Multigene Families: Evolution

Secondary article

Article Contents

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

Wolfgang Stephan, *University of Munich, Munich, Germany*

Groups of genes showing similarity with each other are referred to as a gene family, arising from wholesale or partial gene duplication. The clustering of functionally related members of a gene family reflects their common ancestry and subsequent duplication and divergence. It has been speculated that gene duplication is essential for the creation of major evolutionary novelty. Genome projects will help in understanding the importance of gene duplications underlying macroevolutionary events.

Gene Families and Gene Clusters

A randomly chosen gene is likely to show varying degrees of similarity with a number of other genes within the genome. Groups of genes showing similarity with each other are referred to as a gene family, reflecting the assumption that all arose from a common ancestor. Similarly, groups of related gene families comprise a supergene family, again reflecting a common (although much older) ancestor to all members. Gene families are a direct consequence of the fact that essentially all new genes arise by gene duplication, either by wholesale duplication of entire genes or by duplication and shuffling of exons from different genes. Gene families provide information on how new genes arise and diversify. They are also of interest in that their organization may provide clues to how the genes in certain families are developmentally regulated. Finally, the presence of multiple copies of related genes generates novel evolutionary forces not seen when considering individual genes.

Gene families show enormous diversity in both the number of members and their genomic organization. Many families consist of just a few very similar genes, while others involve a large number of both closely related and more distant genes. Still others exist as hundreds of essentially identical copies. In terms of genomic organization, members of a gene family can be dispersed singly throughout the genome or can show varying degrees of clustering, forming gene clusters. A gene family can consist of several related gene clusters scattered throughout the genome. The most extreme type of clustering is when the entire family is restricted to one or a few tandem arrays where multiple (and essentially identical) copies are arranged together in a tandemly repeating pattern (with copies arrayed in a repeated head-to-tail fashion). More commonly, clustered members are separated by regions of unrelated DNA and the family members show varying degrees of divergence, often reflecting functional divergence.

While the term gene family is most often applied to protein-coding and structural RNA-coding genes, there are two other broad classes of repeated elements found within the genome. Mobile genetic elements (or transposons) are sequences that spread throughout the genome by duplicating themselves. A typical mobile genetic element consists of a stretch of DNA flanked by the appropriate *cis*-acting sequences and perhaps containing one or a few genes required for mobilization. The genomes of most organisms are littered with numerous families of active elements and the decaying remnants of other families that have lost their ability to mobilize. The second broad class of repeated elements are tandem repeats of noncoding DNA. These can consist of tens of repeats of a few nucleotide bases (microsatellites) up to thousands of much larger repeats (satellite DNAs). Such arrays are thought to be generated by errors during DNA replication and/or recombination. Once generated, an array can persist for a modest evolutionary time before being lost. Transposons and noncoding arrays are generally referred to as repetitive DNAs. While we will discuss some of the evolutionary features of such repetitive DNAs here, our main focus is on families coding for protein and structural RNAs.

Paralogous versus orthologous genes

Two fundamentally different types of comparisons can be made between gene family members from different species – those involving the same genes (orthologous comparisons) and those involving different, but related, genes (paralogous comparisons). Suppose we are comparing gene family members A and B in two species, where A and B arose from an ancient duplication that occurred prior to the ancestor of the two species. Let A_1 and B_1 denote these genes from the first species, A_2 and B_2 those from the second. The orthologous comparisons are A_1 versus A_2, and B_1 versus B_2. Differences observed in these comparisons are the result of changes that have occurred since the two species diverged from their common ancestor. In contrast, for the paralogous comparisons (A_1 versus B_2) and (A_2 versus B_1), much of the observed differences are due to the evolutionary divergence between A and B that occurred prior to the two species diverging from their common ancestor. For example, an ancient gene duplication around 400 million years ago in the jawed fishes produced the α- and β-globin families. Comparing human α-globin with chimp β-globin is an example of a paralogous comparison, as most of the differences between the genes are due to almost 400 million years of evolution before the humans and chimps split from their common ancestor. In contrast, the sequence changes in the orthologous comparison of human versus chimp α-globin are the result of the evolution that has occurred since the human and chimp diverged.

