Gene Families: Multigene Families and Superfamilies

Chapter · March 2008

DOI: 10.1002/9780470015902.a0005126.pub2

Tomoko Ohta
National Institute of Genetics

Gene Families: Multigene Families and Superfamilies

Advanced article

doi: 10.1002/9780470015902.a0005126.pub2

Tomoko Ohta, *National Institute of Genetics, Mishima, Japan*

Multigene families are defined as groups of genes with sequence homology and related overlapping functions, whereas superfamilies are defined as a group of proteins or genes of common origin with nonoverlapping functions. If a group of proteins or genes contains a domain of common origin, it is a superfamily; a protein or a gene may belong to two or more superfamilies.

Overview

In the genomes of higher organisms, there are numerous gene families, from large to small, and from those with uniform members to those with diverse members. Research on multigene families started in the 1970s, arising mainly through a curiosity about how the genetic variability generated in an individual or existing in a population is related to gene redundancy. Smith (1974) and Hood *et al.* (1975) studied the evolution of immunoglobulins and showed that this and other families are characterized by concerted evolution; they suggested that unequal crossing-over and gene conversion are contributing to immunoglobulin evolution. To understand the evolution and variation of gene families, therefore, these illegitimate crossing-over events need to be considered along with mutation, selection and random drift. Sequence homology among members of a gene family mainly depends on the relative rate of occurrences of mutation and illegitimate crossing-over events, and is influenced to some extent by selection and drift (Ohta, 1983).

An important feature of a superfamily is the different expression patterns of the gene members. By contrast, genes belonging to a multigene family are usually under common regulatory control. It will, therefore, be interesting to clarify the process of how diverse expression patterns are acquired. In view of the rapid progress in our understanding of genetic regulatory systems, this issue will be a most prominent one in the coming years.

As shown by the examples given later, a large superfamily can contain single genes and several multigene families. Formation of these families must have been fundamental in organism evolution. Almost all genes belong to one or more gene families, and homology search is popular as the first step for identifying gene relationships. Phylogenetic analyses are useful for understanding relationships of member genes of a gene family. Gene trees may clarify the history of gene duplication events and can provide information for identifying orthologues in comparative studies of homologous genes using data from many species. **Table 1** gives some examples of large superfamilies in the human genome. **See also**: Gene Families

Table 1 Examples of large superfamilies in the human genome

Domain	Number of proteins
Developmental and homeostatic regulators	
Cadherin	100
Transforming growth factor β-like domain	27
Hemostasis	
Sushi domain	53
Immune response	
Immunoglobulin domain	381
PI-PY-Rho GTPase signalling	
C2 domain	73
FYVE zinc-finger	28
PH domain	193
Ras family	126
Src homology 2 domain	87
Src homology 3 domain	143
Extracellular matrix adhesion	
EGF-like domain	108
Fibronectin type III domain	106
Protein interaction	
Ank repeat	145
EF hand	83
PH domain	193
WD40 domain	136
Nuclear interaction	
BTB/POZ domain	97
Helix–loop–helix DNA-binding domain	60
Homeobox domain	160
KRAB box	204
RNA recognition motif	224
Zinc-finger, C2H2 type	564
Zinc-finger, C3HC4 type	135

Note: Data are taken from Table 18 of Venter *et al.* (2001).

Genomic Organization of Gene Families

Clustered gene families

Many important gene families exist as clusters, that is, gene members are repeated in tandem in the genome. They have arisen by unequal crossing-over during meiosis or mitosis of germ cell lineages. The large copy number of members of some multigene families, such as ribosomal ribonucleic acid (RNA) or histone families, is due to a need for large amounts of gene product.

Genes of ribosomal RNA usually form a multigene family. In many eukaryotes, a repeating unit contains a transcribed region and a nontranscribed spacer, and the former encodes 18S, 5.8S and 28S ribosomal RNAs. This unit is repeated hundreds of times. This multigene family is characterized by concerted evolution such that homology among repeating units is high even though they show divergence when compared with the sequences of different species (Eickbush and Eickbush, 2007).

In regard to the histone multigene family, four core histone genes are usually clustered in higher organisms and this cluster is repeated tens of times. These genes are expressed as a cohort during the S phase of the cell cycle. There are also a few histone genes that are expressed independently from the cell cycle. These replication-independent genes may behave like ordinary tissue-specific genes.

