# Genome Evolution: Overview

**J Bruce Walsh,** *University of Arizona, Tucson, Arizona, USA*

Introductory article

Article Contents



The genome is the total genetic constitution of an organism. Understanding of the structure and evolution of genomes is undergoing a revolution with ability to sequence entire genomes. A key feature in the evolution of genomes is the creation of new genes, usually by duplication.

## Introduction

Our understanding of the structure and evolution of genomes (the total genetic constitution of an organism) is presently undergoing a revolution on a par with any other in the history of science. We now have the ability to sequence entire genomes and this is happening at a rapidly increasing pace. Further, molecular studies have provided a fairly good understanding of the deep phylogeny of life (how the major groups of life are related to each other), providing the framework for examining the evolution of genomes across all of life. It is now established that rather than the old prokaryotic–eukaryotic division (bacteria versus cells with nuclei) there are instead three very deep, fundamental branches of life, the Bacteria (or eubacteria), the Archaea (or archaebacteria), and the Eucarya (or eukaryotes). Prokaryotes thus consist of two radically different groups (the eubacteria and archaebacteria) that are as distinct from each other as either is from the eukaryotes. While we will still use the term prokaryote when referring to features shared by both eubacteria and archaebacteria, the reality is that the archaebacteria are (slightly) more closely related to eukaryotes than they are to eubacteria. Excluding those of viruses, genomes consist of one or more double-stranded molecules of DNA. Depending on the particular strain, the genomes of viruses use all four potential types of nucleic acids. Some use double-stranded DNA, others use single-stranded DNA, still others either single-stranded or double-stranded RNA.

## How Genes and Genomes Are Organized

The vast majority of genes code for messenger RNAs (mRNAs) that are translated into proteins, with the collection of all protein-coding genes within a genome referred to as the proteome. Proteins can be broadly classified as those involved in basic cellular housekeeping functions common to essentially all cells, and those proteins that have more specialized functions (such as those that appear only in specific tissues and/or only when the cell experiences specific environmental cues). Examples of housekeeping function include information storage and retrieval (DNA replication and repair, transcription of DNA into RNA, and translation of mRNA into proteins), metabolism (synthesis of complex organic molecules from simpler precursors), energy management and storage, cellular transport and cellular division. While such housekeeping proteins are absolutely essential to the cell, the majority of proteins have tissue-specific or environment-specific roles. For example, the bulk of genes in multicellular organisms are probably involved in tissue-specific functions and in regulating the development of different cell types.

The genome also contains a smaller (but certainly no less important) set of genes coding for structural RNAs. In both prokaryotes and eukaryotes, ribosomal and transfer RNAs (rRNAs and tRNAs, respectively) play critical roles in translation. Additionally, eukaryotes have a number of small RNAs that play key roles in rRNA and mRNA processing, cellular transport and other functions. Collectively, all the protein-coding and structural RNA-coding genes constitute the genic DNA of a genome. For prokaryotes, this is the bulk of the entire genome. In eukaryotes, genic DNA comprises only a fraction (and in some cases a very small fraction) of the total genome.

### Eubacteria

The Bacteria as well as the Archaea lack a nuclear membrane and as a consequence there is no clear separation of the genome (DNA) from the cytoplasm. This lack of compartmentalization allows for coupled transcription–translation wherein a mRNA can be translated into a protein on ribosomes while it is still being transcribed from the DNA. Typically, the genomes of these groups exist as a single circular chromosome, with a single replication origin (DNA replication can only initiate at a single site). Genome size is small, with very little nongenic DNA. The genome may also be augmented by a number of plasmids (small circular DNAs), which usually (but not always) lack genes essential for normal cellular function.

Bacterial genes are tightly clustered, often organized in an operon structure in which a single long transcript contains several distinct genes, each with their own initiation and termination codons. Such transcripts are called polycistronic, as they code for multiple cistrons (proteins). Individual genes typically start with a Shine–Dalgarno sequence that facilitates the binding of mRNA to the ribosome for translation. Bacterial genes are uninterrupted, lacking the internal intron sequences that are ubiquitous in eukaryotic protein-coding genes, and do not undergo the extensive mRNA processing found in eukaryotes. Transcriptional control of bacterial genes (and operons) occurs by proteins binding to the promoter and a few adjacent regions to either enhance or block transcription. Bacteria use a single RNA polymerase for transcribing both mRNAs and structural RNAs (rRNAs, tRNAs).