Pseudogenes

Pseudogenes are a common feature of multigene families. These are gene copies that cannot make a functional product (as a result of any number of mutations disrupting regulatory and/or coding regions). Pseudogenes fall into two distinct classes, normal and processed. Normal pseudogenes arise via direct duplication of a stretch of DNA and are typically linked to the gene of origin. There are no restrictions on the size of the duplicated gene and hence it can contain the coding region as well as flanking sequences (in which most of the regulatory regions reside). By contrast, processed pseudogenes are formed by reverse-transcribing an mRNA. As such, they encompass only the coding regions, lacking any regulatory sequences that reside outside the primary transcript. They are usually found on different chromosomes from their putative source genes. Processed pseudogenes are most likely inactive upon creation, unless they are extremely fortunate to be inserted next to appropriate gene regulatory sequences. Normal pseudogenes, on the other hand, often show evidence of being active for some reasonable period of evolutionary time before becoming inactive.

Examples of Gene Families

As the following examples show, families come in an incredible diversity of types, reflecting differences in their evolutionary history, in the roles of their products in the cell (making a large amount of product versus making a diversity of related products), and in their regulation.

The human prolactin family

An example of a typical small gene family is provided by the prolactin genes of placental mammals, which consist of prolactin, growth hormone (GH), and placental lactin (PL) genes. Different family members are expressed in different tissues (GH in the pituitary gland, PL in the placenta), and hence these genes are under different developmental controls. Humans have one prolactin, two GH and three PL genes, with the GH and PL genes tightly linked together. The GH and PL genes show about 90% sequence similarity to each other at the nucleotide level, while the similarity at the amino acid level of either GH or PL to prolactin is about 65%. How did this family evolve? Clues are offered both from sequence similarity and from comparison of the composition of this gene family in other more distantly related vertebrates. Around 400 million years ago, an ancestral vertebrate gene duplicated and diverged, giving rise to the prolactin and GH genes. Around 90 million years ago (shortly after the appearance of placental mammals), a second duplication/divergence event occurred wherein a GH-like ancestor duplicated to give rise to a GH and PL gene (which were tightly linked together). About 60 million years ago, the GH–PL cluster itself duplicated in higher primates, giving rise to multiple linked copies of GH and PL genes seen in humans.

Histone and tRNA genes

A very different type of gene family organization is seen with the histone and tRNA genes. In both cases the families consist of numerous repeated members that are unrelated to each other. The histone genes encode DNA-binding proteins involved in the packaging of eukaryotic chromosomes into nucleosomes. The H2A, H2B, H3 and H4 histone genes are highly conserved and present in all eukaryotes, while the H1 gene is absent in many lower eukaryotes. Even though each of these five histones are unrelated to each other, we still refer to a histone family because each different class of histone genes has multiple copies, which often reside in clusters with the other histone genes. A striking example is seen in some species of sea urchins, which have early histones (required in large amounts in the early embryo) and late histones (used later in development). The early genes exist as around 500 tandem repeats of the cluster H1–H4–H2B–H3–H2A. By contrast, there are approximately ten copies of each late gene and, although some exist in pairs with other histones, many are scattered as single loci throughout the genome. The arrangement of histone genes is similarly diverse as one looks across different species. While clusters containing all five histone genes are common, other clusters contain incomplete sets, and still other histone genes exist as individual copies unlinked to other family members.