Some clustered gene families have more diverse functions than the earlier examples. The best-known case is the globin superfamily. The globin superfamily of mammals consists of three families, that is, the α-like family, the β-like family and myoglobin. The first two families are each encoded by clusters of genes, whereas myoglobin is encoded by a single gene. Each cluster contains the embryonic and adult-type together with pseudogenes. Many adult genes exist in multiples, for example, two genes, α1 and α2, encode adult human α-haemoglobin. These two genes may be said to form a small multigene family and are characterized by concerted evolution. **See also**: Gene Duplication: Evolution

Superfamilies with clustered and dispersed members

Superfamilies often contain both clustered and dispersed members. The clustered genes usually form multigene families with overlapping functions, whereas the dispersed ones may have more diverse functions. In some cases, however, clustered genes may have distinct functions as exemplified by the *Hox* genes. **See also**: *Hox* Genes: Embryonic Development

The immunoglobulin superfamily includes both clustered and dispersed members, with complicated organization in the human genome. It contains members of diverse functions. Many of its members contain domains other than the immunoglobulin domain and are therefore multifunctional. The largest family is the immunoglobulin family, which codes for polypeptides that form antibodies in the bloodstream. Genes of immunoglobulins are known to be encoded by variable (V), diversity (D), joining (J) and constant (C) segments.

Several copies of V, D and J segments exist, and the enormous diversity of immunoglobulins is generated by the combinatorial usage of V, D, J and light and heavy chains. Somatic mutation also contributes to generating diversity. To have efficient generation of diversity by combinatorial usage, however, mutant accumulation in evolution is essential. Detailed analyses of variable region gene families suggest that natural selection for enhancing diversity at the crucial regions of antigen recognition has been operating.

Gene families of the major histocompatibility complex (MHC) belong to the immunoglobulin superfamily. They have attracted much interest because of associated clinical problems and exceptionally high polymorphisms. The MHC has been suggested to be evolving dynamically under various illegitimate recombinations, as alleles have been found that differ by a short segment of sequence that has been apparently converted by a homologous gene. **See also**: Major Histocompatibility Complex (MHC)

Detailed analyses by sequence comparison show that natural selection that enhances amino acid diversity at the antigen recognition site is important for maintaining high polymorphisms (Hughes and Nei, 1988). Simultaneously, gene conversions including the antigen recognition site are generating useful variability for selection to work.

Another remarkable example of a gene family with clustered and dispersed members is that of the olfactory receptor. This receptor is a seven-transmembrane protein and belongs to the large G-protein-coupled receptor (GPCR) superfamily. Functional diversity is encoded in the genome without combinatorial usage or somatic mutation. It is estimated that there are several hundred GPCR genes in mammals; these genes are organized in many clusters and one cluster may have diverse members. This family also contains many pseudogenes. The catfish has a much smaller number of olfactory receptor genes than mammals, and the gene family is thought to have expanded in the ancestral lineage of mammals.

Dispersed gene families

Many of the dispersed gene families are thought to have been formed by reverse transcription of RNA and subsequent integration into the genome. The integrated sequence, or 'retrosequence', is derived from the RNA transcript of a gene, and therefore does not contain introns. It is likely that most retrosequences degenerate and become pseudogenes; however, there are several interesting cases of retrosequences that have retained function or acquired new functions (Long *et al*., 2003). A functional retrosequence is called a retrogene or processed gene. The autosomal gene of the human phosphoglycerate kinase has no intron and is an example of the retrogene. It is interesting that the expression pattern of this gene is different from that of the original gene on the X chromosome. As expected, there are many cases where retrosequences become retropseudogenes, including argininosuccinate synthetase, β-tubulin, cytochrome *c*, glyceraldehyde-3-phosphate dehydrogenase and ribosomal protein L32.

Many noncoding retrogenes are known. For example, substantial fraction of human microRNAs originate from retroposed repetitive elements (Piriyapongsa *et al.*, 2007).

Forces that Shape Gene Families

Gene duplication, point mutation and retrotransposition

The various processes of gene duplication include polyploidization, tandem (segmental) duplication and retrotransposition. All genes in a genome duplicate during polyploidization, whereas a much smaller region is duplicated in the last two processes. There are two types of polyploidization: autopolyploidization, in which the same genome is duplicated; and allopolyploidization, in which two closely related genomes are duplicated, often by hybridization between two species followed by duplication of the whole set of chromosomes. Polyploidization has contributed to the formation of many gene families. Tandem duplication is responsible for the evolution of clustered gene families. Duplication may include a smaller or larger deoxyribonucleic acid (DNA) region than the size of a gene. For gene family evolution, however, duplication of a whole gene is most common.

Tandem duplication is caused by unequal crossing-over during meiosis or mitosis in a germ cell lineage. Once a gene cluster is formed, the rate of unequal crossing-over becomes high. In multigene families with uniform members, the high rate of unequal crossing-over is thought to be responsible for their concerted evolution. In general, the balance between diversification by point mutation and homogenization by unequal crossing-over is important for determining genetic variability contained in a gene family. Selection and drift may also have significant effects on gene diversity of a family.

