CLUSTERS AND REPEATS

1.4.1 Introduction

Key Terms

A **gene family** consists of a set of genes whose exons are related; the members were derived by duplication and variation from some ancestral gene.

A **translocation** is a rearrangement in which part of a chromosome is detached by breakage or aberrant recombination and then becomes attached to some other chromosome.

A **gene cluster** is a group of adjacent genes that are identical or related.

Nonreciprocal recombination (unequal crossing-over) results from an error in pairing and crossing-over in which nonequivalent sites are involved in a recombination event. It produces one recombinant with a deletion of material and one with a duplication.

Satellite DNA (Simple-sequence DNA) consists of many tandem repeats (identical or related) of a short basic repeating unit.

Minisatellite DNAs consist of ~10 copies of a short repeating sequence. the length of the repeating unit is measured in 10s of base pairs. The number of repeats varies between individual genomes.

A set of genes descended by duplication and variation from some ancestral gene is called a **gene family**. Its members may be clustered together or dispersed on different chromosomes (or a combination of both). Genome analysis shows that many genes belong to families; the 40,000 genes identified in the human genome fall into ~15,000 families, so the average gene has a couple of relatives in the genome (see **Figure 3.15**). Gene families vary enormously in the degree of relatedness between members, from those consisting of multiple identical members to those where the relationship is quite distant. Genes are usually related only by their exons, with introns having diverged (see *Molecular Biology 1.2.5 Exon sequences are conserved but introns vary*). Genes may also be related by only some of their exons, while others are unique (see *Molecular Biology 1.2.10 Some exons can be equated with protein functions*).

The initial event that allows related exons or genes to develop is a duplication, when a copy is generated of some sequence within the genome. Tandem duplication (when the duplicates remain together) may arise through errors in replication or recombination. Separation of the duplicates can occur by a **translocation** that transfers material from one chromosome to another. A duplicate at a new location may also be produced directly by a transposition event that is associated with copying a region of DNA from the vicinity of the transposon. Duplications may apply either to intact genes or to collections of exons or even individual exons. When an intact gene is involved, the act of duplication generates two copies of a gene whose activities are indistinguishable, but then usually the copies diverge as each accumulates different mutations.

The members of a well-related structural gene family usually have related or even identical functions, although they may be expressed at different times or in different cell types. So different globin proteins are expressed in embryonic and adult red blood cells, while different actins are utilized in muscle and nonmuscle cells. When genes have diverged significantly, or when only some exons are related, the proteins may have different functions.

Some gene families consist of identical members. Clustering is a prerequisite for maintaining identity between genes, although clustered genes are not necessarily identical. **Gene clusters** range from extremes where a duplication has generated two adjacent related genes to cases where hundreds of identical genes lie in a tandem array. Extensive tandem repetition of a gene may occur when the product is needed in unusually large amounts. Examples are the genes for rRNA or histone proteins. This creates a special situation with regards to the maintenance of identity and the effects of selective pressure.

Gene clusters offer us an opportunity to examine the forces involved in evolution of the genome over larger regions than single genes. Duplicated sequences, especially those that remain in the same vicinity, provide the substrate for further evolution by recombination. A population evolves by the classical recombination illustrated in **Figure 1.31** - **Figure 1.32**, in which an exact crossing-over occurs. The recombinant chromosomes have the same organization as the parental chromosome. They contain precisely the same loci in the same order, but contain different combinations of alleles, providing the raw material for natural selection. However, the existence of duplicated sequences allows aberrant events to occur occasionally, changing the content of genes and not just the combination of alleles.

Unequal crossing-over describes a recombination event occurring between two sites that are not homologous. The feature that makes such events possible is the existence of repeated sequences. **Figure 4.1** shows that this allows one copy of a repeat in one chromosome to misalign for recombination with a different copy of the repeat in the homologous chromosome, instead of with the corresponding copy. When recombination occurs, this increases the number of repeats in one chromosome and decreases it in the other. In effect, one recombinant chromosome has a deletion and the other has an insertion. This mechanism is responsible for the evolution of clusters of related sequences. We can trace its operation in expanding or contracting the size of an array in both gene clusters and regions of highly repeated DNA.

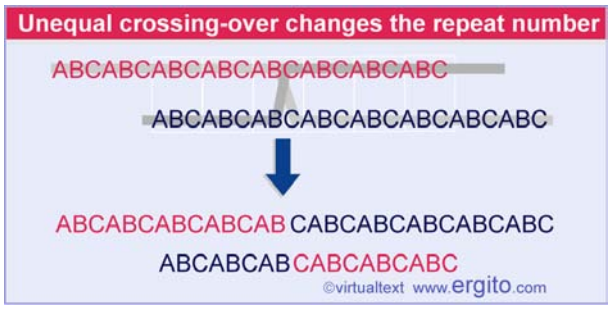


Figure 4.1 Unequal crossing-over results from pairing between non-equivalent repeats in regions of DNA consisting of repeating units. Here the repeating unit is the sequence ABC, and the third repeat of the red allele has aligned with the first repeat of the blue allele. Throughout the region of pairing, ABC units of one allele are aligned with ABC units of the other allele. Crossing-over generates chromosomes with 10 and 6 repeats each, instead of the 8 repeats of each parent.