## Eukaryotes

Eukaryotic genomes contain multiple linear chromosomes, each chromosome containing multiple origins of replication. The presence of multiple linear chromosomes requires specialized sequences for proper replication of the chromosome ends (telomeres) and special sequences to ensure the correct segregation of chromosome pairs during cell division (centromeres). Eukaryotic genomes are considerably larger (by orders of magnitude) than bacterial genomes. Part of this size difference is due to increases in the total number of genes, but the vast majority is due to a great increase in the fraction of nongenic DNA (see below). The presence of multiple replication origins probably accounts for eukaryotes' much larger genome sizes, as the speed of genome replication (and hence cell growth) is far less constrained by genome size relative to prokaryotes with their single origin. A consequence of the much larger genome sizes in eukaryotes is that the DNA in eukaryotic chromosomes must be tightly packaged to fit within the nucleus (in some species, this is equivalent to fitting over a hundred miles of wire into an object the size of a basketball). At the lowest level of packing, the DNA is wrapped around complexes of histone proteins to form nucleosomes. Strings of nucleosomes are themselves further folded to greatly increase the DNA compaction. The result is chromatin, a DNA molecule extensively covered with proteins and highly condensed.

The eukaryotic genome is enclosed in a nuclear membrane that separates the DNA from the cytoplasm of the cell. This nuclear–cytoplasmic separation has profound influences on transcription and translation. Following transcription, extensive RNA processing occurs in the nucleus before the final mRNAs are transported to the cytoplasm for translation into proteins. Both ends of the initial mRNA transcript are modified, the starting (5′) end with a special nucleotide cap, while a run of adenines (the poly(A) tail) is added at the finishing (3′) end. By far the most striking feature of eukaryotic mRNA processing is that the original transcript often contains numerous introns, internal sequences that must be precisely removed (spliced out) to create the final product. Introns can easily make up the vast bulk of a gene, with the coding exons (those regions of the mRNA that remain following splicing) comprising only a small part of the initial transcript.

Transcription in eukaryotes is much more complex than in prokaryotes for several reasons. First, the configuration of the chromatin surrounding a gene has a strong influence on its expression. Second, transcriptional control often requires the binding of several transcription factors at multiple regulatory sites adjacent to the gene to form large multiprotein transcription complexes. Because of the tight packing of eukaryotic DNA, sequences very far apart on a chromosome may in fact actually lie close together in the nucleus. This probably accounts for the fact that transcription can be strongly influenced by enhancer sequences many thousands of bases away from the gene, which can greatly increase the level of transcription. Eukaryotes use three distinct RNA polymerases for transcription: mRNAs (and some small RNAs) are transcribed by RNA polymerase II (Pol II), rRNAs use Pol I, and tRNA and other small RNAs used Pol III. Eukaryotic mRNAs lack the Shine–Dalgarno sequence used by prokaryotes to bind the mRNA to the ribosome. In eukaryotes, mRNA–ribosome binding is facilitated by ribosomal protein interactions with the 5′ cap. Finally, operon-like structures are only rarely found in eukaryotes (the most extreme being in the nematode *Caenorhabditis elegans* in which roughly 25% of genes exist in short operons where a single initial transcript contains two or more separate genes). Even when operons are present, as a result of extensive RNA processing and splicing, the final mRNA that is transported into the cytoplasm codes for only a single protein. Thus, eukaryotic cytoplasmic mRNAs are not polycistronic.

Most eukaryotes contain one or two additional genomes beyond the nuclear DNA, as the two major cellular organelles (the mitochondrion and plastid or chloroplast) each contain their own distinct genomes. Comparative sequencing clearly shows that each of these organelles originally arose by a single endosymbiotic event. First, a very primitive eukaryotic cell engulfed a eubacterium that gave rise to mitochondria. This endosymbiotic event occurred near the base of the eukaryotic tree. Second, the ancestor of the plants and algae later captured a cyanobacterium that gave rise to the plastid. In both cases, the majority of genes from the captured eubacterial ancestor were transferred to the eukaryotic nucleus. Despite their small size, organelle genomes still contain unmistakable eubacterial signatures from their ancestors (for example, most organelle genomes exist as single circular chromosomes).