Genes encoding transfer RNAs have a similar family organization. Each of the 30–40 types of tRNA genes exists as tens to hundreds of copies in higher eukaryotes, with a typical higher eukaryote containing upwards of a thousand tRNA genes (some of which may be pseudogenes). There is no set genomic organization for tRNA genes as, even within the same genome, there are clusters containing numerous different types of tRNA genes as well as many individual tRNA genes scattered throughout genome.

rDNA genes

Perhaps the gene family with the most conserved structure is the ribosomal RNA family. The primary (18S, 28S) rRNAs used in the eukaryotic ribosome are transcribed from a single unit (the rDNA gene), with the initial transcript subsequently processed to yield the individual rRNAs. For most genomes, rDNA genes exist in large tandem arrays. Many organisms (such as yeast, corn and the nematode *Caenorhabditis elegans*) have a single array, while others have several (humans have arrays existing on the tips of five different chromosomes). However, all functional rDNA genes appear to be confined to tandem arrays.

Developmental Regulation and Multigene Families

The clustering of functionally related members of a gene family reflects to some extent their common ancestry and subsequent duplication and divergence. While many such families eventually have some (or all) of their members dispersed throughout the genome, others have kept their clustered organizations for long periods of evolutionary time. Such conservation of organization may reflect constraints imposed by proper regulation of different members. One potential example of this is seen in the globin genes, in which genes are sequentially expressed along a cluster as the organism develops. An example wherein the organization of gene family members is absolutely critical is given by the immunoglobulin genes.

The globin gene family

Members of the globin gene family are involved in oxygen transport and storage. While there are globin-like genes in plants (leghaemoglobin), the family has been best studied in vertebrates, where tetrapods have three subfamilies – myoglobin and the α- and β-haemoglobins. Higher vertebrates have a single myoglobin gene and clusters of α and β genes. Myoglobin is largely used for oxygen storage, while two α- and two β-globins form the tetrameric haemoglobin protein used in oxygen transport. During mammalian development, different forms of haemoglobin are used, reflecting differences in the oxygen transport requirements of the early embryo, fetus and adult. These different haemoglobins result from the expression of different members of the α and β families.

The evolutionary diversification of the α and β families can be followed by looking across successively more recent vertebrates. The primitive hagfish has only a single globin (myoglobin), while frogs also show single α and β genes (which are linked together). In mammals and birds, the α–β linkage is broken and several duplications have occurred, resulting in separate unlinked α and β clusters containing several related genes. In humans, all functional genes in both clusters are transcribed in the same direction and are expressed developmentally in a serial fashion (the earliest expressed genes being at one end of the cluster, the latest expressed at the opposite end). The human α gene cluster spans roughly 30 kb and consists of zeta (ζ) globin, three pseudogenes (one related to zeta, ψζ; two related to alpha, ψα), two α-globins (α2, α1) and theta (θ)-globin, giving the cluster organization ζ–ψζ–ψα–ψα–α2–α1–θ (**Figure 1a**). The zeta gene arose by duplication of an ancestral α about 300 million years ago (Mya), while the θ gene arose around 250 Mya. The β gene cluster in humans spans around 50 kb and consists of an epsilon (ε), two gamma (Gγ , Aγ), a pseudogene (related to beta, ψβ), a delta- (δ), and a β-globin (β) for a gene order of ε–Gγ–Aγ–ψβ–δ–β (**Figure 1b**). The (ε–γ) and (δ–β) families resulted from a duplication around 175 Mya, with δ and β resulting from a duplication around 40–80 Mya, while ε and γ resulted from a duplication around 100 Mya. The order of functional genes with each globin cluster corresponds to the order of gene expression during development. Human embryonic haemoglobin largely consists of ζ–ε tetramers; fetal haemoglobin consists of α–γ tetramers, while adult haemoglobins consist of the α–β and α–δ tetramers (the role of the θ gene is unclear).

