Gene Duplication and Redundancy

Introductory article

Andreas Wagner, *University of New Mexico, Albuquerque, New Mexico, USA*

Gene duplications create one or more copies of a gene in a genome. They are important forces of genome evolution which change genome size and lead to the evolution of new gene functions.

Mechanisms of Gene Duplication

Gene duplications are the accidental byproducts of cellular processes (deoxyribonucleic acid (DNA) replication, recombination and gene expression) that can generate copies of DNA regions (DNA repeats) within a genome. A gene duplication occurs when a DNA repeat thus generated includes at least one gene. Specifically, two major mechanisms can lead to gene duplication: nonhomologous recombination (unequal crossing-over) and retroposition. In unequal crossing-over, two nonhomologous DNA double helices align and undergo recombination, as shown in **Figure 1a**. Unequal crossing-over is greatly facilitated if the two strands already contain repeat units, as shown in the figure, but it can also occur if that is not the case. The second mechanism, retroposition, requires a gene to be transcribed into ribonucleic acid (RNA). From this RNA, the cellular enzyme reverse transcriptase then produces a double-stranded DNA copy, which can then integrate into the genome at some arbitrary location (**Figure 1b**). Genes thus duplicated are also called retrogenes. Retroposition usually does not generate a copy of the regulatory DNA sequences of the original gene, and sometimes does not generate a complete copy of the gene. Such duplicate genes cannot express functional gene product and are called retropseudogenes. A key diagnostic distinction between duplication through nonhomologous recombination and retroposition is that introns are usually eliminated during retroposition.

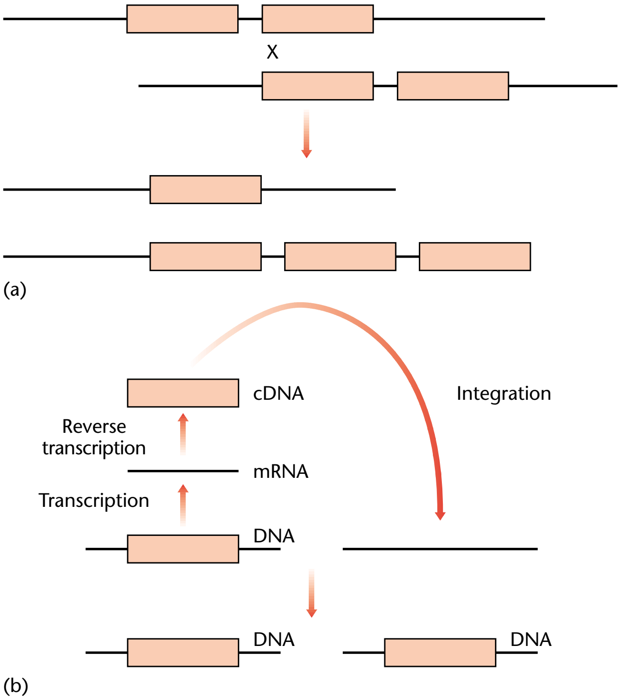

Figure 1 Gene duplications can occur via (a) unequal crossing-over or (b) retroposition (see text for details).

Gene duplications are biologically significant for three reasons. First, they change the number of genes in a genome. Second, they facilitate the evolution of new gene functions. Third, for some genes multiple copies may be necessary to ensure that a sufficient amount of gene product (RNA or protein) can be made. In addition to duplications of individual genes, entire genomes can be duplicated through the failure of chromosome segregation during cell division. Genome duplications will not be addressed here.

Protein-coding Genes

Although biologists have long suspected that gene duplications are important in evolution, their importance has become glaringly obvious with the availability of complete genome sequences. About one-third of the genes in fully sequenced genomes are duplicate genes. Some genes have only one duplicate, others occur in large families of over 100 duplicates that arose through repeated duplication of individual family members. Expansion of particular genes into large families is often specific to organismal lineages and can sometimes be traced to aspects of an organism's biology. Nematodes like *Caenorhabditis elegans*, for example, have collagenous cuticles. The importance of collagen in this lineage is reflected in the genome by a family of some 160 collagen genes. The amplification of genes required for tryptophan synthesis in endosymbiotic bacteria of the genus *Buchnera* is correlated with this endosymbiont's role in providing tryptophan for its aphid host. Natural selection may thus play a role in the expansion of some gene families. However, it is unclear whether it does so for all gene families. Random processes such as genetic drift may well contribute to the establishment and expansion of many gene families.