## Archaea

In many respects, the genomes of Archaea (the archaebacteria) are typical of bacteria: a single major circular chromosome with a single origin of replication, and a small genome consisting mainly of genic DNA. Archaeal protein-coding genes lack introns, are often clustered in operons, and have Shine–Dalgarno ribosome-binding sites. Despite these eubacterial similarities, sequencing studies have shown that archaebacterial genomes also have many eukaryotic features. Much of the cellular informational processing machinery (DNA replication, transcription, translation) is far more like that found in eukaryotes. For example, the DNA replication enzyme (DNA polymerase) used by archaebacteria is homologous to the eukaryotic polymerases, both of which are unrelated to the polymerases used by eubacteria. Most other archaeal replication proteins are also far more eukaryote-like than eubacteria-like. Similar patterns are seen in genes involved in transcription and translation. Conversely, most archaeal metabolic genes are much more like eubacterial genes. This chimaeric distribution of archaeal genomes, with some genes being very eubacteria-like while others are very eukaryote-like is consistent with the Archaea being a third distinct branch of life about equally distant from the other two groups. However, this distribution has also led to suggestions that the Archaea may be the result of an ancient fusion between a eubacterium and eukaryote.

## The first genomes

While the Earth did not have a stable crust until about 4.0 billion years ago (Bya), unmistakable and unambiguous evidence of cells is seen at 3.5 Bya and reasonable evidence at 3.7–3.8 Bya. Life thus exploded onto the Earth in what amounts to a cosmic instant. Given the complexity of even the smallest genomes (Mushegia and Koonin argue that the minimal number of genes required for a cell is around 250–300): how did the first genome (the progenote) arise and what was its nature?

While present-day cells rely absolutely on proteins for almost all cellular functions, there are suggestions that primitive cells may have been entirely RNA-based. RNA can both store genetic information and catalyse biological reactions, circumventing the historical problem of which came first, proteins or DNA. The observation that DNA replication in all present-day genomes requires an RNA primer is consistent with an RNA-based progenote, potentially being a vestigial relict of an RNA genome. If the progenote was indeed RNA-based, how did it transform into a DNA-based genome that uses proteins for most cellular functions? The concept of hypercycles provides a solution. Suppose factor 1 is required to make factor 2, factor 2 is required to make factor 3, and factor 3 is required to make 1. This is a simple example of a hypercycle in which the replication of each component depends on replication of the others. If an additional factor (such as a protein component) can increase the speed of reaction, there is strong selection pressure to incorporate it into the hypercycle. Very complex biological interactions can be built up in this matter. Thus a complex RNA–protein hypercycle can evolve from an initially RNA-only hypercycle with far fewer components. At some point a primitive reverse-transcriptase enzyme converted RNA into DNA, allowing an RNA-based genome to be turned into a chemically more stable DNA genome.

Even allowing for hypercycles, the simplest biochemical pathways in present-day genomes are still extraordinarily complex. In a typical reaction the cell starts with factor A and, through a complex series of protein-mediated enzymatic interactions, converts it to the final product E (say) required by the cell: e.g. A→B→C→D→E. How do such complex metabolic pathways evolve? It has been suggested that they evolved backwards. Suppose that initially E was common in the environment, as might be expected on a primeval Earth rich in organic (i.e. carbon-based) molecules but largely devoid of life. The first life forms would use what is present (E). As the supply of E became rarer, there would be a great selective advantage to those cells that could use D (also present in the environment) and convert this to E. Proceeding in this fashion, the complex pathways of today could have evolved backwards by a series of such steps.

Another central question is what can be said about the genomic structure of the last common ancestor (the cenancestor) of the three major domains of life. We can draw inferences about its nature by examining the genes and cellular processes shared by all three groups. First, the translational code was in place in the cenancestor because all present-day genomes are based on the same genetic code (specifying which codons code for which amino acids). Second, while the progenote may have been RNA-based, the cenancestor likely had a DNA-based genome. Evidence for this is that all three domains share homologous enzymes for dealing with DNA (such as DNA topoisomerases, gyrases, and DNA-dependent RNA polymerases). However, the replication machinery may not have been finalized in the cenancestor, as the DNA polymerases used by eubacteria are unrelated to those used by archaebacteria and eukaryotes. Comparing the fully-sequenced genome of a eukaryote (yeast) with that of a eubacterium and an archaebacterium shows 80 detectable orthologues (DNA sequences showing signs of common ancestry) common to all three groups. This is probably an underestimate, as some genes are expected to evolve to the point where it is very difficult to detect any signature of past common ancestry. Further, nonorthologous gene displacement has likely occurred in some lineages, with unrelated genes taking over the roles of other genes, displacing the original orthologous genes. A more detailed comparison between a eubacterial and archaebacterial genome found around 260 genes in common, of which 130 were involved in

informational processing (95 for translation, 18 for DNA replication, 8 for recombination/repair, 9 for transcription). The remainder were involved in nucleotide metabolism (23), general cellular metabolism (86), and other functions (18).