Immunoglobulin genes

Vertebrates have an immune response system tasked with producing an antibody response to an almost infinite number of antigens. Antibodies are tetramers formed by the products of the heavy and light chain immunoglobulin genes. Any particular antibody consists of two identical light and two identical heavy chains joined to form a Y-shaped molecule. Antigen binding occurs at the tips of the Y, where highly variable regions on both the light and heavy chains interact to form the antigen-binding pocket. Variation in the variable regions of each chain is generated

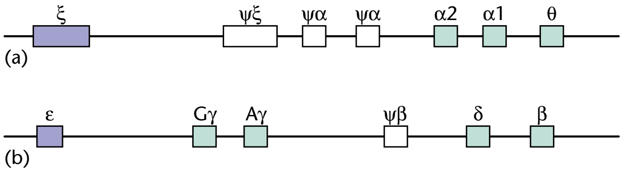

Figure 1 Order and arrangement of the human globin genes. (a) Ã globin cluster, human chromosome 16; (b) ã globin cluster, human chromosome 11. Coloured boxes represent functional genes, white boxes are pseudogenes.

from clusters of linked gene segments that are joined in different combinations in different antibody-producing cells. This organization of the immunoglobulin genes allows an organism to produce literally millions of different types of antibodies, allowing the potential to recognize the vast array of novel antigens that an organism may face.

To illustrate these points, consider the organization of the lambda (λ) light chain gene. The final λ protein is produced by joining a V gene segment with a JC segment, where the V and JC elements exist in tandem arrays. In humans, there are around 300 V elements linked to roughly 10 copies of JC elements, giving the lambda gene structure as V_1–V_2…V_{300}–JC_1–JC_2…JC_{10}. Recombination between a particular V and JC unit removes the intervening segments, producing a functional light chain gene. By joining different segments in a combinatorial fashion, this gene structure encodes for a tremendous amount of variation, with the potential for over 3000 different V–JC combinations. This same type of organization occurs in the heavy chain gene. Since antigen binding is determined by light–heavy chain interactions, similar combinatorial variation in the heavy chain results in over $3000^2 = 9\,000\,000$ potential antibodies. This same strategy of creating diversity through the combinatorial use of tandemly arrayed gene segments is also used by the T-cell receptor genes (which are involved in the control of antibody production). The heavy, light and T-cell receptor genes all share sequence similarity and comprise (along with other immune system genes) a major supergene family.

The λ gene in chickens uses an interesting variation of this combinatorial generation of diversity. In chickens, there is only one functional V segment linked to a single functional J and C gene. However, upstream of the functional V segment are around 30 V pseudogenes, existing in different orientations and missing the complete sequence at one (or both) ends of the gene segment. Diversity is generated by apparent short (10–100 bp) gene conversion events (see later) between these pseudogenes and the functional V gene.

Processes of Evolution in Multigene Families

Concerted evolution

The special evolutionary feature of multigene families – concerted evolution – arises as a consequence of recombination events (such as gene conversion and unequal crossing-over) between family members showing sequence similarity. As a result of repeated rounds of such sequence exchanges, some gene families appear to evolve in concert (hence the term concerted evolution), with members within a species being more similar to each other than they are to members of closely related species.

A classic example of concerted evolution is the ribosomal RNA gene clusters in two species of the frog *Xenopus*. The rRNA genes are encoded in a long tandem array, with adjacent genes being separated by intergenic spacer (IGS) regions. Within each species, the IGS regions are very similar, yet they are quite different between the species. The actual rRNA genes themselves show little change, either within or between species. Differences in the IGS regions imply that mutations have been fixed between the species. But what accounts for the within-species homogenization? If each family member behaved as an independent locus, the mutations observed as being fixed between the two species would have to arise independently at hundreds of copies. It thus appears as if all family members evolve more like a single locus than independent copies, and the most likely explanation for this is cycles of unequal crossing-over.

