# Eukaryote Genomes

# 8

## Chapter Summary

General overviews of eukaryote genomes are first discussed, including organelle genomes, introns, and junk DNAs. We then discuss the evolutionary features of eukaryote genomes, such as genome duplication, C-value paradox, and the relationship between genome size and mutation rates. Genomes of multicellular organisms, plants, fungi, and animals are then briefly discussed.

## 8.1 Major Differences Between Prokaryote and Eukaryote genomes

A eukaryotic cell has a nucleus, surrounded by the nuclear membrane. There are other membrane systems in their cells, such as endoplasmic reticulum, Golgi apparatus, and vacuole. Prokaryotes do not have these membranes nor organella. Therefore, existence of membrane systems and organella, particularly mitochondria, are the two major characteristics of eukaryotes. It should be noted that some parasitic eukaryotes lost mitochondria. Genome sizes of eukaryotes became much bigger than those of prokaryotes. Accordingly, gene numbers are also more abundant in eukaryotes than prokaryotes. It is not clear if the formation of nucleus triggered the increase of the genome size.

There are various differences of genome structures between prokaryotes and eukaryotes, and they are listed in Table 8.1. Most of bacterial genomes are circular, while all eukaryotic genomes so far known are linear (here organelle genomes are not considered). The main reason for a large genome size in eukaryotes is the existence of many repeat sequences, which are minority in prokaryotes. Pseudogenes and introns are also few in prokaryotic genomes, while both are abundant in eukaryotic genomes. High occurrences of gene duplications in eukaryotes prompted production of many pseudogenes. Horizontal gene transfers are known to be quite frequent in prokaryotes, and they are rare in eukaryotes. Finally, genome

N. Saitou, *Introduction to Evolutionary Genomics*, Computational Biology 17,
DOI 10.1007/978-1-4471-5304-7_8, © Springer-Verlag London 2013

**Table 8.1** Comparison of prokaryotic and eukaryotic genomes

| Category | Prokaryotes | Eukaryotes |
|---|---|---|
| Size | Small (1–10 Mb) | Large (3–5,000 Mb) |
| Gene number | Small (<10,000) | Many (often > 10,000) |
| Topology | Mostly circular | Linear |
| Intergenic region | Short (<100 bp) | Long (often >100 kb) |
| Repeat sequence | Minor component | Major component |
| Pseudogene | Few | Many |
| Intron | Few | Usually exit |
| Complexity | Low | High |
| Horizontal gene transfer | Frequent | Rare |
| Gene duplication | Rare | Frequent |
| Genome duplication | None | Frequent (especially in plants and vertebrates) |

**Table 8.2** Examples of genes shared among most of eukaryote genomes but nonexisting in prokaryote genomes

| |
|---|
| DNA polymerase subunit γ 1 |
| DNA topoisomerase 1 |
| Histone H2B |
| Microtubule binding protein RP/EB family 2 |
| Myosin |
| Nucleolin |
| Translation initiating factor |
| Ubiquitin |

duplications sometimes occur in eukaryotes, especially in plants and in vertebrates, but genome duplication is so far not known for prokaryotic genomes.

Because the gene number of typical eukaryotic genomes is much larger than that of prokaryotes, there are many genes shared among most of eukaryote genomes but nonexisting in prokaryote genomes. Some examples are listed in Table 8.2. For example, myosin is located in animal muscle tissues, and its homologous protein exists in cytoskeleton of all eukaryotes, but not found in prokaryotes.

Recently, Kryukov et al. (2012; [1]) constructed a new database on oligonucleotide sequence frequencies and conducted a series of statistical analyses. Frequencies of all possible 1–10 oligonucleotides were counted for each genome, and these observed values were compared with expected values computed under observed oligonucleotide frequencies of length 1–4. Deviations from expected values were much larger for eukaryotes than prokaryotes, except for fungal genomes. Figure 8.1 shows the distribution of the deviation for various organismal groups. The biological reason for this difference is not known.

## 8.2 Organelle Genomes

There are two major types of organella in eukaryotes: mitochondria and plastids. Figure 8.2 shows schematic views of mitochondria and chloroplasts. These two organella has their independent genomes. This suggests that they were initially

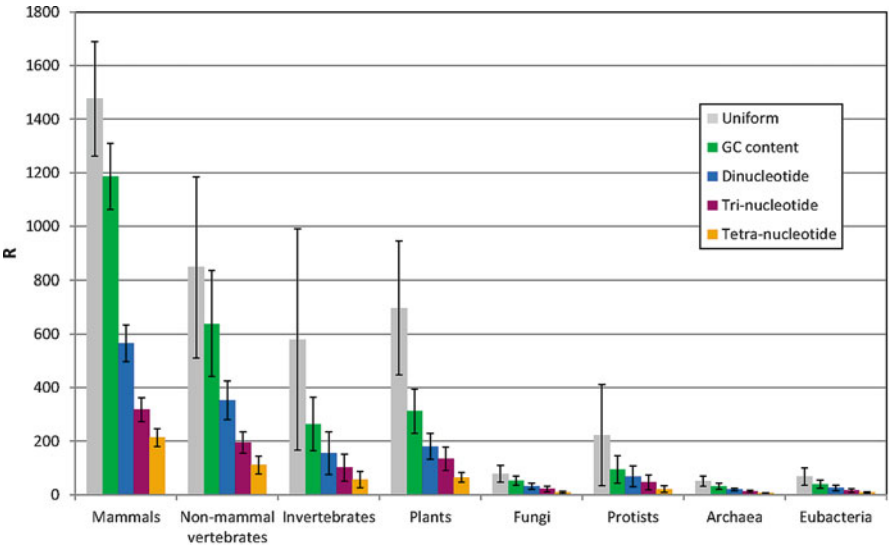


**Fig. 8.1** Comparison of genome complexity among eukaryote genomes (From Kryukov et al. 2012; [1])

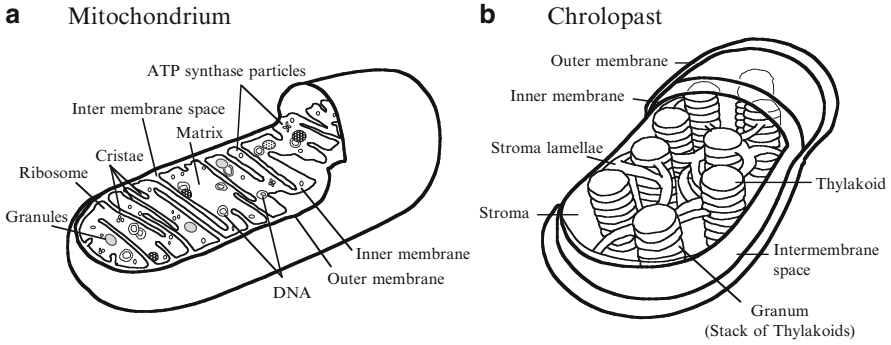


**Fig. 8.2** Schematic views of mitochondrion and chloroplast

independent organisms which started intracellular symbiosis with primordial eukaryotic cells. Because most eukaryotes have mitochondria, the ancestral eukaryotes, a lineage that emerged from Archaea, most probably started intracellular symbiosis with mitochondrial ancestor. A parasitic *Rickettsia prowazekii* is so far phylogenetically closest to mitochondria [2], and a rickettsia-like bacterium is the best candidate as the mitochondrial ancestor. However, there is an alternative "hydrogen hypothesis" [3]. Plastids include chloroplasts, leucoplasts, and chromoplasts and exist in land plants, green algae, red algae, glaucophyte algae, and some protists like euglenoids.

Mitochondrial genome sizes of some representative eukaryotes are listed in Table 8.3. Most of animal mitochondrial genomes are less than 20 kb, and sizes of protist and fungi mitochondrial genomes are somewhat larger. Mitochondrial genome size of plants is much larger than those of other eukaryotic lineages, yet the size is mostly less than 500 kb.

**Table 8.3** Size of mitochondrial genomes

| Organism | Genome size (kb) |
|---|---|
| Animals | |
| *Homo sapiens* (human) | 16.5 |
| *Takifugu rubripes* (Torafugu fish) | 16.5 |
| *Ciona intestinalis* (ascidian) | 14.8 |
| *Drosophila melanogaster* (fruit fly) | 19.5 |
| *Apis mellifer*a (honey bee) | 16.3 |
| *Limulus polyphemus* (horseshoe crab) | 15.0 |
| *Caenorhabditis elegans* (nematode) | 13.8 |
| *Schistosoma mansoni* (parasitic flatworm) | 14.4 |
| *Aplysia californica* (mollusk) | 14.1 |
| *Hydra magnipapillata* (freshwater polyp hydra) | 8.2 + 7.7 |
| Fungi | |
| *Moniliophthora perniciosa* | 109.1 |
| *Saccharomyces cerevisiae* (baker's yeast) | 75 |
| *Suillus grisells* (basidiomycete fungus) | 121 |
| Protists | |
| *Acanthamoeba castellanii* [Acanthamoebidae] | 41.6 |
| *Paramecium aurelia* [Alveolata] | 40.5 |
| *Plasmodium falciparum* [Alveolata] | 5.9 |
| *Tetrahymena thermophila* [Alveolata] | 47.6 |
| *Phytophthora infestans* [Stramenopiles] | 39.8 |
| *Reclinomonas americana* [Jakobida] | 69.0 |
| *Trypanosoma brucei brucei* [Euglenozoa] | 23.0 |
| Plants | |
| *Arabidopsis thaliana* (Wall cress) | 366.9 |
| *Oryza sativa indica* (indica rice) | 434.7 |
| *Oryza sativa japonica* (japonica rice) | 490.5 |
| *Brassica oleracea* (cabbage) | 160.0 |
| *Nicotiana tabacum* (tobacco) | 430.6 |
| *Zea mays* (corn) | 570.0 |
| *Cucumis melo* (melon) | 2,880 |
| *Chlamydomonas reinhardtii* (green alga) | 15.8 |
| *Chondrus crispus* (red alga) | 26 |

### 8.2.1 Mitochondria

An ancestral eukaryotic cell, probably an archaean lineage, hosted a bacterial cell, and intracellular symbiosis started. Initially, Archaea and Bacteria shared genes responsible for basic metabolism, and the situation is a sort of gene duplication for many genes, though homologous genes are not identical but already diverged long time ago. In any case, division of labor followed, and only limited metabolic pathways were left in the bacterial system, which eventually became mitochondria.