A widely debated issue is whether the cenancestor (and perhaps even the progenote) had interrupted genes (introns). The 'introns-early' view holds that introns were present in the cenancestor but were lost by eubacteria and archaebacteria, while the 'introns-late' hypothesis holds that introns invaded the protein-coding genes after the eukaryotic ancestor branched off from the cenancestor. The phylogenetic distribution of introns is far more consistent with the introns-late view. For example, nuclear-encoded genes originally from organelles contain introns, yet phylogeny shows that the ancestors of these organelles lacked introns, implying that introns arose following the transfer of these genes to the nucleus.

## Variation in Gene Number

Estimates of the number of protein-coding genes in sequenced genomes are obtained by counting the number of open reading frames (ORFs) of sufficient length and with sufficient other signatures (such as start and stop codons and appropriate regulatory regions). Some ORFs can immediately be assigned to known families of proteins while others remain as unidentified reading frames (URFs), potentially indicating genes (of unknown function). The current numbers of such detected ORFs in eubacterial genomes range from 500 to 700 for certain intercellular parasites to 1000–4300 in free-living bacteria. Archaeal genomes show gene numbers ranging from 1800 to 2500, although this range is expected to grow as more species are sequenced. The range of gene number in eukaryotes is considerably higher: yeast (*Saccharomyces*) with 6154; nematodes (*C. elegans*) with 19 100; the fruit fly (*Drosophila*) with around 12 000; and humans with around 70 000. These numbers are probably underestimates as small genes are often overlooked by ORF-searching programs as they scan for long open reading frames.

One very sobering, and exciting, observation is that even in the two best-characterized organisms (the human gut eubacteria *Escherichia coli* and the yeast *Saccharomyces*), the majority of ORFs have unknown function – 60% (2600) of *E. coli* and 56% (3500) of yeast ORFs have unknown functions. Further, in yeast a large fraction of the ORFs can be individually deleted with no obvious effect. As the genomes of higher eukaryotes (such as flies, mice and humans) are sequenced, we expect the percentage of unknown ORFs are likely to be much higher.

.

## Organelle genomes and genome miniaturization

Mitochondrial genomes (mtDNA) also show considerable variation in gene number, especially given that the eubacterial ancestor was expected to contain between 2000 and 4000 genes at the time of capture. Fully sequenced mtDNAs show between 3 and 62 protein-coding genes and between 5 and 28 RNAs (all contain the 16S and 23S rRNAs, and most contain 22–25 tRNAs), with most animal mitochondria containing no more than a dozen or so protein-coding genes. Looking across all sequenced mtDNAs, only four genes seem to have been conserved in all genomes – the *cob* and *cox1* genes involved in cellular respiration, and the large and small rRNAs. Over a thousand nuclear-encoded proteins are imported into the mitochondria, mostly produced by genes originally from the eubacterial ancestor. For example, the ribosomal proteins and most other machinery to translate the mtDNA-encoded mRNAs are prokaryotic in nature, and distinct from those required for cytoplasmic ribosomes. Thus, following the initial endosymbiotic event, there was massive gene-transfer of eubacterial genes into the eukaryotic nucleus. Given this transfer, a key unresolved issue is why mitochondrial genomes exist at all. Indeed, there are a few eukaryotes that lack mitochondria. While these amitochondrial eukaryotes were originally thought to be very deep-branching lineages predating the original endosymbiotic event, it is now known that some have a far more recent origin and have lost their mitochondria fairly recently. Other amitochondrial eukaryotes do appear to be very deep-branching, but sequencing has shown that they contain one or more key mitochondrial genes of eubacterial origin in their nucleus. Thus, the mitochondrial endosymbiotic event must have occurred very early in the eukaryotic radiation, with even very deep-branching amitochondrial eukaryotes originally containing mitochondria, which transferred at least some genes to the nucleus before the organelle itself was lost in these lineages.