What is the fate of original and duplicate genes after duplication? First, one of the duplicates may suffer a loss-of-function mutation and disappear from the genome, thus reinstating the genome's state before duplication. Second, both duplicates may remain in the genome. If so, they might both become indispensable, either because their functions become partitioned among them, or because advantageous mutations create new functions in one of them. To address the question which of these processes is prevalent in genome evolution, it is necessary to analyse how the DNA sequence of duplicate genes diverges. To this end, two useful indicators of DNA sequence divergence will be briefly introduced.

The first indicator is the fraction of synonymous (silent) nucleotide substitutions, K_s, per nucleotide site. Silent substitutions do not lead to amino acid changes in the protein encoded by a gene. The second indicator is the fraction of nonsynonymous (amino acid replacement) nucleotide substitutions, K_a, per nucleotide site. These quantities are useful for two reasons. First, K_s provides a crude measure of time since duplication for each gene pair. The reason is that synonymous nucleotide substitutions are not subject to the same strong selection pressures as nonsynonymous substitutions that change amino acids in a protein. They thus accumulate at a stochastic rate proportional to time. In organisms where fossil or other evidence can be used to calibrate this molecular clock, it is even possible to assign a crude absolute time scale to observed values of K_s. Another important use of these indicators of sequence divergence derives from the ratio $K_a{:}K_s$. This ratio provides a measure of the selection pressure a gene pair is subject to. If a duplicate gene pair shows $K_a{:}K_s \approx 1$, that is, amino acid replacement substitutions occur at the same rate as synonymous substitutions, then few or no amino acid replacement substitutions have been eliminated since the gene duplication. In other words, the duplicate genes are under few or no selective constraints. The gene pair is said to be under 'purifying selection' if $K_a{:}K_s < 1$. Here, some replacement substitutions have been purged by natural selection, presumably because of their deleterious effects. The smaller $K_a{:}K_s$, the greater this number of eliminated substitutions, and the greater is the selective constraint under which two genes evolve. The converse case of $K_a{:}K_s > 1$ indicates that replacement substitutions occur at a rate higher than expected by chance alone. It indicates that advantageous mutations occur in the evolution of two duplicates.

What fraction of duplicate genes gets lost after duplication? This question can be addressed by studying duplicates in different age classes K_s. One can bin closely related gene duplicates into several categories according to K_s. If gene duplications occur at an approximately constant rate, and if duplication products survive indefinitely, then each bin should contain the same number of gene pairs. But if genes get lost after duplication, the number of duplicates per bin should decrease with increasing K_s. The faster this number decreases, the greater the rate of gene loss. Michael Lynch and John Conery carried out such an analysis for multiple fully sequenced genomes. They found a rapid and nearly exponential decrease in the number of duplicates per bin, from which one can infer that more than 90% of duplicates get eliminated in the first 50 million years after duplication. This result is in good agreement with earlier predictions from theoretical models. Their observation suggests that the rate of gene duplication must be high to account for the many duplicate genes retained in eukaryotic genomes. This is indeed the case. Duplication rate estimates for fully sequenced eukaryotic genomes range from 0.002 (fruitfly) to 0.02 (nematode) per gene and million years. (An important caveat to reporting such average rates is that different genes may have very different duplication probabilities.)

What is the role of natural selection in the diversification of gene duplicates? Is selection absent immediately after selection, where two duplicates have identical functions, and where one might thus be eliminated without consequences? Is selection mostly 'purifying', eliminating deleterious variants in either gene? Or are there a large number of beneficial changes to the DNA sequences, changes that are driven to fixation by natural selection? The statistics of the ratio $K_a{:}K_s$ for many genes provides important information in this regard. They show that the vast majority of duplicate genes in both prokaryotic and eukaryotic genomes experience purifying selection. Even very closely related gene duplicates, duplicates no older than a few million years, experience selective constraints, as indicated by a ratio of $K_a{:}K_s < 0.5$ for such young duplicates. On the other hand, recent duplicates appear to tolerate up to ten times more replacement amino acid substitutions than older duplicates. Although the ratio of $K_a{:}K_s$ may vary widely according to gene family and organism, these statistics show that the vast majority of duplicates are under purifying selection, whose strength increases as duplicates age.

