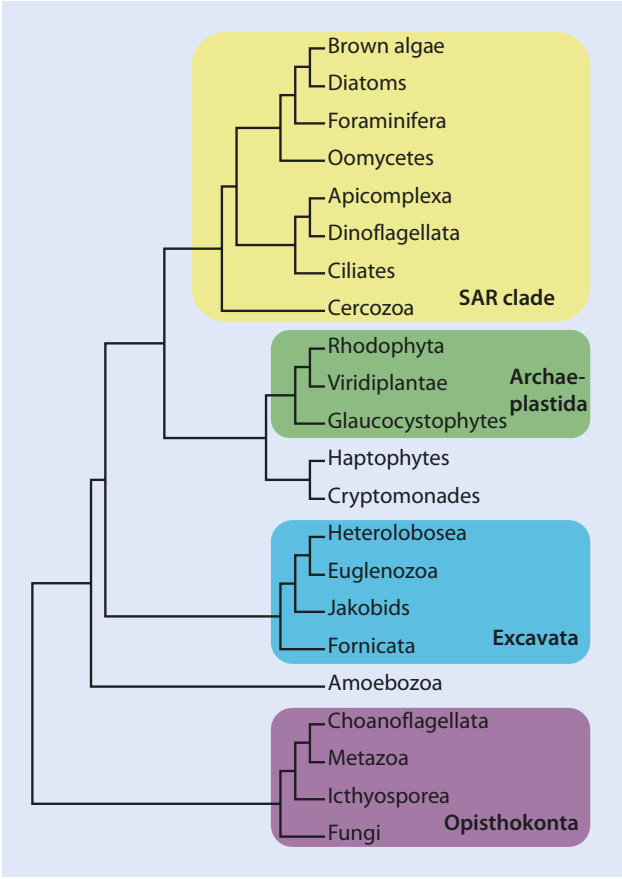

Fig. 1.2 Phylogenetic relationships of eukaryotes based on the phylogenomic analyses of Katz and Grant (2014)

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).

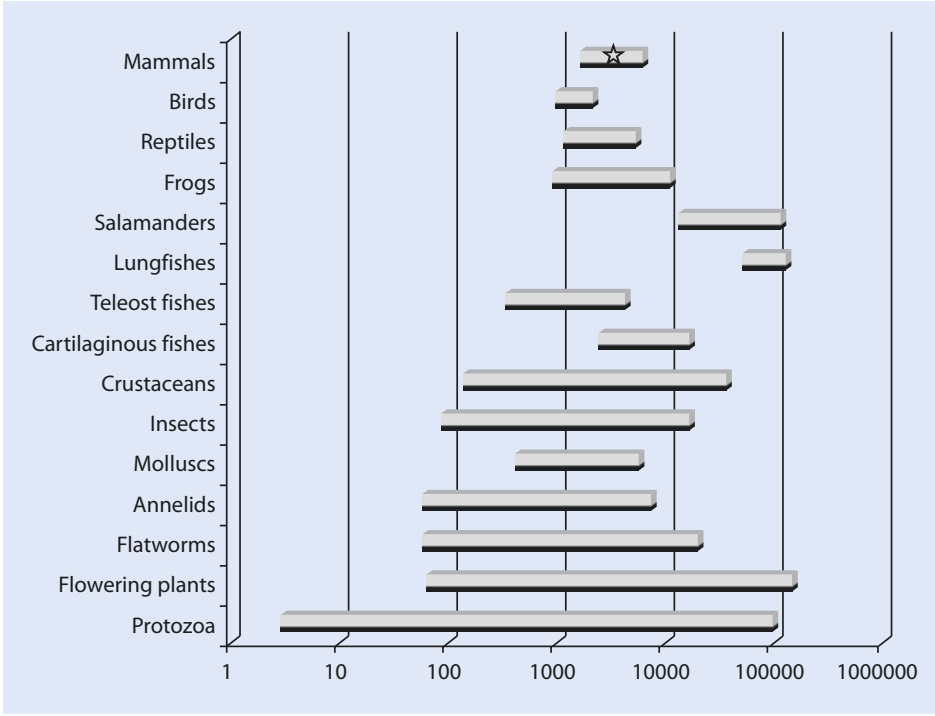

Fig. 1.3 Variation of genome size (given in Kb) across eukaryotes (Reprinted from Palazzo and Gregory (2014))

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.

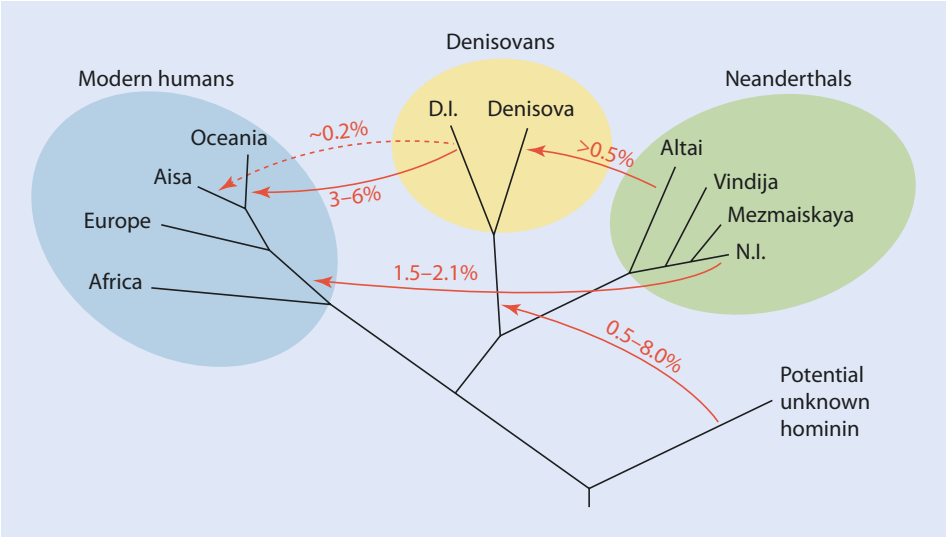

Fig. 1.4 Human relationships and possible model of gene flow. Note that the age of archaic genomes does not allow detection of gene flow of modern humans towards them (Reprinted by permission from Macmillan Publishers Ltd: Nature (Prüfer et al. 2014), Copyright 2014)

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