Animal mitochondrial genomes contain very small number of genes; 13 for peptide subunits, 20 for tRNA, and 2 for rRNA [4]. Figure 8.3 shows gene orders of five

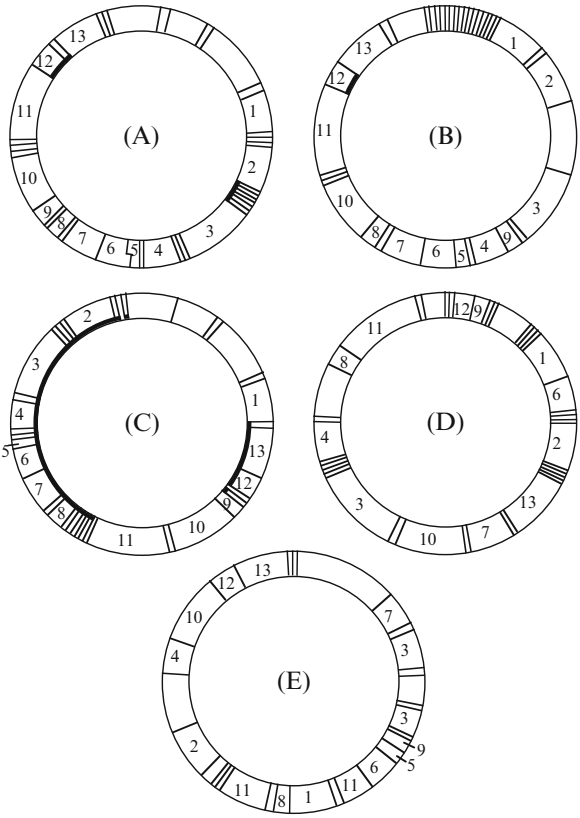


**Fig. 8.3** Gene orders of five animal mitochondrial DNA genomes (From Saitou 2007; [103])

representative animal species mitochondrial DNA genomes. Although most of vertebrate mitochondrial DNA genomes have the same gene order as in human (Fig. 8.3a), gene order may vary from phylum to phylum. Yet the gene content and the genome size are more or less constant among animals. It is not clear why animal mitochondrial genomes are so small. One possibility is that animal individuals are highly integrated compared to fungi and plants, and this might have influenced a drastic reduction of the mitochondrial genome size. Another interesting feature of animal mitochondrial DNA genomes is the heterogeneous rates of gene order change. For example, platyhelminthes exhibit great variability in mitochondrial gene order (Sakai and Sakaizumi, 2012; [5]).

In contrast, plant mitochondrial genomes are much larger (see Table 8.3). Figure 8.4 shows the genome structure of tobacco mitochondrial genome (from Sugiyama et al. 2005; [6]). Horizontal gene transfers are also known to occur in plant mitochondrial DNAs even between remotely related species [7].

The melon (*Cucumis melo*) mitochondrial genome size, ca. 2.9 Mb, is exceptionally large, and recently its draft genome was determined [8]. Interestingly, melon mitochondrial genome looks like the vertebrate nuclear genome in its contents, in

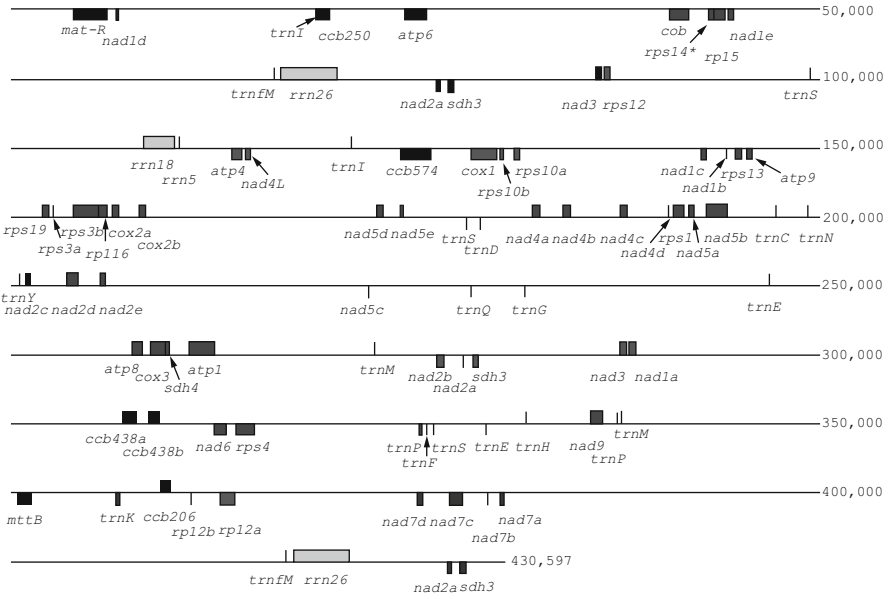


**Fig. 8.4** Genome structure of tobacco mitochondria (From Sugiyama et al. 2005; [6])

spite of its genome size being similar to that of bacteria. The protein coding gene region accounted for only 1.7 % of the genome, and about half of the genome is composed of repeats. The remaining part is mostly homologous to melon nuclear DNA, and 1.4 % is homologous to melon chloroplast DNA. Most of the protein coding genes of melon mitochondrial DNAs are highly similar to those of its congeneric species, which are watermelon and squash whose mitochondrial genome sizes are 119 kb and 125 kb, respectively. This indicates that the huge expansion of its genome size occurred only recently. Interestingly, cucumber (*Cucumis sativus*), another congeneric species, also has ~1.8-Mb mitochondrial genome with many repeat sequences [9]. It will be interesting to study whether the increase of mitochondrial genomes of melon and cucumber is independent or not.

### 8.2.2 Chloroplasts

Chloroplasts exist only in plants, algae, and some protists. It may change to leucoplasts and chromoplasts. Because of this, a generic name "plastids" may also be used. The origin of chloroplast seems to be a cyanobacterium that started intracellular symbiosis as in the case of mitochondria.

A unique but common feature of chloroplast genome is the existence of inverted repeats [10], and they mainly contain rRNA genes. Chloroplast DNA contents may

change during the plant growth, and matured leaves are devoid of DNA in their chloroplasts [11].

Chloroplast genomes were determined for more than 340 species as of December 2013 [106]. Their genome sizes range from 59 kb (*Rhizanthella gardneri*) to 521 kb (*Floydiella terrestris*). Although the largest chloroplast genome is still much smaller than atypical bacterial genome, its average intergenic length is 4 kb, much longer than that for bacterial genomes.

### 8.2.3 Interaction Between Nuclear and Organelle Genomes

Fractions of mitochondrial DNA may sometimes be inserted to nuclear genomes, and they are called "numts." An extensive analysis of the human genome found over 600 numts [12]. Their sequence patterns are random in terms of mitochondrial genome locations. This suggests that mitochondrial DNAs themselves were inserted, not via cDNA reverse-transcribed from mitochondrial mRNA. A possible source is sperm mitochondrial DNA that were fragmented after fertilization [12]. The reverse direction, from nucleus to mitochondria, was observed in melon, as discussed in subsection 8.2.1.

## 8.3 Intron

Intron is a DNA region of a gene that is eliminated during splicing after transcription of a long premature mRNA molecule. Intron was discovered by Phillip A. Sharp and Richard J. Roberts in 1977 as "intervening sequence" [13], but the name "intron" coined by Walter Gilbert in 1978 [14] is now widely used. It should be noted that some description on intron by Kenmochi [15] was used for writing this section.

### 8.3.1 Classification of Intron

There are various types of introns, but they can be classified into two: those requiring spliceosomes (spliceosome type) and self-splicing type. Figure 8.5 shows the splicing mechanisms of these two major types. Most of introns in nuclear genomes of eukaryotes are spliceosome type, and there are common GU–AG type and rare AU–AC type, depending on the nucleotide sequences of the intron–exon boundaries [16]. Spliceosomes involving these two types differ [17].

Self-splicing introns are divided into three groups: groups I, II, and III. Group I introns exist in organellar and nuclear rRNA genes of eukaryotes and prokaryotic tRNA genes. Group II are found in organellar and some eubacterial genomes. Cavalier-Smith [18] suggested that spliceosome-type introns originated from group II introns because of their similarity in splicing mechanism and structural similarity between group II introns and spliceosomal RNA. Group III introns exist in organellar genomes, and its splicing system is similar with that of group II intron, though they are smaller and have unique secondary structure.

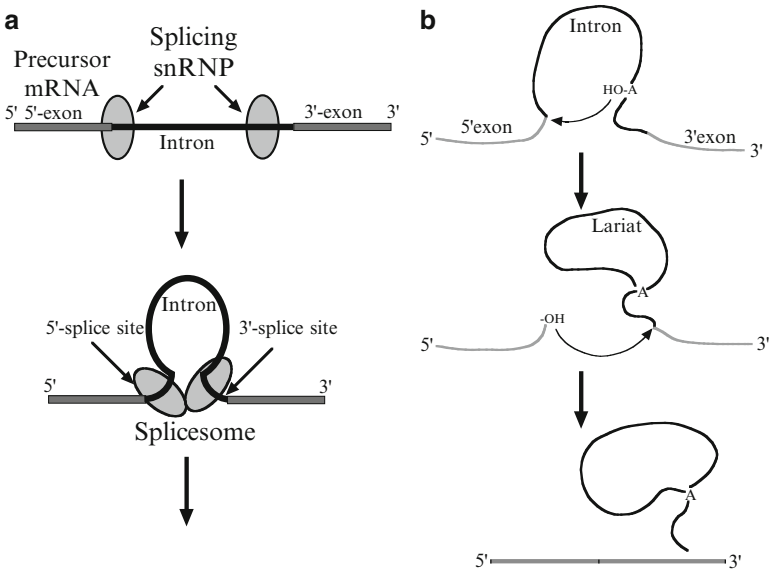


**Fig. 8.5** Two major types of introns. (**a**) Spliceosome type. (**b**) Self-splicing type

There is yet another type of introns which exist only in tRNAs of single-cell eukaryotes and Archaea [19]. These introns do not have self-splicing functions, but endonuclease and RNA ligase are involved in splicing. The location of this type of introns is often at a certain position of the tRNA anticodon loop.

### 8.3.2 Introns Early/Late Controversy

After the discovery of introns, their probable functions and evolutionary origin have long been argued (e.g., [20, 21]). Because self-splicing introns can occur at any time, even in the very early stage of origin of life, we consider only spliceosome-type introns. For brevity, we hereafter call this type of introns as simply "intron." There are mainly two major hypotheses: introns early and introns late. The former claims that exon existed as a functional unit from the common ancestor of prokaryotes and eukaryotes, and "exon shuffling" was proposed for creating new protein functions [14]. Introns which separate exons should also be quite an ancient origin [14, 22]. In contrast, introns are considered to emerge only in the eukaryotic lineage according to the introns-late hypothesis [23, 24].

The protein "module" hypothesis proposed by Go [25] is related to be introns-early hypothesis. Pattern of intron appearance and loss has been estimated by various methods (e.g., [21, 26]). Kenmochi and his colleagues analyzed introns of ribosomal proteins of mitochondrial genomes and eukaryotic nuclear genomes in details [27–29]. These studies supported the introns-late hypothesis, because introns in mitochondrial and cytosolic ribosomal proteins seem to be independent

origins and introns seem to emerge in many ribosomal protein genes after eukaryotes appeared.

### 8.3.3 Functional Regions in Introns

Introns do not code for amino acid sequences by definition. In this sense, most of introns may be classified as junk DNAs (see the next section). There are, however, evolutionarily conserved regions in introns, suggesting the existence of some functional roles in introns.

## 8.4 Junk DNAs

Ohno (1972; [30]) proclaimed that the most part of mammalian genomes are nonfunctional and coined the term "junk DNA." With the advent of eukaryotic genome sequence data, it is now clear that he was right. There are in fact so much junk DNAs in eukaryotic genomes. Junk DNAs or nonfunctional DNAs can be divided into repeat sequences and unique sequences. Repeat sequences are either dispersed type or tandem type. Unique sequences include pseudogenes that keep homology with functional genes.

### 8.4.1 Dispersed Repeats

Prokaryote genomes sometimes contain insertion sequences; however, this kind of dispersed repeats constitutes the major portion of many eukaryotic genomes. These interspersed elements are divided into two major categories according to their lengths: short ones (SINEs) and long ones (LINEs).