Whether beneficial mutations are frequent is a question more difficult to answer. Genome-scale studies are likely to have the most limited impact in answering this question, because the approach of finding genes with $K_a{:}K_s > 1$ does not work in general. While a genome may contain some duplicates with $K_a{:}K_s > 1$, the observed difference from unity does often not hold up to statistical scrutiny. Does this indicate the absence of positive selection after gene duplication? No. Positively selected amino acid substitutions often occur only in a small region of the coding region, too small to be detectable by an elevated $K_a{:}K_s$. Individual case studies have suggested positive selection for opsin visual pigments, members of primate ribonuclease genes, and triosephosphate isomerase, among others. They show that a strong case for positive selection generally requires not only information about gene divergence, but also about protein structure, protein function and phylogeny.

Gene Redundancy

Do many gene duplicates retain similar functions a long time after duplication? Such gene redundancy might shelter an organism from otherwise deleterious mutations in one of the duplicates, and may thus be sustained by natural selection for precisely this reason. However, population genetic theory suggests that the benefits of such redundancy for an organism are very weak, mainly because the probability that a mutation affects any particular pair of genes is very small. Only a very large population would experience a number of mutations sufficient for natural selection to sustain redundancy.

To find out whether many gene pairs retain redundant functions long after duplication is difficult, mainly because it cannot be accomplished with genome sequence information alone. Even gene duplicates with very similar DNA sequence may have undergone changes in key nucleotides that altered their protein products' function. To address this question, one thus needs to study biochemically characterized gene products of duplicate genes. Examples of biochemically characterized old gene duplicates with similar biochemical functions certainly exist. They include the budding yeast gene families encoding the catalytic subunits of cyclic adenosine monophosphate (cAMP)-dependent protein kinase, as well as the cyclin (CLN) cell cycle regulators. However, such examples are counterbalanced by many others where functional divergence (subtle or profound) has been rapid.

An observation potentially relevant to address this question is that many synthetic-null ('knockout') mutations of individual genes, mutations that completely eliminate a gene's function, have very weak phenotypic effects in standard laboratory assays. Often, the mutated genes are members of gene families, which raises the possibility that other family members compensate for the lost gene's function. Such an explanation is particularly attractive given the observation that genomes are full of duplicate genes. However, this explanation encounters some problems upon closer inspection. For instance, although more than one-third of the genes in the yeast *Saccharomyces cerevisiae* have exceedingly weak phenotypic effects when knocked out, almost one-half of these potentially redundant genes are not part of gene families. They are single copy genes. Furthermore, laboratory assays may not assess accurately whether a synthetic-null mutation has no effect on the organism. They measure the effects of loss-of-function mutations in only one or a few laboratory environments. While nearly neutral in such environments, loss-of-function mutations may have formidable effects in the wild. Furthermore, such assays are often not designed to assess subtle but evolutionarily important fitness differences between a mutant and its wild-type ancestor.

In sum, while gene redundancy certainly occurs, it is questionable whether it is sustained to protect an organism against mutations. Also, it cannot account for all or most cases of mutations with weak phenotypic effects.

RNA-coding Genes and Concerted Evolution

Genes coding for RNA, while accounting for only a small fraction of a genome, are involved in critical biological processes, such as translation and splicing. The most prominent RNA-coding genes are those encoding transfer RNAs (tRNAs) and ribosomal RNAs (rRNAs).

The total number of tRNA genes in a genome correlates with genome size. For instance, the human mitochondrial genome (1.7×10^4 bp), the genome of the bacterium *Escherichia coli* (4.6×10^6 bp) and the human nuclear genome (3×10^9 bp) have one complete set, (approximately) 100 complete sets and over 1000 complete sets of tRNA genes, respectively. tRNA genes are scattered throughout the genome and occur also as part of rRNA genes. During translation, tRNAs are responsible for associating one of 61 codons on an mRNA molecule with one of 20 amino acids. For one amino acid, there may thus be different tRNAs, recognizing different (synonymous) codons. Such tRNAs are called isoacceptor tRNAs. Because extant tRNAs are very similar in structure, much work has focused on the question of how they might have arisen from a common ancestor. Gene duplication has probably played an important role in this process. This holds especially for the evolution of isoacceptor tRNAs, where few nucleotide changes can be sufficient to transform one isoacceptor tRNA into another. However, alternative scenarios involving tRNA recruitment across isoacceptor families have also been proposed.

rRNAs are important components of ribosomes. Each ribosome contains several rRNAs which are transcribed from one gene as a precursor rRNA. Similar to tRNA genes, the number of rRNA genes also correlates highly with genome size. Prokaryotic genomes typically contain less than 10 rRNA genes, which are often dispersed throughout the genome. In contrast, eukaryotic genomes contain one or a few clusters of hundreds of tandemly arrayed rRNA genes. The yeast *S. cerevisiae* has 100–200 rRNA genes organized in a single tightly clustered array.