Plastids (chloroplasts) also contain a very reduced set of genes relative to their cyanobacterial ancestor, although they contain many more genes than mitochondrial genomes (typically around 100 or so protein-coding genes). The plastids from certain groups prove very interesting examples of multiple rounds of gene transfer. It is clear for both structural and phylogenetic reasons that some eukaryotes have obtained their plastids from secondary endosymbiotic events by engulfing a plastid-containing alga. In at least two such cases of secondary endosymbiosis, the current plastid contains a vestigial remnant of the nucleus of the captured eukaryotic cell, the nucleomorph. These are the smallest eukaryotic genomes known, being on the order of 350 000 to 600 000 bases and containing an estimated 200 or so genes. Plastids with nucleomorphs represent cases of two massive gene transfers – first from the cynanobacterial ancestor to the eukaryotic nucleus in

the primary endosymbiotic event and then the transfer of these genes again from the nucleus of the engulfed alga to the nucleus of the new host eukaryote. Note that these eukaryotes have four separate genomes: nuclear, mitochondrial, plastid and nucleomorph.

## Variation in Genome Size: The *C*-value Paradox

For historical reasons, the (haploid) genome size of an organism is often referred to as its *C* value. The genome size of eubacteria and archaebacteria shows over a 30-fold range, from 0.4 megabases (1 Mb = $10^6$ bases ) to 13 Mb of DNA. Most of this variation must be due to differences in the number of genes, as the correlation between genome size and number of protein-coding genes in fully sequenced prokaryotic genomes is extremely high.

Considerable size variation is also seen in mtDNA genomes, which range from a low of 6000 bases to over 3 Mb. This is in contrast to the total number of mitochondrially encoded genes, which show an 18-fold range, compared to the 500-fold range in the mitochondrial *C* value. Plastid genomes show a much smaller range in *C*-values and this variation is similar to the variation in gene number.

Eukaryotic nuclear genomes show an almost 80 000-fold range in genome size (from 3.5 Mb to over 600 000 Mb), with plants showing a 6000-fold range across different species, animals a 2000-fold range, and mammals a 4-fold range. This huge range has been referred to as the *C*-value paradox. While the minimal genome size increases with phylogenetic complexity (the smallest insect, fish, and mammal *C* values are 10, 38, and 1400 Mb, respectively), the upper limits within each group show no such correlation. For example, humans have 3400 Mb, while some species of pines have 68 000 Mb, lungfish have 140 000 Mb, certain ferns 160 000 Mb, and two species of amoeba have the largest genomes (300 000 and 670 000 Mb). The huge variation in *C* values is only partly due to differences in the number of genes, as there is only about a 50-fold variation among eukaryotes in the number of protein-coding genes. Rather, most of the variation is due to differences in the amount of nongenic DNA. The fraction of nongenic DNA in eukaryotes ranges from under 30% to over 99.9%, which (when differences in genome size are taken into account) translates into a 300 000-fold range in the total amount of nongenic DNA. Humans have around 10% genic DNA, implying that we carry in our genomes around 3000 Mb of DNA with no obvious cellular function. What is all this extra DNA doing? Some possible explanations follow from the fact that large fractions of nongenic DNA are repetitive DNAs of one sort or another.

## Types of repetitive DNA

Repetitive DNAs are sequences that exist as multiple copies within a genome. Gene families, related genes coding for mRNAs or structural RNAs, are ubiquitous features of genomes, but the vast majority of repetitive sequences are nongenic. We can classify these by their genomic organization, in that many families of nongenic repeated sequences are clustered in one or a few localized repeats, while other families exist as dispersed repeats with their copies scattered around the genome. These organizations provide clues as to the origin and maintenance of such nongenic families.

Dispersed repeats are typically the result (directly or indirectly) of mobile genetic element or transposons – DNA sequences that code for the ability to make additional copies of themselves at other genomic locations. These elements are genomic parasites. Transposons make up a large fraction of many eukaryotic genomes (they are also present in prokaryotes, but make up a much smaller fraction of the genome). For example, 36% of the human genome consists of sequences with similarity to known mobile genetic elements. This figure is almost certainly an underestimate as ancient families of mobile elements that are no longer active have decayed through mutation to the point of no longer being recognizable. Just two element types comprise almost 25% of the entire human genome – ALU elements (1 200 000 copies for 10% of the genome) and longer LINE (long interspersed DNA element)-1 elements (600 000 copies for 15% of the genome).