One well-known example of SINE is Alu elements in primate genomes. It is about 300-bp length, and originated from 7SL ribosomal RNA gene. Let us see the real Alu element sequence from the human genome sequence. If we retrieve the DDBJ/EMBL/GenBank International Sequence Database accession number AP001720 (a part of chromosome 21), there are 128 Alu elements among the 340-kb sequence. The density is 0.38 Alu elements per 1 kb. If we consider the whole human genome of ~3 billion bp, Alu repeats are expected to exist in ~1.13 million copies. One example of Alu sequence is shown below from this entry coordinates from 133600 to 133906:

```
ggcgggagcg atggctcacg cctgtaatgc cagcactttg ggaggccgag
gtgggtggat cacaaggtca ggagatagag accatcctgg ctaacacggt
gaaacactgt ctctactaaa aacacaaaaa actagccagg cgtggtggcg
ggtgcctgta atcccagcta ctcgggaggc tgaggcagga gaatggtgtg
aacccaggaa gtggagcttg cagtgagctc agattgcgcc actgcactcc
agcctgggtg acagagtgag actccatctc aaaaaaaata aaataaataa
aaaaaa
```

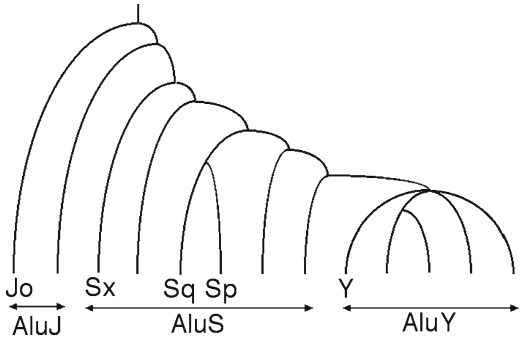


**Fig. 8.6** An overall pattern of Alu element evolution (From Saitou 2007; [103])

If we do BLAST homology search (see Chap. 14) using DDBJ system (http://blast.ddbj.nig.ac.jp/blast/blastn) targeted to nonhuman primate sequences (PRI division of DDBJ database), the best hit was obtained from chimpanzee chromosome 22, which is orthologous to human chromosome 21. I suggest interested readers to do this homology search practice.

Alu elements were first classified into J and S subfamilies [31]. It is not clear about the reason of selection of two characters (J and S), but probably two authors (Jurka and Smith) used initials of their surnames. In any case, this division was based on the distance from Alu consensus sequence; Alu elements which are more close to the consensus were classified as S and those not as J. Later, a subset of the S subfamily were found to be highly similar with each other, and they were named as Y after 'young," for they appeared relatively in young or recent age. Rough estimates of the divergence time of Alu elements are as follows: J subfamily appeared about 60 million years ago, and S subfamily separated from J at 44 million years ago, followed by further separation of Y at 32 million years ago [32]. Figure 8.6 shows the overall pattern of Alu element evolution (based on [32]).

### 8.4.2 Tandem Repeats

Tandemly repeated sequences are also abundant in eukaryotic genomes, and the representative ones are heterochromatin regions. Heterochromatins are highly condensed nonfunctional regions in nuclear DNA, in contrast to euchromatins, in which many genes are actively transcribed. Heterochromatins usually reside at teromeres, terminal parts of chromosomes, and at centromeres, internal parts of chromosomes, that connect spindle fibers during cell division. A more than 1-Mb teromeric regions of *Arabidopsis thaliana* were found to be tandem repeats of ca. 180-bp repeat unit [33, 34]. The nucleotide sequence below is *Arabidopsis thaliana* tandemly repeated sequence AR12 (International Sequence Database accession number X06467):

```
aagcttcttc ttgcttctca atgctttgtt ggtttagccg aagtccatat
gagtctttgt ctttgtatct tctaacaagg aaacactact taggctttta
ggataagatt gcggtttaag ttcttatact taatcataca catgccatca
agtcatattc gtactccaaa acaataacc
```

The human genome also has a similar but nonhomologous sequence in centromeres, called "alphoid DNA" with the 171-bp repeat unit [35]. The following is the sequence (International Sequence Database accession number M21746):

```
catcctcaga aacttctttg tgatgtgtgc attcaagtca cagagttgaa
cattcccttt cgtacagcag tttttaaaca ctctttctgt agtatctgga
agtgaacatt aggacagctt tcaggtctat ggtgagaaag gaaatatctt
caaataaaaa ctagacagaa g
```

If we do BLAST homology search (see Chap. 13) targeted to the human genome sequences of the NCBI database, there was no hit with this alphoid sequence. This clearly shows that the human genome sequences currently available are far from complete, for they do not include most of these tandem repeat sequences.

Telomores of the human genome are composed of hundreds of 6-bp repeats, ttaggg. If we search the human genome as 36-bp long 6 tandem repeats of this 6-repeat units as query using the NCBI BLAST, many hits are obtained.

### 8.4.3 Pseudogenes

As we already discussed in Chap. 4, authentic pseudogenes have no function, and they are genuine members of junk DNAs. When a gene duplication occurs, one of two copies often become a pseudogene. Because gene duplication is prevalent in eukaryote genomes, pseudogenes are also abundant. Pseudogenes are, by definition, homologous to functional genes. However, after a long evolutionary time, many selectively neutral mutations accumulate on pseudogenes, and eventually they will lose sequence homology with their functional counterpart. There are many unique sequences in eukaryote genomes, and majority of them may be this kind of homology-lost pseudogenes.

### 8.4.4 Junk RNAs and Junk Proteins

A long RNA is initially transcribed from a genomic region having an exon–intron structure, and then RNAs corresponding to introns are spliced out. These leftover RNAs may be called "junk" RNAs, for they will soon be degraded by RNAse. Only a limited set of genes are transcribed in each tissue of multicellular organisms, but leaky expression of some genes may happen in tissues in which these genes should not be expressed. Again these are "junk" RNAs, and they are swiftly decomposed. A series of studies (e.g., [36, 37]) claimed that many noncoding DNA regions are transcribed. However, van Bakel et al. [38] showed that most of them were found to be artifact of chip-chip technologies used in these studies. If nonsense or frameshift mutations occur in a protein coding sequences, that gene cannot make proteins. Yet its mRNA may be produced continuously until the promoter or its enhancer will become nonfunctional. In this case, this sort of mutated genes produces junk RNAs.

If only a small quantity of RNAs are found from cells and when they are not evolutionarily conserved, they are probably some kind of junk RNAs.

As junk DNAs and junk RNAs exist, cells may also have "junk" proteins. If mature mRNAs are not produced in the expected way, various aberrant mRNA molecules will be produced, and ribosomes try to translate them to peptides based on these wrong mRNA information. Proteins produced in this way may be called "junk" proteins, for they often have no or little functions. Even if one protein is correctly translated and is moved to its expected cellular location, it can still be considered as "junk" protein. One good example is the ABCC11 transporter protein of dry-type cerumen (earwax), for one nonsynonymous substitution at this gene caused that protein to be essentially nonfunctional [39].

## 8.5 Evolution of Eukaryote Genomes

There are various genomic features that are specific to eukaryotes other than existence of introns and junk DNAs, such as genome duplication, RNA editing, C-value paradox, and the relationship between genome size and mutation rates. We will briefly discuss them in this section.

### 8.5.1 Genome Duplication

The most dramatic and influential change of the genome structure is genome duplications. Genome duplications are also called polyploidization, but this term is tightly linked to karyotypes or chromosome constellation.

Prokaryotes are so far not known to experience genome duplications, which are restricted to eukaryotes. Interestingly, genome duplications are quite frequent in plants, while it is relatively rare in the other two multicellular eukaryotic lineages. An ancient genome duplication was found from the genome analysis of baker's yeast [40], and *Rhizopus oryzae*, a basal lineage fungus, was also found to experience a genome duplication [41]. Among protists, *Paramecium tetraurelia* is known to have experienced at least three genome duplications [42]. Because we human belongs to vertebrates and the two-round genome duplications occurred at the common ancestor of vertebrates (see Chap. 9), we may incline to think that genome duplications often happen in many animal species. It is not the case. So far, only vertebrates and some insects are known to experience genome duplications. The reason for this scattered distribution of genome duplication occurrences is not known.

If we plot the number of synonymous substitutions between duplogs in one genome, it is possible to detect a relatively recent genome duplication. This is because all genes duplicate when a genome duplication occurs, while only a small number of genes duplicate in other modes of gene duplications (see Chap. 3). Figure 8.7 shows the schematic view of two cases: with and without genome duplication. Lynch and Conery (2000; [44]) used this method to various genome sequences and found that the *Arabidopsis thaliana* genome showed a clear peak indicative of relatively recent genome duplication, while the genome sequences of

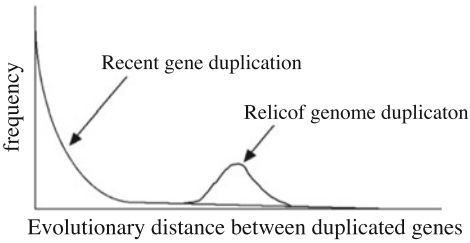


**Fig. 8.7** A schematic view of synonymous distance distribution of duplogs with and without genome duplication (From Saitou 2007; [103])

nematode *Caenorhabditis elegans* and yeast *Saccharomyces cerevisiae* showed the curves of exponential reduction. It is interesting to note that before the genome sequence was determined, the genome duplication was not known for *Arabidopsis thaliana,* while the genome of *Saccharomyces cerevisiae* was later shown to be duplicated long time ago [40].

When genome duplications occurred in some ancient time, the number of synonymous substitutions may become saturated and cannot give appropriate result. In this case, the number of amino acid substitutions may be used, even if each protein may have varied rates of amino acid substitutions. In any case, accumulation of mutations will eventually cause two homologous genes to become not similar with each other. Therefore, although the possibility of genome duplications in prokaryotes are so far rejected [45], it is not possible to infer the remote past events simply by searching sequence similarity. We should be careful to reach the final conclusion.

### 8.5.2 RNA Editing

Modification of particular RNA molecules after they are produced via transcription is called RNA editing. All three major RNA molecules (mRNA, tRNA, and rRNA) may experience editing [46]. There are various patterns of RNA editing; substitutions, in particular between C and U, and insertions and deletions, particularly U, are mainly found in eukaryote genomes. Guide RNA molecules exist in one of the main RNA editing mechanisms, and they specify the location of editing, but there are some other mechanisms [47].

It is not clear how the RNA editing mechanism evolved. Tillich et al. [47] studied chloroplast RNA editing and concluded that suddenly many nucleotide sites of chloroplast DNA genome started to have RNA editing, but later the sites experiencing RNA editing constantly decreased via mutational changes. They claimed that there was no involvement of RNA editing on gene expression. This result does not give RNA editing a positive significance.

Because there are many types of RNA molecules inside a cell, there also exist many sorts of enzymes that modify RNAs. It may be possible that some of them suddenly started to edit RNAs via a particular mutation. RNA editing which did not cause deleterious effects to the genome may have survived by chance at the initial phase. This view suggests the involvement of neutral evolutionary process in the evolution of RNA editing.