Repetitive sequences derived from reverse transcription are numerous in the human genome. The most abundant repetitive sequences are short interspersed nuclear elements (SINEs) and long interspersed nuclear elements (LINEs). Among these, Alu repeats are the most profuse, with more than 500 000 in the genome. The human Alu family has sequence similar to 7SL RNA, which is abundant in cytoplasm. In other species, there is homology between transfer RNA and SINEs, and small RNAs with a polymerase III promoter are thought to be the templates for SINEs.

LINEs in the human genome have a reverse transcriptase gene and can therefore retrotranspose by themselves. Their copy number is about 100 000 in the human genome. Most LINE sequences are truncated and not functional.

These classes of repetitive sequences indicate that retrotransposition may be efficient if a condition is met such that appropriate RNA transcripts and reverse transcriptase activity are available in a germ cell lineage. Most retrosequences become retropseudogenes. However, retrotransposition is an important mechanism for creating new genes in new genomic positions. When a retroposed gene copy acquires a new regulatory element, a chimaera gene may appear that is useful for the organism (Long *et al.*, 2003). **See also**: Homologous, Orthologous and Paralogous Genes; Human Genetic Diversity

Natural selection and random drift

Any duplicate genes are under purifying selection if they are expressed and functional. The rate of their evolution is therefore lower than the mutation rate of single genes. But the degree of such a selective constraint may differ between duplicate and single genes. It is expected that the constraint weakens with gene redundancy, because deleterious mutations may accumulate as long as at least one gene remains functional. For large multigene families, such as those of ribosomal RNA and histones, accumulation of deleterious mutations will be prevented by purifying selection that detects the number of intact and functional genes. Concerted evolution that increases or decreases the number of gene copies with deleterious mutations is helpful for selection to work.

Positive selection is thought to be operating when duplicate genes acquire new functions. In fact, immediately after their origination many duplicate genes show an acceleration of amino acid substitution, which can be detected by calculating the rates of synonymous and nonsynonymous divergence. Examples of such genes include embryonic haemoglobin in primates, stomach lysozyme of ruminants and visual pigments of mammals. In some cases, however, it is difficult to determine whether or not an increased substitution rate is due to positive selection. The acceleration may simply be caused by relaxation of the selective constraint owing to gene redundancy.

Another significance of gene duplication is the differentiation of expression patterns, that is, the subfunctionalization or specialization of duplicate genes. This is particularly important for transcription factors and other proteins that participate in developmental pathways, as changes are related directly to morphological evolution.

Analyses of the regulatory element of a transcription factor from *Drosophila melanogaster* indicate that the sequence of the element is turning over while the function is kept under stabilizing selection. Here again, both random drift and selection are operating. When the stabilizing selection is not strong, there is a chance for the transcription factor to be replaced by another, newly recruited factor, and a change in expression pattern takes place. Interaction of drift and selection is thought to be working during such a transition stage. **See also**: Gene Families; Gene Families: Formation and Evolution

References

Eickbush TH and Eickbush DG (2007) Finely orchestrated movements: evolution of the ribosomal RNA genes. *Genetics* **175**: 477–485.

Hood L, Campbell JH and Elgin SCR (1975) The organization, expression and evolution of antibodies and other multigene families. *Annual Reviews in Genetics* **9**: 305–353.

Hughes AL and Nei M (1988) Pattern of nucleotide substitution at major histocompatibility complex class-I loci reveals overdominant selection. *Nature* **335**: 167–170.

Long M, Betran E, Thornton K and Wang W (2003) The origin of new genes: glimpses from the young and old. *Nature Review Genetics* **4**: 865–875.

Ohta T (1983) On the evolution of multigene families. *Theoretical Population Biology* **23**: 216–240.

Piriyapongsa J, Marino-Ramirez L and Jordan IK (2007) Origin and evolution of human microRNAs from transposable elements. *Genetics* **176**: 1323–1337.

Smith GP (1974) Unequal crossover and the evolution of multigene families. *Cold Spring Harbor Symposia on Quantitative Biology* **38**: 507–513.

Venter JC, Adams MD, Myers EW *et al*. (2001) The human genome. *Science* **291**: 1304–1351.

Further Reading

International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. *Nature* **409**: 860–921.

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

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

Ohta T (2000) Evolution of gene families. *Gene* **259**: 45–52.

Wagner A (2001) Birth and death of duplicate genes in completely sequenced eukaryotes. *Trends in Genetics* **17**: 237–239.