The human genome contains some 500 rRNA genes distributed among five such arrays. The individual rRNA genes in such large arrays are almost certainly redundant, in that their products, ribosomal RNAs, have identical functions. Their large numbers probably reflect a high demand for the gene product.

Especially in large DNA repeats, the individual repeat units no longer evolve independently. Instead they display concerted evolution, resulting in sequence similarity of rRNA genes that is greater within a species than among species. In other words, concerted evolution homogenizes the genes within an rRNA gene array. Two main mechanisms, gene conversion and unequal crossing-over, are thought to be responsible for concerted evolution.

Gene conversion is a byproduct of DNA recombination, where the DNA strand-breaks that initiate and conclude a DNA recombination event often occur at different positions within two recombining molecules. The result is that part of the sequence of one recombining molecule can become identical to that of the other molecule. Gene conversion is thought to act mainly over short stretches of DNA comprising several hundred base pairs. In the second mechanism, unequal crossing-over, one of the arrays participating in the recombination event contracts, i.e. it loses genes, whereas the other array expands (**Figure 1a**). It is thought that natural selection maintains an optimal gene number in an rDNA array, such that arrays that have become too short or too long are eliminated from a population. In repeated rounds of unequal crossing-over, array contractions result in the loss of divergent genes, whereas array expansions lead to the replacement of these genes with similar genes, and thus to homogenization of the array.

Which of these two processes is more important for the concerted evolution of rRNA gene arrays is unknown. It has been argued that unequal crossing-over should be prevalent, because it can lead to duplication or elimination of many genes at once. This is consistent with the observation that the number of rRNA genes in an array can vary widely among individuals in a species. However, some empirical evidence suggests an important role for gene conversion. In species hybrids, the homogenization of rRNA genes may occur preferentially in the direction of one hybrid, inconsistent with a mechanism of unequal crossing-over that would not generate such bias. In prokaryotic genomes, regions of exceptional similarity among (nonclustered) RNA genes often comprise short DNA stretches, consistent with gene conversion as a mechanism for homogenization.

Limits of Genome Analysis to Study Genome Evolution

Whole genome sequences provide a wealth of information about duplicate genes. They allow inferences about selection pressures, rates of gene loss and mechanisms of concerted evolution. However, they also have serious limitations. First, the information they provide is often not conclusive in deciding whether genes diversify through advantageous mutations generating new functions. Second, they cannot answer the question of how and if two gene duplicates have diverged in function. Third, many gene duplicates of equal age (time since duplication) show vastly different divergence on the amino acid level, indicating that different gene products are subject to very different selection pressures related to their function. Such differences cannot be understood without understanding protein function. To address these and other issues regarding gene duplications, genome sequence information – however valuable – must be complemented by detailed biochemical analyses of gene function.

Further Reading

Elder JF and Turner BJ (1995) Concerted evolution of repetitive DNA sequences in eukaryotes. *Quarterly Review of Biology* **70**: 297–320.

Hillis DM and Dixon MT (1991) Ribosomal DNA: molecular evolution and phylogenetic inference. *Quarterly Review of Biology* **66**: 410–453.

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

Liao DQ (2000) Gene conversion drives within genic sequences: concerted evolution of ribosomal RNA genes in bacteria and archaea. *Journal of Molecular Evolution* **51**: 305–317.

Lynch M and Conery JS (2000) The evolutionary fate and consequences of duplicate genes. *Science* **290**: 1151–1155.

Ohno S (1970) *Evolution by Gene Duplication*. New York: Springer.

Romero D and Palacios R (1997) Gene amplification and genomic plasticity in prokaryotes. *Annual Review of Genetics* **31**: 91–111.

Saks ME, Sampson JR and Abelson J (1998) Evolution of a transfer RNA gene through a point mutation in the anticodon. *Science* **279**: 1665–1670.

Wagner A (1999) Redundant gene functions and natural selection. *Journal of Evolutionary Biology* **12**: 1–16.

Wagner A (2000) Mutational robustness in genetic networks of yeast. *Nature Genetics* **24**: 355–361.