### 8.5.3 C-Value Paradox

Organisms with complex metabolic pathways have many genes. Multicellular organisms are such examples. Generally speaking, their genome sizes are expected to be large. In contrast, viruses whose genomes contain only a handful of genes have small genome sizes. Therefore, their possibility of genome evolution is rather limited. Even if amino acid sequences are rapidly changing because of high mutation rates, the protein function may not change. Unless the gene number and genome size increase, viruses cannot evolve their genome structures. It is thus clear that the increase of the genome size is crucial to produce the diversity of organisms. However, genomes often contain DNA regions which are not indispensable. Organisms with large genome sizes have many such junk DNA regions. Because of their existence, the genome size and the gene number are not necessarily highly correlated. This phenomenon was historically called C-value paradox (e.g., [48]), after the constancy of the haploid DNA amount for one species was found, yet their values were found to vary considerably among species at around 1950 (e.g., [49–51]). "C-value" is the amount of haploid DNA, and C probably stands as acronym of "constant" or "chromosomes." We now know that the majority of eukaryote genome DNA is junk, and there is no longer a paradox in C-values among species.

### 8.5.4 Conserved Noncoding Regions

While bacterial genomes are mostly consisting of protein coding genes, a considerable region of eukaryote genomes is noncoding. Most of them are junk DNA and do not have functions. If we find evolutionary conservation, however, these conserved regions should have some function through purifying selection. From the initial stage of molecular evolutionary studies, protein noncoding regions were suspected to be involved in gene regulation (Zuckerkandl and Pauling 1965; Britten and Davidson 1971; King and Wilson 1975). Now it is becoming clear that at least some noncoding regions play important roles in gene regulation (e.g., Carroll 2005; [55]). The functional elements are expected to evolve more slowly than surrounding nonfunctional DNA, as they are under purifying selection. Therefore, conserved noncoding sequences (CNSs) are likely to be important from the functional point of view.

Animal CNSs were discovered by comparison of human and fugu fish genome sequences by How et al. (1996; [52]). CNS analyses have been proved to be powerful for detecting regulatory elements (e.g., Hardison 2000; [53], Levy et al. 2001; [54]). Bejarano et al. (2004; [102]) found highly conserved noncoding sequences through comparison of human, mouse, and rat genomes. Siepel et al. (2005; [56]) found conserved noncoding DNA sequences from insects, nematodes, and yeasts by comparing closely related species. We will discuss more on conserved noncoding sequences of vertebrates in Chap. 9.

As for plants, Kaplinsky et al. (2002; [57]) found six short (<60 bp) CNSs from seven DNA regions related to protein coding gene orthologs between rice and maize. Guo et al. (2003; [58]) identified 20 bp as the minimal criterion for a CNS in

grasses. Inada et al. (2003; [59]) examined 3,000 bases upstream and downstream of 52 orthologous protein coding genes of rice and maize and found that most CNSs were less than 20 bases. Thomas et al. (2007; [60]) compared *Arabidopsis thaliana* paralogous sequences, and found 14,944 intronic conserved noncoding sequences, ranging their lengths from 15 to 285 bp. D'Hont et al. (2012; [61]) determined banana genome and found 116 CNSs from genome sequences of commelinid monocot (banana, palm, and grasses). Kristas et al. (2012; [62]) compared genome sequences of Arabidopsis, grape rice, and Brachypodium and found >100 times more abundant CNSs from monocots than dicots. Hettiarachchi and Saitou; [63] compared genome sequences of 15 plant species and searched lineage-specific CNSs. They found 2 and 22 CNSs shared by all vascular plants and angiosperms, respectively, and also confirmed that monocot CNSs are much more abundant than those of dicots.

### 8.5.5 Mutation Rate and Genome Size

What kind of the relationship exists between the genome size and mutation rates? If all the genetic information contained in the genome of one organism are necessary for survival of that organism, the individual will die even if only one gene of its genome lost its function by a mutation. An organism with a small genome size and hence with a small number of genes, such as viruses, can survive even if the mutation rate is high. In contrast, organisms with many genes may not be able to survive if highly deleterious mutations often happen. Therefore, such organisms must reduce the mutation rate.

Rajic et al. (2005; [43]) compared the rate of synonymous substitutions per year from virus to human and the protein coding region size and found a clear negative correlation, as shown in Fig. 8.8. Sunjan et al. (2010; [64]) compared many studies on viral mutations and found a clear negative correlation between the substitution type mutation rate per nucleotide site per cell infection and viral genome size.

However, when the nucleotide substitution type mutation rate per generation was compared with the whole-genome size, Lynch (2006; [65]) found a positive correlation. More recently, Lynch (2010; [66]) admitted that for organisms with small-sized genomes, these two values were in fact negatively correlated. However, when large-genome-sized eukaryotes are compared, now a positive correlation was observed.

We have to be careful when we discuss these two contradictory reports. One considered the rate using unit as physical year, while the other used one generation as the unit. Another difference is to use either only protein coding gene region DNA sizes or the whole-genome sizes. The relationship between the mutation rate and genome size is not simple. Drake et al. (1998; [67]) examined this problem and found that the mutation rate per genome per replication was approximately 1/300 for bacteria, while mutation rates of multicellular eukaryotes vary between 0.1 and 100 per genome per sexual or individual generation. Table 8.4 shows the list of the mutation rate and the genome size for various organisms. Apparently there is no clear tendency.

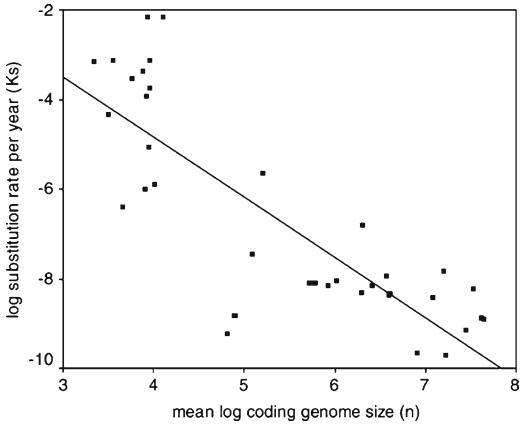


**Fig. 8.8** A negative correlation between the rate of synonymous substitutions and the protein-coding region size (From Rajic et al. 2005; [43])

**Table 8.4** Mutation rates and genome sizes of various organisms

| Organism | Genome | Mutation rate ($\times 10^{-9}$) per | | |
|---|---|---|---|---|
| (Organism group) | Size (bp) | Year | Generation | Reference |
| By direct method: | | | | |
| Human | $3.2\times10^{9}$ | ~0.4 | 11–12 | 1–4 |
| *Drosophila* | $1.7\times10^{8}$ | 35 | 3.5 | 5 |
| *C. elegans* | $8.0\times10^{7}$ | – | 2.7 | 6 |
| *Neurospora* | $4.2\times10^{7}$ | – | 0.072 | 7 |
| Baker's yeast | $1.2\times10^{7}$ | – | 0.22 | 7 |
| *E. coli* | $4.6\times10^{6}$ | 50 | 0.5 | 8 |
| Phage T2,T4 | $1.7\times10^{5}$ | – | 24 | 7 |
| Phage λ | $4.9\times10^{4}$ | – | 77 | 7 |
| mtDNA (*C. elegans*) | $1.5\times10^{4}$ | – | 160 | 9 |
| Phage M13 | $6.4\times10^{3}$ | – | 720 | 7 |
| By indirect method: | | | | |
| Human-Chimpanzee | $3\times10^{9}$ | 1.0 | 15 | 10 |
| Mouse-Rat | $3\times10^{9}$ | 5.3 | 5.3 | 11 |
| *E. coli-Salmonella* | $4\times10^{6}$ | 4.5 | 0.04 | 7 |
| mtDNA (Plants) | $4\times10^{5}$ | 0.34 | – | 12 |
| mtDNA (Mammals) | $1.7\times10^{4}$ | 34 | – | 12 |
| mtDNA (Birds) | $1.7\times10^{4}$ | 17 | – | 12 |
| RNA virus | ~$10^{4}$ | ~$10^{6}$ | ~$10^{4}$ | 13 |

1: Roach, J. C., Glusman, G., Smit, A. F., Huff, C. D., Hubley, R., Shannon, P. T., Rowen, L., Pant, K. P., Goodman, N., Bamshad, M., Shendure, J., Drmanac, R., Jorde, L .B., Hood, L., & Galas, D. J. (2010). *Science*, *328*, 636–639

2: Conrad, D. F., et al. (2011). Variation in genome-wide mutation rates within and between human families. *Nature Genetics*, *43*, 712–715

3: Kong, A., et al. (2012). Rate of de novo mutations and the importance of father's age to disease risk. *Nature*, *488*, 471–475

4: Campbel, C. D. (2012). Estimating the human mutation rate using autozygosity in a founder population. *Nature Genetics*, *44*, 1277–1283

5: Keightley, P. D., Trivedi, U., Thomson, M., Oliver, F., Kumar, S., & Blaxter, M. L. (2009). Analysis of the genome sequences of three *Drosophila melanogaster* spontaneous mutation accumulation lines. *Genome Research*, *19*, 1195–1201

(continued)

**Table 8.4** (continued)

6: Denver, D. R., Dolan, P. C., Wilhelm, L. J., Sung, W., Lucas-Lledó, J. I., Howe, D. K., Lewis, S. C., Okamoto, K., Thomas, W. K., Lynch, M., & Baer, C. F. (2009). A genome-wide view of *Caenorhabditis elegans* base-substitution mutation processes. *Proceedings of the National Academy of Sciences of the United states of America*, *106* 16310–16314
7: Drake, J. W., Charlesworth, B., Charlesworth, D., & Crow, J. F. (1998). Rates of spontaneous mutation. *Genetics*, *148*, 1667–1686
8: Ochman, H. (2003). Neutral mutations and neutral substitutions in bacterial genomes. *Molecular Biology and Evolution*, *20*, 2091–2096
9: Denver, D. R., Morris, K., Lynch, M., & Thomas, W. K. (2004). High mutation rate and predominance of insertions in the *Caenorhabditis elegans* nuclear genome. *Nature*, *430*, 679–682
10: Fujiyama, A., Watanabe, H., Toyoda, A., Taylor, T. D., Itoh, T., Tsai, S.-F., Park, H.-S., Yaspo, M.-L., Lehrach, H., Chen, Z., Fu, G., Saitou, N., Osoegawa, K., de Jong, P. J., Suto, Y., Hattori, M., & Sakaki, Y. (2002). Construction and analysis of a human-chimpanzee comparative clone map. *Science*, *295*(5552), 131–134
11: Abe, K., Noguchi, H., Tagawa, K., Yuzuriha, M., Toyoda, A., Kojima, T., Ezawa, K., Saitou, N., Hattori, M., Sakaki, Y., Moriwaki, K., & Shiroishi, T. (2004). Contribution of Asian mouse subspecies Mus musculus molossinus to genomic constitution of strain C57BL/6J, as defined by BAC end sequence-SNP analysis. *Genome Research*, *14*, 2239–2247
12: Lynch, M., Koskella, B., & Schaack, S. (2006). Mutation pressure and the evolution of organelle genomic architecture. *Science*, *311*, 1727–1730
13: Hanada, K., Suzuki, Y., & Gojobori, T. (2004). A large variation in the rates of synonymous substitution for RNA viruses and its relationship to a diversity of viral infection and transmission modes. *Molecular Biology and Evolution*, *21*(6), 1074–1080

## 8.6 Genome of Multicellular Eukaryotes

We will discuss genomes of three multicellular lineages of eukaryotes: plants, fungi, and animals in this section. Unfortunately, there seems to be no common feature of genomes of multicellular organisms, so each lineage is discussed independently.