Functionally, transposons can be classified into those that use an RNA intermediate to move about the genome (retrotransposons) and those that move entirely through DNA intermediates. The ALU and LINE elements in humans are both retrotransposons, but are evolutionarily unrelated. Full-length LINE elements encode the reverse transcriptase enzyme necessary for retrotransposition and use Pol II for transcription. ALU RNAs are Pol III transcripts and do not code for reverse transcriptase, rather they rely on LINE elements to produce this enzyme. Thus, in a sense ALUs are parasitic on LINE elements, and both are parasitic in using other machinery from the human genome (such as the appropriate transcription factors) to make additional copies of themselves.

Examples of localized repeats include satellite DNAs of various types (the term following from the fact that such sequences form a band or satellite when whole genomic DNA is broken into small fragments and centrifuged at high speed). Such DNAs exist as a moderate to very large numbers of tandem repeats of a common core sequence. Satellite DNAs are further divided into micro- and minisatellites depending on the length of the repeat unit, with minisatellites having a repeat unit of just a few bases while microsatellites have longer repeat units. The fraction of satellite DNA within a genome can be very large. For example, mammalian genomes typically consist of between

5% and 30% satellite DNAs, while in plants the figure is around 40%. There are also more extreme cases, one example being the Kangaroo rat. Here, over half the genome consists of families of just three basic repeats, with over two billion copies of a three-base repeat, two billion copies of a six-base repeat, and a billion copies of a ten-base repeat. Satellite DNAs typically do not code for RNAs (although there are rare exceptions), and they are generally found in genomic regions showing reduced recombination. They are probably generated by a combination of replication slippage (where DNA polymerase slips when copying a small repeat, generating excess) and incorrect recombination between multiple adjacent copies (unequal crossing-over). Other mechanisms of local gene amplification may also be involved.

## Nongenic DNA: selfish, junk or structural?

Three not necessarily exclusive hypotheses have been proposed to account for the huge fraction, and variation therein, of nongenic DNA in eukaryotes. The selfish DNA hypothesis states that most nongenic DNA consists of sequences that exist solely to make additional copies of themselves (i.e. they are transposons). Such repetitive DNAs will spread even if there is some cost to the host. The junk DNA hypothesis states that much of the nongenic DNA is simply a byproduct of DNA replication, recombination and mutation (such as satellite DNAs). Rather than actively spreading copies through the genome (as under the selfish DNA model), nongenic DNA simply piles up like junk in a closet. This hypothesis assumes that the cost to the cell of carrying this extra DNA is negligible. Finally, the structural DNA hypothesis states that much of this nongenic DNA has important cellular functions. In particular, it has been argued that genome size influences cell size, and that selection for increased (or decreased) cell size can indirectly influence genome size. At present, there is little evidence to support the structural DNA hypothesis, and the vast majority of nongenic DNA is thought to be either junk or selfish.

## Concerted evolution of repeated sequences

One surprising observation is that gene family members within a species are often more similar to each other than they are to members from different species. This is especially true for nongenic repeats. This phenomenon is referred to as concerted evolution, as the individual repeats do not seem to evolve independently, but rather appear to evolve in concert within a species. Given that most nongenic DNA is not thought to be under selection, what mechanism(s) account for concerted evolution? The key is that several recombination-related processes allow for sequence exchange between repeats. One process is gene conversion, where one DNA region converts another region to its sequence. Conversion can occur between both localized and dispersed repeats. When repetitive sequences occur as tandem arrays, unequal crossing-over also results in some members of the array being over- or under-represented following the crossover event. Akin to genetic drift removing variation in a finite population, multiple rounds of gene conversion and/or unequal crossing-over result in a collection of repeats becoming more similar. These sequence exchange processes can thus produce concerted evolution even in the absence of any selection.

## Expansion of genomes: the origin of new genes

One key feature in the evolution of genomes is the creation of new genes. Most new genes arise by duplication of existing genes, with the duplicate copies diverging and acquiring new functions. Exon shuffling is a variation on the theme of gene duplication, wherein the exons from two or more genes are joined (shuffled) together to create a new gene. While shuffling is an important mechanism for gene creation in eukaryotes, its role in the other domains of life is less clear. Proponents of the introns-early school have argued that many early genes were formed by similar processes in the progenote, by shuffling between a small set of minigenes coding for different protein domains.