To see how multiple cycles of unequal crossing-over can result in concerted evolution, consider a cluster of six tandemly repeated gene family members, denoted by 1–2–3–4–5–6. A round of unequal crossing-over can give rise to an array with (say) genes 1–2–3–3–4–5–6. Several additional rounds could give rise to a cluster of (say) 1–3–3–3–3–6. The original six different genes have now been reduced to three. Repeated cycles of unequal crossing-over generates a resampling process that is very similar to genetic drift and results in the members of a gene family having all their members tracing back to a single recent ancestral copy. For the *Xenopus* example, the common ancestor of all IGS elements within each species is far more recent than the time of divergence between species. Thus the two frog species show divergence from each other, yet the IGS copies within each species remain rather similar to each other. Even if gene family members do not exist as tandem arrays, resampling of family members can occur via gene conversion (the nonreciprocal exchange between similar DNA sequences) which can occur between sequences even on different chromosomes.

Both unequal crossing-over and gene conversion have potential restrictions on their ability to homogenize a gene family. Unequal crossing-over generally requires that gene family members occur in clusters and with the same orientation, while gene conversion has no such restriction. While unequal crossovers can occur between similar sequences on nonhomologous chromosomes, the resulting products are chromosome translocations, which are often deleterious. Even if restricted to exchanges within a cluster, unequal crossing-over changes the number of genes, which can have potentially deleterious effects if the genes are sensitive to dosage effects. While conversion has fewer restrictions in terms of gene position, it has the disadvantage that a single conversion event typically homogenizes only part of a gene (as conversion tracks are usually much shorter than the average gene). Unequal crossing-over, on the other hand, often involves units of entire genes.

Concerted evolution can also occur in dispersed families of mobile genetic elements. While gene conversion can contribute to this, a more likely scenario is that elements are being deleted from the genome while new copies are created by transposition. This is also a resampling process akin to genetic drift and (again) results in all genomic copies being the descendants of a single ancestral copy. If this ancestor is more recent than the divergence time between the species being compared, the within-species copies will be more similar than the between-species copies. Likewise, cycles of unequal crossing-over can result in concerted evolution in families of noncoding arrays.

Selection on multigene families

For a gene family experiencing weak selection (as is likely to be the case for many noncoding tandemly repeated DNA families), concerted evolution is expected to be the norm, unless the family members have diverged sufficiently to stop sequence exchange. For a family under moderate to strong selection, concerted evolution can be either helped or hindered by selection. In a family consisting of functionally distinct members, selection counters the effects of concerted evolution. Conversely, in a family of functionally identical genes, the presence of multiple copies diffuses the amount of selection on any given locus. Here, the forces of concerted evolution can greatly increase the effectiveness of selection.

There are several human examples of potentially deleterious mutations caused by exchange events between functionally distinct members of gene families. One example is red–green colour blindness, which results from the lack of both functional red and green opsin genes. These genes are the result of a recent duplication and are tightly linked on the X chromosome. Unequal crossing-over between them can result in an X chromosome missing one of these opsins. Sequence exchange also accounts for many cases of thalassaemia (blood diseases caused by the reduction or absence of one of the haemoglobin chains). Unequal crossing-over can again remove functional copies, creating chromosomes with deficiencies in certain α or β genes. Likewise, exchange events occurring within two genes can create fusion proteins that function poorly. α-Thalassaemias are more common than β-thalassaemias, which may be a consequence of the larger intron size and more divergent flanking sequences found in the β genes. Both of these reduce the amount of sequence similarity, which can reduce the rate of sequence exchange.

While sequence exchange between the coding regions of functionally distinct members can result in deleterious mutations, concerted evolution can still occur in noncoding regions. The IGS regions of *Xenopus* rRNA genes discussed above are a good example. Likewise, striking convergence in the intron sequences of otherwise functionally distinct genes has been seen in several families. Presumably, selection against homogenization events is at best weak in such regions. Against a background of sequence exchange trying to homogenize sequences, natural selection acts as a filter that lets some events through while removing others.