### 8.6.1 Plant Genomes

*Arabidopsis thaliana* was the first plant species whose 125-Mb genome was determined in 2000 [68]. A. *thaliana* is a model organism for flowering plants (angiosperms), with only 2-month generation time. In spite of its small genome size, only 4 % of the human genome, it has 32,500 protein coding genes. The genome sequence of its closely related species, *A. lyrata*, was also recently determined [69].

Angiosperms are divided into monocots and dicots. A. *thaliana* is a dicot, and genome sequences of six more species were determined as of December 2013 (see Table 8.5).

Rice, *Oryza sativa*, is a monocot, and its genome size, 370 ~ 410 Mb, is much smaller than that of the wheat genome. Its japonica and indica subspecies genomes were determined [70] and [71], and the origin of rice domestication is currently in great controversy, particularly in single or multiple domestication events (e.g., [72, 73]). The number of protein coding genes in the rice genome is 37,000 ~ 40,000 [74].

**Table 8.5** List of plant species whose genome sequences were determined

| Species name | English common name | Genome size (Mb) | Reference |
|---|---|---|---|
| Dicots: | | | |
| *Arabidopsis thaliana* | Thale Cress | 135 | 1 |
| *Brassica rapa* | Turnip mustard | 273 | 2 |
| *Cucumis sativus* | Cucumber | 203 | 3 |
| *Ricinus communis* | Castor bean | 400 | 4 |
| *Populus tricocarpa* | Cottonwood | 422 | 5 |
| *Vitis vinifera* | Grape | 487 | 6 |
| *Aquilegia coerulea* | Blue columbine | 293 | Unpublished |
| Monocots: | | | |
| *Oryza sativa japonica* | Rice (japonica variety) | 372 | 7 |
| *Brachypodium distachyon* | Purple false brome | 272 | 8 |
| *Setaria italica* | Foxtail millet | 405 | 9 |
| *Sorghum bicolor* | Sorghum | 697 | 10 |
| *Musa acuminata* | Banana | 523 | 11 |
| *Phyllostachys heterocycla* | Bamboo | 2,000 | 12 |
| Non-seed plants: | | | |
| *Selaginella moellendorffii* | Spikemoss | 212 | 13 |
| *Physcomitrella patens* | Moss | 480 | 14 |
| *Chlamydomonas reinhardtii* | *Chlamydomonas* | 120 | 15 |

References

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

2: The Brassica rapa Genome Sequencing Project Consortium. (2011). The genome of the mesopolyploid crop species *Brassica rapa*. *Nature Genetics*, *43*, 1035–1039

3: Huang, S., et al. (2009). The genome of the cucumber, *Cucumis sativu*s L. *Nature Genetics, 41*(12), 1275–1281

4: Chan, A. P., et al. (2010). Draft genome sequence of the oilseed species *Ricinus communis*. *Nature Biotechnology*, *28*, 951–956

5: Tuskan, G., et al. (2006). The Genome of black cottonwood, Populus trichocarpa. *Science*, *313*(5793), 1596–1604

6: Jaillon, O., et al. (2007). The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature*, *449*, 463–467

7: Goff, S. A., et al. (2002). A draft sequence of the rice genome (*Oryza sativa* L. ssp. japonica). *Science*, *296*, 92–100

8: Vogel, J. P., et al. (2010). Genome sequencing and analysis of the model grass *Brachypodium distachyon*. *Nature*, *463*, 763–768

9: Zhang, G., et al. (2012). Genome sequence of foxtail millet (*Setaria italica*) provides insights into grass evolution and biofuel potential. *Nature Biotechnology*, *30*, 549–554

10: Paterson, A. H., et al. (2009). *Nature, 457,* 551–556

11: D'Hont, A., et al. (2012). The banana (*Musa acuminata*) genome and the evolution of monocotyledonous plants. *Nature*, *488*, 213–217

12: Peng, Z., et al. (2013). The draft genome of the fast-growing non-timber forest species moso bamboo (*Phyllostachys heterocycla*). *Nature Genetics.* doi:10.1038/ng.2569

13: Banks, J. A., et al. (2011). The *Selaginella* genome identifies genetic changes associated with the evolution of vascular plants. *Science*, *332*, 960–963

14: Rensing, S. A., et al. (2007). The *Physcomitrella* genome reveals evolutionary insights into the conquest of land by plants. *Science*, *319*, 64–69

15: Merchant, S. S., et al. (2007). The *Chlamydomonas* genome reveals the evolution of key animal and plant functions. *Science*, *318*, 254–250

Wheat corresponds to genus *Triticum*, and there are many species in this genus. The typical bread wheat is *Triticum aestivum*, and it is a hexaploid with 42 (7×6) chromosomes. Its genome arrangement is conventionally written as AABBDD [75]. Because it is now behaving as diploid, genomic sequencing of 21 chromosomes (A1–A7, B1–B7, and D1–D7) is under way (see http://www.wheatgenome.org/ for the current status). The hexaploid genome structure emerged by hybridization of diploid (DD) cultivated species *T. durum* and tetraploid (AABB) wild species *Aegilops tauschii* [75]. A genome duplication followed hybridization.

Non-seedling land plants are ferns, lycophytes, and bryophytes, in the order of closeness to seed plants (e.g., [76]). A draft genome sequence of a moss, *Physcomitrella patens* was reported in 2008 [77], followed by genome sequencing of a lycophyte, *Selaginella moellendorffii,* in 2011 [78]. These genome sequences of different lineages of plants are deciphering stepwise evolution of land plants.

### 8.6.2 Fungi Genomes

The genome sequence of baker's yeast (*Saccharomyces cerevisiae*) was determined in 1996, as the first eukaryotic organism [79]. There are 16 chromosomes in *S. cerevisiae,* and its genome size is about 12 Mb. There are a total of 8,000 genes in its genome: 6,600 ORFs and 1,400 other genes. The genome-wide GC content is 38 %, slightly lower than that of the human genome. The proportion of introns is very small compared to that of the human genome, and the average length of one intron is only 20 bp, in contrast to the 1,440-bp average length of exons [80]. As we already discussed, the ancestral genome of baker's yeast experienced a genome-wide duplication [40]. Pseudogenes, which are common in vertebrate genomes, are rather rare in the genome of baker's yeast; they constitute only 3 % of the protein coding genes [80]. The baker's yeast is often considered as the model organisms for all eukaryotes; however, their genome may not be a typical eukaryote genome.

As of December 2013, genome sequences of more than 400 fungi species are available (see NCBI genome list at http://www.ncbi.nlm.nih.gov/genome/browse/ for the present situation). Figure 8.9 shows the relationship between the genome size and gene numbers for 88 genomes. There is a clear positive correlation between them. However, there are some outliers. The Perigord black truffle (*Tuber melanosporum*), shown as A in Fig. 8.9, has the largest genome size (~125 Mb) among the 88 fungi species whose genome sequences were so far determined, yet the number of genes is only ~7,500 [81].

Three other outlier species are *Postia placenta*, *Ajellomyces dermatitidis*, and *Melampsora laricipopulina*, shown as B, C, and D in Fig. 8.9, respectively. Interestingly, these four outlier species are phylogenetically not clustered well; two are belonging to Pezizomycotina of Ascomycota and the other two are Agaricomycotina and Pucciniomycotina of Basidiomycota. If we exclude these four outlier species, a good linear regression is obtained, as shown in Fig. 8.9. This straight line indicates that in average, one gene size corresponds to 2.9 kb in a typical fungi genome. If we apply this average gene size to the truffle genome, its genome

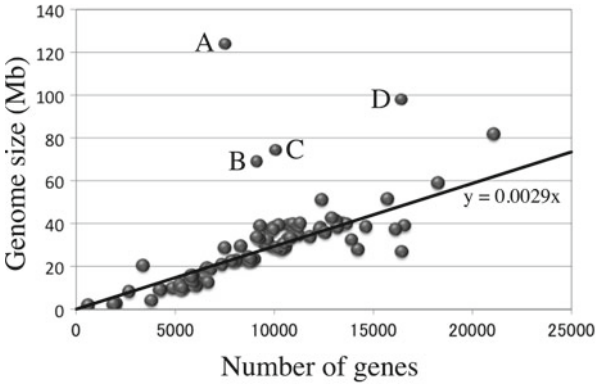


**Fig. 8.9** The relationship between the genome size and gene numbers among 88 fungi genomes

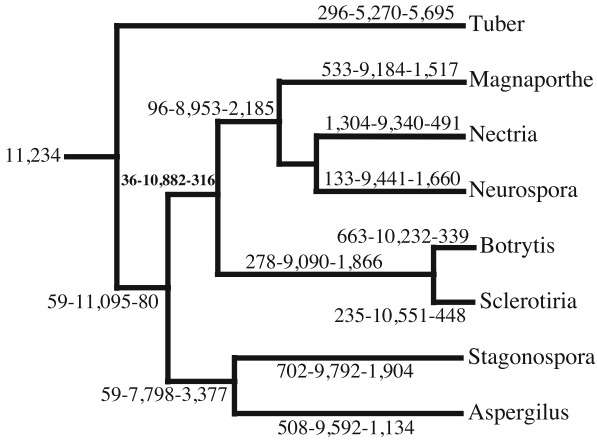


**Fig. 8.10** Gain and loss of genes in each branch of the phylogenetic tree for fungi species (Based on [81])

size should be ~22 Mb, but the real size is 103 Mb larger. This suggests that there is unusually large number of junk DNA in this genome. In fact, 58 % of its genome consists of transposable elements [81]. The truffle genome must still have 24 % more junk DNA region. Gain and loss of genes in each branch of the phylogenetic tree for fungi species are shown in Fig. 8.10 (based on [81]). It will be interesting to examine genome sizes of species related to the Perigord black truffle, so as to infer the evolutionary period when the genome size expansion occurred.

### 8.6.3 Animal Genomes

Animals, or metazoa, are the most integrated multicellular organisms. Genome sequences of four 35 invertebrate species and 32 vertebrate species were determined by end of 2011 according to the GCDB of Kryukov et al. (2012; [1]). As of December 2013, 35 invertebrate and 43 vertebrate species were deteremined according to KEGG database (http://www.genome.jp/kegg/catalog/org_list.html). A major gene

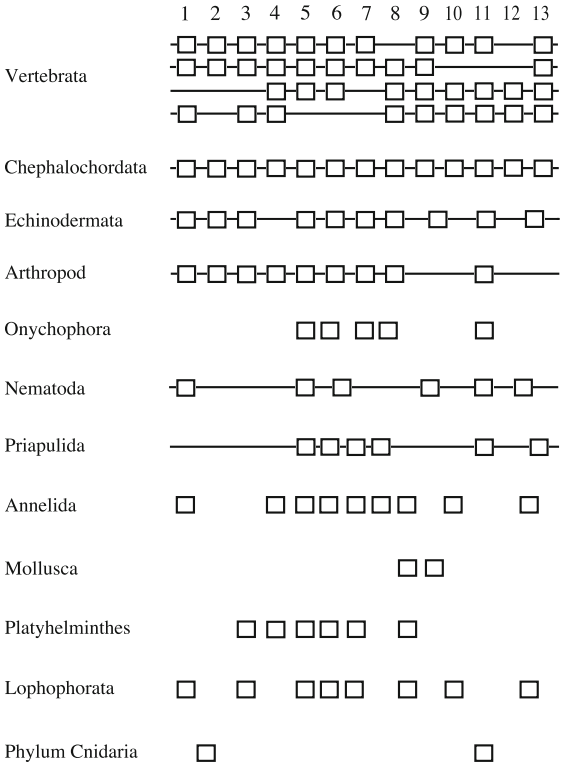


**Fig. 8.11** The Hox gene clusters found in each animal phylum (From Saitou 2007; [103])

system that is responsible for this is Hox genes. We thus first discuss this gene system in this subsection. The genome of *C. elegans*, first determined genome among animals, will be discussed next, followed by genomes of insects and those of deuterostomes. Because genomes of many vertebrate species were determined, we discuss them in Chap. 9, and in particular, on the human genome in Chap. 10.