Not all duplications result in new genes, as one of the copies can acquire one or more inactivating mutations, becoming a pseudogene no longer under selection so that its sequence signature decays away over time. Pseudogenes are common features of eukaryotic genomes, and are often found in clusters of related genes. Eukaryotic genomes also contain processed pseudogenes, which are reverse-transcribed mRNAs that have been inserted back into the genome. Such pseudogenes are inactive at the time of their formation, while the duplicate gene that eventually becomes a traditional pseudogene often remains functional for tens of millions of years before becoming inactivated.

While duplications can be local, involving one gene or a few genes, duplication of the entire genome also occurs. For example, the complete sequence of yeast (*Saccharomyces*) indicates that is has undergone a whole-genome duplication, but that only about 8% of the duplicated genes survived. There is also clear evidence for whole-genome duplication in many plants and in some bacteria. It has also been argued that two successive genome duplications occurred during the transition to the higher vertebrates, one between the invertebrates and the jawless fishes and the second occurring immediately after the jawless fishes. Consistent with this is the observation that several genes found as four copies in humans are present as two copies in amphibians and as single copies in insects. While there is still debate on whether these successive whole-genome duplications did indeed occur in verte-

brates, the complete sequencing of the human and mouse genomes will go a long way towards resolving the issue.

## Further Reading


Cavalier-Smith T (ed.) (1985) *The Evolution of Genome Size*. New York: Wiley.

Clayton RA, White O, Ketchum KA and Venter JC (1997) The first genome from the third domain of life. *Nature* **387**: 459–460.

Douglas SE (1998) Plastid evolution: origins, diversity, trends. *Current Opinion in Genetics and Development* **8**: 655–661.

Forterre P (1997) Archaea: what can we learn from their sequences? *Current Opinion in Genetics and Development* **7**: 764–770.

Gilson PR, Maier U-G and McFadden GI (1997) Size isn't everything, lessons in genetic miniaturization from nucleomorphs. *Current Opinion in Genetics and Development* **7**: 800–806.

Gruar D and Li W-H (1999) *Fundamental of Molecular Evolution*, 2nd edn. Sunderland, MA: Sinauer Associates.

Keeling PJ (1998) A kingdom's progress: archezona and the origin of eukaryotes. *BioEssays* **20**: 87–95.

Koonin EV and Galperin MY (1997) Prokaryotic genomes: the emerging paradigm of genome-based microbiology. *Current Opinion in Genetics and Development* **7**: 757–763.

Li W-H (1997) *Molecular Evolution*. Sunderland, MA: Sinauer Associates.

Logsdon JM Jr (1998) The recent origins of spliceosomal introns revisited. *Current Opinion in Genetics and Development* **8**: 637–648.

Mushegia AR and Koonin EV (1996) A minimal gene set for cellular life derived by comparisons of complete bacterial genomes. *Proceedings of the National Academy of Sciences of the USA* **93**: 10268–10273.

Olsen GJ and Woese CR (1997) Archaeal genomics. An overview. *Cell* **89**: 991–994. [Four additional reviews covering different aspects of Archaeal genomics follow this overview.]

Page RDM and Holmes EC (1998) *Molecular Evolution: A Phylogenetic Approach*. Oxford: Blackwell Science.

Palmer JD (1997) The mitochondrion that time forgot. *Nature* **387**: 454–455.

Sidow A (1996) Gen(om)e duplications in the evolution of early vertebrates. *Current Opinion in Genetics and Development* **6**: 715–722.

Skrabanek L and Wolfe KH (1998) Eukaryote genome duplication – where's the evidence? *Current Opinion in Genetics and Development* **8**: 694–700.

Singer M and Berg P (1991) *Genes and Genomes*. Herndon, VA: University Science Books.

Smit AFA (1996) The origin of interspersed repeats in the human genome. *Current Opinion in Genetics and Development* **6**: 743–748.

TIGR (2000) *TIGR – The Institute for Genomic Research*. [http://www.tigr.org/tdb/index.shtml] [The Institute for Genomic Research website provides a listing of all fully sequenced genomes.]

OGMP (2000) OGMP – The Organelle Genome Megasequencing Program [http://megasun.bch.umontreal.ca/ogmpproj.html] [The Organelle Genome Megasequencing Program website provides a listing of fully sequenced mitochondrial and chloroplast genomes.]