Although concerted evolution has little beneficial effect in clusters of functionally distinct genes, it facilitates selection acting to maintain functionality in a gene family consisting of a large number of functionally identical genes. If selection acted independently on each member, one would expect a fair number of copies to be inactive owing to the diffuse nature of selection acting on hundreds of identical copies. However, such families often contain a very high percentage of functional copies and this is probably the result of concerted evolution interacting with selection. Cycles of sequence exchange (especially unequal crossing-over acting within a tandem array) will either amplify a rare mutation up to higher frequencies (i.e. spread it across different family members) or remove it entirely. If the new mutation is deleterious, increasing its frequency will allow selection to act more efficiently in removing it. If a new mutation is advantageous, increasing its frequency will allow selection to act more efficiently in spreading it throughout a family. The efficiency of natural selection depends on the variation in fitness among family members. The forces of concerted evolution, by randomly increasing and decreasing the number of family members carrying a particular variant, inflate the variance in fitness and increase the efficiency of selection.

Finally, what role does selection play on the organization of gene families? In situations where there are regulatory and/or developmental constraints, selection may act to conserve the organization of a cluster. In the absence of such forces, the members of a cluster may be slowly dispersed over the genome as such events may be selectively neutral.

Evolution of new gene functions

There has been much speculation that gene duplication is essential for the creation of major new evolutionary novelties, such as complex developmental switches. The basis for such thinking is that it may be very difficult to modify an existing gene that acts deep within a developmental programme owing to very strong selective constraints. However, by duplicating a copy, one can keep the original function while allowing the duplicate copy to be removed from such constraints, allowing it potentially to be used as raw material for new novelties. Sequence comparisons show that it is not unusual for a recently duplicated gene to exhibit a burst of very rapid sequence change followed by long periods of much slower (and more normal) rates of change. Such bursts of change may reflect either an initial relaxation of selection constraints and/or directional selection for particular changes in the duplicate copy. The eventual slower rates of response are consistent with the duplicate copy acquiring a new gene function and hence itself being under selective constraints on which further changes can occur. As the current genome programmes reach their goals of providing us with the complete sequences of numerous diverse organisms, we will certainly be in a better position to resolve the relative importance of gene duplications underlying major macroevolutionary events.

Further Reading

Brookfield J (1992) Can genes be truly redundant? *Current Biology* **2**: 553–554.

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

Li W-H and Graur D (1991) *Fundamentals of Molecular Evolution*. Sunderland, MA: Sinauer.

Ohta T (1980) *Evolution and Variation of Multigene Families*. Berlin: Springer-Verlag.

Ohta T (1987) A model of evolution for accumulating genetic information. *Journal of Theoretical Biology* **124**: 199–211.

Ohta T (1988) Further simulation studies on evolution by gene duplication. *Evolution* **42**: 375–386.

Ohta T (1988) Multigene and supergene families. *Oxford Surveys in Evolutionary Biology* **5**: 41–65.

Ohta T (1991) Multigene families and the evolution of complexity. *Journal of Molecular Evolution* **33**: 34–41.

Ruddle FH, Bartles JL, Bentleys KL *et al*. (1994) Evolution of *Hox* genes. *Annual Review of Genetics* **28**: 423–442.

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

Singer M and Berg P (1991) *Genes and Genomes*. Mill Valley, CA: University Science Press.

Tachida H and Kuboyama T (1988) Evolution of multigene families by gene duplication: a haploid model. *Genetics* **149**: 2147–2158.

Walsh JB (1985) Interaction of selection and biased gene conversion in a multigene family. *Proceedings of the National Academy of Sciences of the USA* **82**: 153–157.

Walsh JB (1985) How many processed pseudogenes are accumulated in a gene family? *Genetics* **110**: 345–364.

Walsh JB (1987) Sequence-dependent gene conversion: can duplicated genes diverge fast enough to escape conversion? *Genetics* **117**: 543–557.

Walsh JB (1995) How often do duplicated genes evolve new functions? *Genetics* **139**: 421–428.

Yokoyama S (1997) Molecular genetic basis of adaptive selection: examples from color vision in vertebrates. *Annual Review of Genetics* **31**: 315–336.