### Hox Code

Hox genes were initially found through studies of homeotic mutations that dramatically change segmental structure of *Drosophila* by Edward B. Lewis [82]. They code for transcription factors, and a DNA-binding peptide, now called homeobox domain, was later found in almost all animal phyla [83]. Figure 8.11 shows the Hox gene clusters found in 12 animal groups. There are four Hox clusters in mammalian and avian genomes, and they are most probably generated by the two-round genome duplication in the common ancestor of vertebrates (see Chap. 9).

Interestingly, the physical order of Hox genes in chromosomes and the order of gene expression during the development are corresponding, called "collinearity" [84]. This suggests that some sort of cis-regulation is operating in Hox gene clusters, and in fact, many long transcripts are found, and some of their transcription start sites are highly conserved among vertebrates [85]. Figure 8.12 shows highly conserved

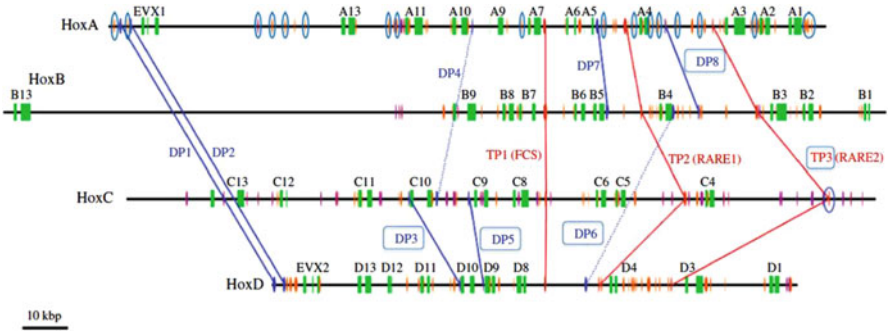


**Fig. 8.12** Highly conserved noncoding sequences found from comparison of Hox A cluster regions of many vertebrate species (From Matsunami et al. 2010; [85])

noncoding sequences found from comparison of Hox A cluster regions of many vertebrate species (from Matsunami et al. 2010; [85]).

The Hox genes control expression of different groups of downstream genes, such as transcription factors, elements in signaling pathways, or genes with basic cellular functions. Hox gene products interact with other proteins, in particular, on signaling pathways, and contribute to the modification of homologous structures and creation of new morphological structures [87].

There are other gene families that are thought to be involved in diverse animal body plan. One of them is the Zic gene family [88]. The Zic gene family exists in many animal phyla with high amino acid sequence homology in a zinc-finger domain called ZF, and members of this gene family are involved in neural and neural crest development, skeletal patterning, and left–right axis establishment. This gene family has two additional domains, ZOC and ZF-BC. Interestingly, Cnidaria, Platyhelminthes, and Urochordata lack the ZOC domain, and their ZF-BC domain sequences are quite diverged compared to Arthropoda, Mollusca, Annelida, Echinodermata, and Chordata. This distribution suggests that the Zic family genes with the entire set of the three conserved domains already existed in the common ancestor of bilateralian animals, and some of them may be lost in parallel in the platyhelminthes, nematodes, and urochordates [88]. Interestingly, phyla that lost ZOC domains have quite distinct body plan although they are bilateralian.

### Genome of *C. elegans*

*Caenorhabditis elegans* was the first animal species whose 97-Mb draft genome sequence was determined in 1998 [89]. This organism belongs to the Nematoda phylum which includes a vast number of species [90]. Brenner (1974; [91]) chose this species as model organism to study neuronal system, for its short generation time (~ 4 days) and its size (~1 mm). Figure 3.3 in Chap. 3 shows the cell genealogy of this species.

The following description of this section is based on the information given in online "WormBook" [86]. There are 22,227 protein coding genes in *C. elegans* including 2,575 alternatively spliced forms, with 79 % confirmed to be transcribed

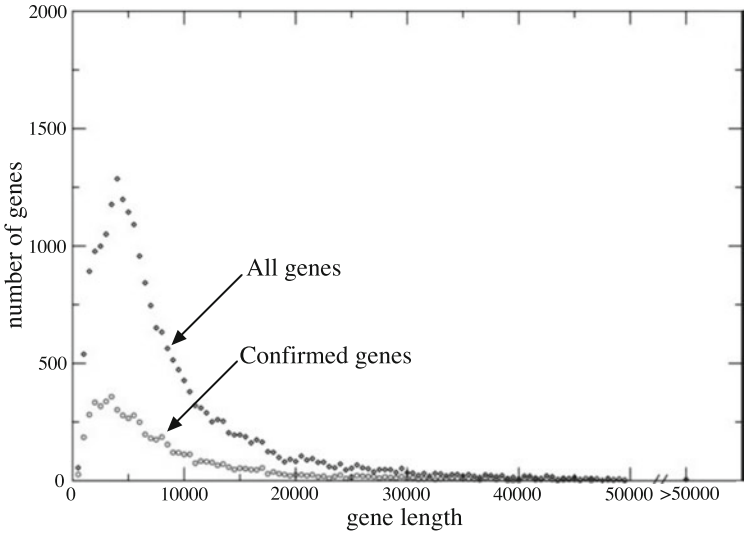


**Fig. 8.13** Distribution of the protein coding genes in the genome of *Caenorhabditis elegans* (From [86])

at least partially. The number of tRNA genes is 608, and 274 are located in X chromosome. The three kinds of rRNA genes (18S, 5.8S, and 26S) are located in chromosome I in 100–150 tandem repeats, while ~100 5S rRNA genes are also in tandem form but located in chromosome V. The average protein coding gene length is 3 kb, with the average of 6.4 coding exons per gene. In total, protein coding exons constitute 25.6 % of the whole genome. Figure 8.13 shows the distribution of the protein coding genes, and Fig. 8.14 the distribution of exon numbers per gene. Both distributions have long tails. The median sizes of exons and introns are 123 bp and 65 bp, respectively. Intron lengths of *C. elegans* are quite short compared to these of vertebrate genes (see Chap. 9). The distribution of protein coding genes varies depending on chromosomes, slightly more dense for five autosomes than X chromosome and more dense in the central region than the edge of one chromosome. Processed, i.e., intronless, pseudogenes are rare, and a total of 561 pseudogenes were reported at the Wormbase version WS133. About half of them are homologous to functional chemoreceptor genes.

Genome sequences of four congeneric species of *C. elegans* (*C. brenneri*, *C. briggsae*, *C. japonica*, and *C. remanei*) were determined (http://www.ncbi.nlm.nih.gov/genome/browse/).

### Insect Genomes

A fruit fly *Drosophila melanogaster* was used by Thomas Hunt Morgan's group in the early twentieth century and has been used for many genetic studies. Because of this importance, its genome sequence was determined at first among Arthropods in 2000 [92]. Heterochromatin regions of ~50 Mb were excluded from sequencing,

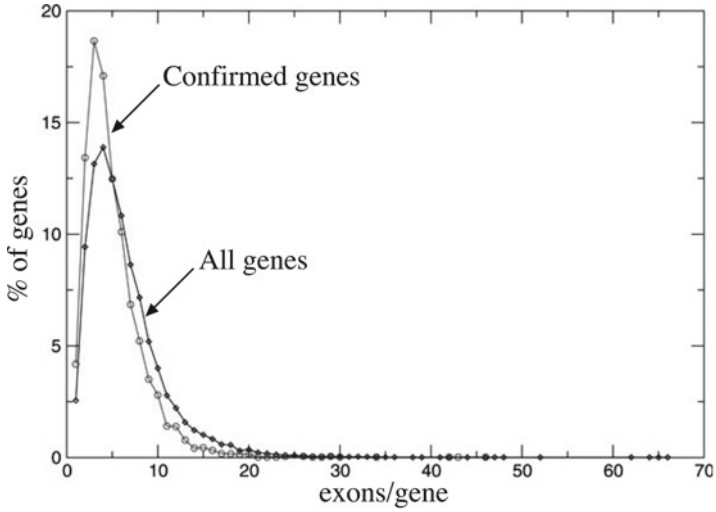


**Fig. 8.14** Distribution of exon numbers per gene in the genome of *Caenorhabditis elegans* (From [86])

and only 120-Mb euchromatin regions were determined. Genome sequences of 12 Drosophila species (D. *ananassae*, *D. erecta*, *D. grimshawi*, *D. melanogaster*, *D. mojavensis*, *D. persimilis*, *D. pseudoobscura*, *D. sechellia*, *D. simulans*, *D. virilis*, *D. willistoni*, and *D. yakuba*) were determined in 2007 [93]. Their genome sizes vary from 145 to 258 Mb, and the number of genes is 15,000–18,000. Interestingly, *D. melanogaster* has the largest genome size and the smallest number of genes.

A total of 12 insect species other than *Drosophila* 12 species were sequenced by end of 2011 [1]. As of December 2013, their genome sizes are in the range of 108 Mb and 540 Mb, more than five times difference, and the gene numbers are from 9,000 to 23,000.

### Genomes of Deuterostomes

Deuterostomes contain five phyla: Echinodermata, Hemichordata, Chaetognatha, Xenoturbellida, and Chordata. The genome of sea urchin *Strongylocentrotus purpuratus* [94] was determined in 2006. Its genome size is 814 Mb with 23,300 genes. Genomes of another sea urchins, *Lytechinus variegatus* and *Patiria miniata*, are also under sequencing, as well as hemicordate *Saccoglossus kowalevskii*.

Chordata is classified into Urochordata (ascidians), Cephalochordata (lancelets or amphioxus), and Vertebrata (vertebrates). Because we will discuss genomes of vertebrates in Chap. 9, let us discuss genomes of ascidians and lancelets only. The genome of ascidian *Ciona intestinalis* was determined in 2002 [95], and the genome sequence of its congeneric species, *C. savignyi*, was also determined three years later [96]. The genome size of *C. intestinalis* is ~155 Mb with ~16,000 genes. Interestingly it contains a group of cellulose synthesizing enzyme genes, which were probably introduced from some bacterial genomes via horizontal gene transfer [8, 97].

The *C. intestinalis* genome also contains several genes that are considered to be important for heart development ([95]), and this suggests that heart of ascidians and vertebrates may be homologous. Through the superimposition of phylogenetic trees (see Chapter A2) for five genes coding muscle proteins, OOta and Saitou ([98]) estimated that vertebrate heart muscle was phylogenetically closer to vertebrate skeletal muscles. If both results are true, muscles used in heart might have been substituted in the vertebrate lineage. The genome sequences of an amphioxus (Cephalochordate *Branchiostoma floridae*) was determined in by Holland et al. (2008; [104]), and they provide good outgroup sequence data for vertebrates.

## 8.7 Eukaryote Virus Genomes

Eukaryotic viruses are relying most of metabolic pathways to their eukaryote host species. Therefore, the number of genes in virus genomes is usually very small. For example, influenza A virus has 8 RNA fragments coding for 11 protein genes, and the total genome size is ~13.6 kb.

As in bacteriophages, there are both DNA type and RNA type genomes in eukaryotic viruses. Table 8.6 shows one example of classification of eukaryotic viruses based on their genome structure [99]. Genomes of double-strand DNA genome viruses have four types: circular, simple linear, linear with proteins covalently attached to both ends, and linear but both ends were closed. Genomes of single-strand DNA genome viruses are either circular or linear.

Genomes of RNA genomes are all linear in both single- and double-strand type. Those of single-strand RNA genomes are classified into two types: plus strand and minus strand. A subset of single-plus strand RNA genome type is experiencing

**Table 8.6** Classification of eukaryotic viruses based on their genome structure (From Sadaie et al. eds. 2004; [99])

| Shape | Example virus |
|---|---|
| DNA genome: | |
| Double strand & circular | SV40, polyomavirus |
| Double strand & linear | T4 bacteriophage, herpes virus |
| Double strand & linear, proteins attached at both ends | Adenovirus, φ29 bacteriophage |
| Double strand & linear, both ends are closed | Poxvirus |
| Single strand & circular | φX174 bacteriophage, M13 bacteriophage |
| Single strand & linear | Parvovirus |
| RNA genome: | |
| Double strand & linear | Reovirus |
| Single plus strand & linear | Tobacco mosaic virus, poliovirus, coronavirus, norovirus, Japanese encephalitis virus |
| Single plus strand & linear, including DNA replication intermediate | Retrovius, human T cell leukemia virus |
| Single minus strand & linear | Rabies virus, measles virus, influenza virus, mumps virus, ebola virus |

DNA intermediate during replication, such as retroviruses and human T-cell leukemia virus (HTLV).

Some DNA genome viruses are unusually large and similar to a small bacterial genome. Megavirus, parasitic to amoeba, has 1.26-Mb genome size and there are 1,120 protein coding genes [100]. Megavirus is phylogenetically close to mimivirus [101], a member of nucleoplasmic large DNA viruses, including pox virus. Recently, a larger genome size virus, Pandoravirus, with more than 2.5-Mb genome, was discovered [105]. The phylogenetic status of these large genome size DNA viruses is unknown at this moment.

## References


1. Kryukov, K., Sumiyama, K., Ikeo, K., Gojobori, T., & Saitou, N. (2012). A new database (GCD) on genome composition for eukaryote and prokaryote genome sequences and their initial analyses. *Genome Biology and Evolution, 4*, 501–512.
2. Andersson, S. G., et al. (1998). The genome sequence of *Rickettsia prowazekii* and the origin of mitochondria. *Nature, 396*, 133–140.
3. Martin, W., & Muller, M. (1998). The hydrogen hypothesis for the first eukaryote. *Nature, 392*, 37–41.
4. Wolstenholme, D. R., & Jeon, K. W. (Eds.) (1992). *Mitochondrial genome*. San Diego: Academic Press.
5. Sakai, M., & Sakaizumi, M. (2012). The complete mitochondrial genome of *Dugesia japonica* (Platyhelminthes; Order Tricladida). *Zoological Science, 29*, 672–680.
6. Sugiyama, Y., Watase, Y., Nagase, M., Makita, M., Yagura, S., Hirai, A., & Sugiura, M. (2005). The complete nucleotide sequence of the tobacco mitochondrial genome: Comparative analysis of mitochondrial genomes in higher plants and multipartite organization. *Molecular and General Genomics, 272*, 603–615.
7. Bergthorsson, U., Adams, K. L., Thomason, B., & Palmer, J. (2003). Widespread horizontal transfer of mitochondrial genes in flowering plants. *Nature, 424*, 197–201.
8. Rodriguez-Moreno, L., Benjak, A., Marti, M. C., Puigdomenech, P., Aranda, M. A., & Garcia-Mas, J. (2011). Determination of the melon chloroplast and mitochondrial genome sequences reveals that the largest reported mitochondrial genome in plants contains a significant amount of DNA having a nuclear origin. *BMC Genomics, 12*, 424.
9. Lilly, J. W., & Havey, M. J. (2001). Small, repetitive DNAs contribute significantly to the expanded mitochondrial genome of cucumber. *Genetics, 159*, 317–328.
10. Shinozaki, K., Ohme, M., Tanaka, M., Wakasugi, T., Hayashida, N., Matsubayashi, T., Zaita, N., Chunwongse, J., Obokata, J., Yamaguchi-Shinozaki, K., Ohto, C., Torazawa, K., Meng, B. Y., Sugita, M., Deno, H., Kamogashira, T., Yamada, K., Kusuda, J., Takaiwa, F., Kato, A., Tohdoh, N., Shimada, H., & Sugiura, M. (1986). The complete nucleotide sequence of the tobacco chloroplast genome: Its gene organization and expression. *EMBO Journal, 5*, 2043–2049.
11. Oldenburg, D. J., & Bendich, A. J. (2004). Changes in the structure of DNA molecules and the amount of DNA per plastid during chloroplast development in maize. *Journal of Molecular Biology, 344*, 1311–1330.
12. Woischnik, M., & Moraes, C. T. (2002). Pattern of organization of human mitochondrial pseudogenes in the nuclear genome. *Genome Research, 12*, 885–893.
13. The Nobel Prize in Physiology or Medicine. (1993). (http://www.nobelprize.org/nobel_prizes/medicine/laureates/1993/press.html)
14. Gilbert, W. (1978). Why genes in pieces? *Nature, 271*, 501.
15. Kenmochi, N. (2012). Introns. In *Encyclopedia of evolution*. Tokyo: Kyoritsu Shuppan (in Japanese).

16. Sheth, N., Roca, X., Hastings, M. L., Roeder, T., Krainer, A. R., & Sachidanandam, R. (2006). Comprehensive splice-site analysis using comparative genomics. *Nucleic Acids Research, 34*, 3955–3967.
17. Tycowski, K. T., Kolev, N. G., Conrad, N. K., Fok, V., & Steitz, J. A. (2006). The ever-growing world of small nuclear ribonucleoproteins. In R. F. Gesteland, T. R. Cech, & J. F. Atkins (Eds.), *The RNA World*, 3rd ed. (pp. 327–368). Woodbury: Cold Spring Harbor Laboratory Press
18. Cavalier-Smith, T. (1991). Intron phylogeny: A new hypothesis. *Trends in Genetics, 7*, 145–148.
19. Marck, C., & Grosjean, H. (2002). tRNomics: Analysis of tRNA genes from 50 genomes of Eukarya, Archaea, and Bacteria reveals anticodon-sparing strategies and domain-specific features. *RNA, 8*, 1189–1232.
20. Koonin, E. V. (2006). The origin of introns and their role in eukaryogenesis: A compromise solution to the introns-early versus introns-late debate? *Biology Direct, 1*, 22.
21. Roy, S. W., & Gilbert, W. (2006). The evolution of spliceosomal introns: Patterns, puzzles and progress. *Nature Reviews Genetics, 7*, 211–221.
22. Doolittle, W. F. (1978). Genes in pieces: Were they ever together? *Nature, 272*, 581–582.
23. 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. *Journal of Cell Science, 34*, 247–278.
24. Logsdon, J. M., Jr. (1998). The recent origins of spliceosomal introns revisited. *Current Opinion in Genetics & Development, 8*, 637–648.
25. Go, M. (1981). Correlation of DNA exonic regions with protein structural units in haemoglobin. *Nature, 291*, 90–92.
26. Rogozin, I. B., Wolf, Y. I., Sorokin, A. V., Mirkin, B. G., & Koonin, E. V. (2003). Remarkable interkingdom conservation of intron positions and massive, lineage-specific intron loss and gain in eukaryotic evolution. *Current Biology, 13*, 1512–1517.
27. Nguyen, D. H., Yoshihama, M., & Kenmochi, N. (2005). New maximum likelihood estimators for eukaryotic intron evolution. *PLoS Computational Biology, 1*, e79.
28. Yoshihama, M., Nakao, A., Nguyen, H. D., & Kenmochi, N. (2006). Analysis of ribosomal protein gene structures: Implications for intron evolution. *PLoS Genetics, 2*, 237–242.
29. Yoshihama, M., Nguyen, H. D., & Kenmochi, N. (2007). Intron dynamics in ribosomal protein genes. *PLoS One, 1*, e141.
30. Ohno, S. (1972). So much "junk" DNA in our genome. *Brookhaven Symposium in Biology, 23*, 366–370.
31. Jurka, J., & Smith, T. (1988). A fundamental division in the Alu family of repeated sequences. *Proceedings of the National Academy of Sciences of the United States of America, 85*, 4775–4778.
32. Price, A. L., Eskin, E., & Pevzner, P. A. (2004). Whole-genome analysis of Alu repeat elements reveals complex evolutionary history. *Genome Research, 14*, 2245–2252.
33. Simoens, C. R., Gielen, J., Van Montagu, M., & Inze, D. (1988). Characterization of highly repetitive sequences of *Arabidopsis thaliana*. *Nucleic Acids Research, 16*, 6753–6766.
34. Murata, M., Ogura, Y., & Mototoshi, F. (1994). Centromeric repetitive sequences in *Arabidopsis thaliana*. *Japanese Journal of Genetics, 69*, 361–370.
35. Wu, J. C., & Manuelidis, L. (1980). Sequence definition and organization of a human repeated DNA. *Journal of Molecular Biology, 142*, 363–386.
36. Yamada, K., et al. (2002). Empirical analysis of transcriptional activity in the *Arabidopsis* genome. *Science, 302*, 842–846.
37. The ENCODE Project Consortium. (2007). Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature, 447*, 799–816.
38. van Bakel, H., Nislow, C., Blencowe, B. J., & Hughes, T. R. (2010). Most "dark matter" transcripts are associated with known genes. *PLoS Biology, 8*, e1000371.
39. Yoshiura, K., et al. (2006). A SNP in the ABCC11 gene is the determinant of human earwax type. *Nature Genetics, 38*, 324–330.

40. Wolfe, K. H., & Shields, D. C. (1997). Molecular evidence for an ancient duplication of the entire yeast genome. *Nature, 387*, 708–713.
41. Ma, L.-J., et al. (2009). Genomic analysis of the basal lineage fungus *Rhizopus oryzae* reveals a whole-genome duplication. *PLoS Genetics, 5*, e1000549.
42. Aury, J. M., et al. (2006). Global trends of whole-genome duplications revealed by the ciliate *Paramecium tetraurelia*. *Nature, 444*, 171–178.
43. Rajic, Z. A., Jankovic, G. M., Vidovic, A., Milic, N. M., Skoric, D., Pavlovic, M., & Lazarevic, V. (2005). Size of the protein-coding genome and rate of molecular evolution. *Journal of Human Genetics, 50*, 217–229.
44. Lynch, M., & Conery, J. S. (2000). The evolutionary fate and consequences of duplicated genes. *Science, 302*, 1401–1404.
45. Gregory, T. R., & DeSalle, R. (2005). Comparative genomics in prokaryotes. In T. R. Gregory (Ed.), *The evolution of the genome*, Chapter 10, Burlington: Elsevier Academic Press.
46. Gott, J. M., & Emeson, R. B. (2000). Functions and mechanisms of RNA editing. *Annual Review of Genetics, 34*, 499–531.
47. Tillich, M., Lehwark, P., Morton, B. R., & Maier, U. G. (2006). The evolution of chloroplast RNA editing. *Molecular Biology and Evolution, 23*, 1912–1921.
48. Gall, J. G. (1981). Chromosome structure and the C-value paradox. *Journal of Cell Biology, 91*, 3s–14s.
49. Vendrely, R., & Vendrely, C. (1948). La teneur du noyau cellulaire en acide désoxyribonucléique à travers les organes, les individus et les espèces animales (in French). *Cellular and Molecular Life Sciences, 4*, 434–436.
50. Pollister, A. W., & Ris, H. (1947). Nucleoprotein determination in cytological preparations. *Cold Spring Harbor Symposia on Quantitative Biology, 12*, 147–157.
51. Swift, H. (1950). The constancy of deoxyribose nucleic acid in plant nuclei. *Proceedings of the National Academy of Sciences of the United States of America, 36*, 643–654.
52. How, G. F., Venkatesh, B., & Brenner, S. (1996). Conserved linkage between the puffer fish (Fugu rubripes) and human genes for platelet-derived growth factor receptor and macrophage colony-stimulating factor receptor. *Genome Research, 6*, 1185–1191.
53. Hardison, R. C. (2000). Conserved noncoding sequences are reliable guides to regulatory elements. *Trends in Genetics, 16*, 369–372.
54. Levy, S., Hannenhalli, S., & Workman, C. (2001). Enrichment of regulatory signals in conserved non-coding genomic sequence. *Bioinformatics, 17*, 871–877.
55. Carroll, S. B. (2005). Evolution at two level: On genes and form. *PLoS Biology, 3*, e245.
56. Siepel, A., Bejerano, G., Pedersen, J. S., Hinrichs, A. S., Hou, M., Rosenbloom, K., Clawson, H., Speith, J., Hillier, L. W., Richards, S., Weinstock, G. M., Wilson, R. K., Gibbs, R. A., Kent, W. J., Miller, W., & Haussler, D. (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Research, 15*, 1034–1050.
57. Kaplinsky, N. J., Braun, D. M., Penterman, J., Goff, S. A., & Freeling, M. (2002). Utility and distribution of conserved noncoding sequences in the grasses. *Proceedings of the National Academy of Sciences of the United States of America, 99*, 6147–6151.
58. Guo, H., & Moose, S. P. (2003). Conserved noncoding sequences among cultivated cereal genomes identify candidate regulatory sequence elements and patterns of promoter evolution. *Plant Cell, 15*, 1143–1158.
59. Inada, D. C., et al. (2003). Conserved noncoding sequences in the grasses. *Genome Research, 13*, 2030–2041.
60. Thomas, B. C., Rapaka, L., Lyons, E., Pedersen, B., & Freeling, M. (2007). Arabidopsis intragenomic conserved noncoding sequence. *Proceedings of the National Academy of Sciences of the United States of America, 104*, 3348–3353.
61. D'Hont, A., et al. (2012). The banana (*Musa acuminata*) genome and the evolution of monocotyledonous plants. *Nature, 488*, 213–217.
62. Kritsas, K., Samuel, E., Wuest, S. E., Hupalo, D., Kern, A. D., Wicker, T., & Grossniklaus, U. (2012). Computational analysis and characterization of UCE-like elements (ULEs) in plant genomes. *Genome Research, 22*, 2455–2466.

63. Hettiarachchi, N., & Saitou, N. (2013). Identification and analysis of conserved noncoding sequences in plants. *Genome Biology and Evolution* (in revision).
64. Sanjuan, R., Nebot, M. R., Chirico, N., Mansky, L. M., & Belshaw, R. (2010). Viral mutation rates. *Journal of Virology, 84*, 9733–9748.
65. Lynch, M. (2006). The origins of eukaryotic gene structure. *Molecular Biology and Evolution, 23*, 450–468.
66. Lynch, M. (2010). Evolution of the mutation rate. *Trends in Genetics, 26*, 345–352.
67. Drake, J. W., Charlesworth, B., Charlesworth, D., & Crow, J. F. (1998). Rates of spontaneous mutation. *Genetics, 148*, 1667–1686.
68. Arabidopsis Genome Initiative. (2000). Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature, 408*, 796–815.
69. Hu, T. T., et al. (2011). The *Arabidopsis lyrata* genome sequence and the basis of rapid genome size change. *Nature Genetics, 43*, 476–481.
70. Goff, S. A., & others. (2002). A draft sequence of the rice genome (*Oryza sativa* L. ssp. japonica). *Science*, *296,* 92–100.
71. Yu, J., & others. (2002). A draft sequence of the rice genome (*Oryza sativa* L. ssp. indica). *Science*, *296*, 79–92.
72. Londo, J. P., et al. (2006). Phylogeography of Asian wild rice, *Oryza rufipogon*, reveals multiple independent domestications of cultivated rice, *Oryza sativa*. *Proceedings of the National Academy of Sciences of the United States of America, 103*, 9578–9583.
73. Yang, C.-C., Kawahara, Y., Mizuno, H., Wu, J., Matsumoto, T., & Itoh, T. (2012). Independent domestication of Asian rice followed by gene flow from japonica to indica. *Molecular Biology and Evolution, 29*, 1471–1479.
74. The Rice Annotation Project. (2007). Curated genome annotation of *Oryza sativa* ssp. japonica and comparative genome analysis with *Arabidopsis thaliana*. *Genome Research, 17*, 175–183.
75. "Chromosomes" of KOMUGI database (http://www.shigen.nig.ac.jp/wheat/komugi/chromosomes/chromosomes.jsp)
76. Nickrent, D. L., et al. (2000). Multigene phylogeny of land plants with special reference to bryophytes and the earliest land plants. *Molecular Biology and Evolution, 17*(12), 1885–1895.
77. Rensing, S. A., et al. (2008). The *Physcomitrella* genome reveals evolutionary insights into the conquest of land by plants. *Science, 319*, 64–69.
78. Banks, J. A., et al. (2011). The *Selaginella* genome identifies genetic changes associated with the evolution of vascular plants. *Science, 332*, 960–963.
79. Mewes, H. W., Albermann, K., Bahr, M., Frishman, D., Gleissner, A., Hani, J., Heumann, K., Kleine, K., Maierl, A., Oliver, S. G., Pfeiffer, F., & Zollner, A. (1997). Overview of the yeast genome. *Nature, 387*, 7–65.
80. Lynch, M. (2007). *Origin of genome architecture*. Sunderland: Sinaur Associates.
81. Martin, F., et al. (2010). Perigord black truffle genome uncovers evolutionary origins and mechanisms of symbiosis. *Nature, 464*, 1033–1038.
82. Biography of E. B. Lewis (http://www.nobelprize.org/nobel_prizes/medicine/laureates/1995/lewis.html)
83. Gehring, W. J. (1999). *Master control genes in development and evolution: The homeobox story*. New Haven: Yale University Press.
84. Carroll, S. B., Grenier, J. K., & Weatherbee, S. D. (2005). *From DNA to diversity*. Malden: Blackwell Publishing.
85. Matsunami, M., Sumiyama, K., & Saitou, N. (2010). Evolution of conserved non-coding sequences within the vertebrate Hox clusters through the two-round whole genome duplications revealed by phylogenetic footprinting analysis. *Journal of Molecular Evolution, 71*, 427–436.
86. Chalfie, M. (Ed.). WormBook – The online review of *C. elegans* biology (http://www.wormbook.org/)
87. Foronda, D., de Navas, L. F., Garaulet, D. L., & Sanchez-Herrero, E. (2009). Function and specificity of Hox genes. *International Journal of Developmental Biology, 53*, 1409–1419.

88. Aruga, J., Kamiya, A., Takahashi, H., Fujimi, T. J., Shimizu, Y., Ohkawa, K., Yazawa, S., Umesono, Y., Noguchi, H., Shimizu, T., Saitou, N., Mikoshiba, K., Sakaki, Y., Agata, K., & Toyoda, A. (2006). A wide-range phylogenetic analysis of Zic proteins: Implications for correlations between protein structure conservation and body plan complexity. *Genomics, 87*, 783–792.
89. C. elegans Sequencing Consortium. (1998). Genome sequence of the nematode *C. elegans*: A platform for investigating biology. *Science, 282*, 2012–2018.
90. Meldal, B. H. M., et al. (2007). An improved molecular phylogeny of the Nematoda with special emphasis on marine taxa. *Molecular Phylogenetics and Evolution, 42*, 622–636.
91. Brenner, S. (1974). The genetics of *Caenorhabditis elegans*. *Genetics, 77*, 71–94.
92. Adams, M. D., & others. (2000). The genome sequence of *Drosophila melanogaster*. *Science*, *287*, 2185–2195.
93. Drosophila 12 Genomes Consortium. (2007). Evolution of genes and genomes on the Drosophila phylogeny. *Nature, 450*, 203–218.
94. Sea Urchin Genome Sequencing Consortium. (2006). The genome of the sea urchin *Strongylocentrotus purpuratus*. *Science, 314*, 941–952.
95. Dehal, P., & others. (2002). The draft genome of *Ciona intestinalis*: Insights into chordate and vertebrate origins. *Science*, *298*, 2157–2167.
96. Vinson, J. P., Jaffe, D. B., O'Neill, K., Karlsson, E. K., Stange-Thomann, N., Anderson, S., Mesirov, J. P., Satoh, N., Satou, Y., Nusbaum, C., Birren, B., Galagan, J. E., & Lander, E. S. (2005). Assembly of polymorphic genomes: Algorithms and application to Ciona savignyi. *Genome Research, 15*, 1127–1135.
97. Matthysse, A. G., Deschet, K., Williams, M., Marry, M., White, A. R., & Smith, W. C. (2004). A functional cellulose synthase from ascidian epidermis. *Proceedings of the National Academy of Sciences of the United States of America, 101*, 986–991.
98. OOta, S., & Saitou, N. (1999). Phylogenetic relationship of muscle tissues deduced from superimposition of gene trees. *Molecular Biology and Evolution, 16*, 856–867.
99. Sadaie, Y., et al. (Eds.). (2004). *Genome science and microorganismal molecular genetics (in Japanese)*. Tokyo: Baifukan.
100. Arslan, D., Legendre, M., Seltzer, V., Abergel, C., & Claverie, J. M. (2011). Distant Mimivirus relative with a larger genome highlights the fundamental features of Megaviridae. *Proceedings of the National Academy of Sciences of the United States of America, 108*, 17486–17491.
101. Raoult, D., & others. (2004). The 1.2-megabase sequence of mimivirus. *Science*, *306*, 1344–1350.
102. Bejerano, G., Pheasant, M., Makunin, I., Stephen, S., Kent, W. J., Mattick, J. S., & Haussler, D. (2004). Ultraconserved elements in the human genome. *Science, 304*, 1321–1325.
103. Saitou, N. (2007). *Genomu Shinkagaku Nyumon (written in Japanese, meaning 'Introduction to evolutionary genomics')*. Tokyo: Kyoritsu Shuppan.
104. Holland, L. G. (2008). The amphioxus genome illuminates vertebrate origins and cephalochordate biology. *Genome Research, 18*, 1100–1111.
105. Phillipe, N., et al. (2013). Pandoraviruses: Amoeba viruses with genomes up to 2.5 Mb reaching that of parasitic eukaryotes. *Science, 341*, 281–286.
106. The Chloroplast Genome Database. http://chloroplast.ocean.washington.edu/