The highly repetitive fraction of the genome consists of multiple tandem copies of very short repeating units. These often have unusual properties. One is that they may be identified as a separate peak on a density gradient analysis of DNA, which gave rise to the name **satellite DNA**. They are often associated with inert regions of the chromosomes, and in particular with centromeres (which contain the points of attachment for segregation on a mitotic or meiotic spindle). Because of their repetitive organization, they show some of the same behavior with regard to evolution as the tandem gene clusters. In addition to the satellite sequences, there are shorter stretches of DNA that show similar behavior, called **minisatellites**. They are useful in showing a high degree of divergence between individual genomes that can be used for mapping purposes.

All of these events that change the constitution of the genome are rare, but they are significant over the course of evolution.

This content is available online at http://www.ergito.com/main.jsp?bcs=MBIO.1.4.1

CLUSTERS AND REPEATS

1.4.2 Gene duplication is a major force in evolution

Key Concepts

- Duplicated genes may diverge to generate different genes or one copy may become inactive.

Exons behave like modules for building genes that are tried out in the course of evolution in various combinations. At one extreme, an individual exon from one gene may be copied and used in another gene. At the other extreme, an entire gene, including both exons and introns, may be duplicated. In such a case, mutations can accumulate in one copy without attracting the adverse attention of natural selection. This copy may then evolve to a new function; it may become expressed in a different time or place from the first copy, or it may acquire different activities.

Figure 4.2 summarizes our present view of the rates at which these processes occur. There is ~1% probability that a given gene will be included in a duplication in a period of 1 million years. After the gene has duplicated, differences develop as the result of the occurrence of different mutations in each copy. These accumulate at a rate of ~0.1% per million years (see *Molecular Biology 1.4.4 Sequence divergence is the basis for the evolutionary clock*).

Figure 4.2 After a gene has been duplicated, differences may accumulate between the copies. The genes may acquire different functions or one of the copies may become inactive.

The organism is not likely to need to retain two identical copies of the gene. As differences develop between the duplicated genes, one of two types of event is likely to occur.

- Both of the genes become necessary. This can happen either because the differences between them generate proteins with different functions, or because they are expressed specifically in different times or places.
- If this does not happen, one of the genes is likely to be eliminated, because it will by chance gain a deleterious mutation, and there will be no adverse selection to eliminate this copy. Typically this takes ~ 4 million years. In such a situation, it is purely a matter of chance which of the two copies becomes inactive. (This can contribute to incompatibility between different individuals, and ultimately to speciation, if different copies become inactive in different populations.)

Analysis of the human genome sequence shows that ~5% comprises duplications of identifiable segments ranging in length from 10-300 kb (2847). These have arisen relatively recently, that is, there has not been sufficient time for divergence between them to eliminate their relationship. They include a proportional share (~6%) of the expressed exons, which shows that the duplications are occurring more or less irrespective of genetic content. The genes in these duplications may be especially interesting because of the implication that they have evolved recently, and therefore could be important for recent evolutionary developments (such as the separation of Man from the monkeys).

Last updated on 8-15-2002

References

2847. Bailey, J. A., Gu, Z., Clark, R. A., Reinert, K., Samonte, R. V., Schwartz, S., Adams, M. D., Myers, E. W., Li, P. W., and Eichler, E. E. (2002). *Recent segmental duplications in the human genome.* Science 297, 1003-1007.

CLUSTERS AND REPEATS

1.4.3 Globin clusters are formed by duplication and divergence

Key Terms

Nonallelic genes are two (or more) copies of the same gene that are present at *different* locations in the genome (contrasted with alleles which are copies of the same gene derived from different parents and present at the same location on the homologous chromosomes).

Pseudogenes are inactive but stable components of the genome derived by mutation of an ancestral active gene. Usually they are inactive because of mutations that block transcription or translation or both.

Key Concepts

- All globin genes are descended by duplication and mutation from an ancestral gene that had three exons.
- The ancestral gene gave rise to myoglobin, leghemoglobin, and α- and β-globins.
- The α- and β-globin genes separated in the period of early vertebrate evolution, after which duplications generated the individual clusters of separate α-like and β-like genes.
- Once a gene has been inactivated by mutation, it may accumulate further mutations and become a pseudogene, which is homologous to the active gene(s) but has no functional role.

The most common type of duplication generates a second copy of the gene close to the first copy. In some cases, the copies remain associated, and further duplication may generate a cluster of related genes. The best characterized example of a gene cluster is presented by the globin genes, which constitute an ancient gene family, concerned with a function that is central to the animal kingdom: the transport of oxygen through the bloodstream.

The major constituent of the red blood cell is the globin tetramer, associated with its heme (iron-binding) group in the form of hemoglobin. Functional globin genes in all species have the same general structure, divided into three exons as shown previously in **Figure 2.7**. We conclude that all globin genes are derived from a single ancestral gene; so by tracing the development of individual globin genes within and between species, we may learn about the mechanisms involved in the evolution of gene families.

In adult cells, the globin tetramer consists of two identical α chains and two identical β chains. Embryonic blood cells contain hemoglobin tetramers that are different from the adult form. Each tetramer contains two identical α-like chains and two identical β-like chains, each of which is related to the adult polypeptide and is later replaced

by it. This is an example of developmental control, in which different genes are successively switched on and off to provide alternative products that fulfill the same function at different times.

The division of globin chains into α-like and β-like reflects the organization of the genes. Each type of globin is coded by genes organized into a single cluster. The structures of the two clusters in the higher primate genome are illustrated in **Figure 4.3**.

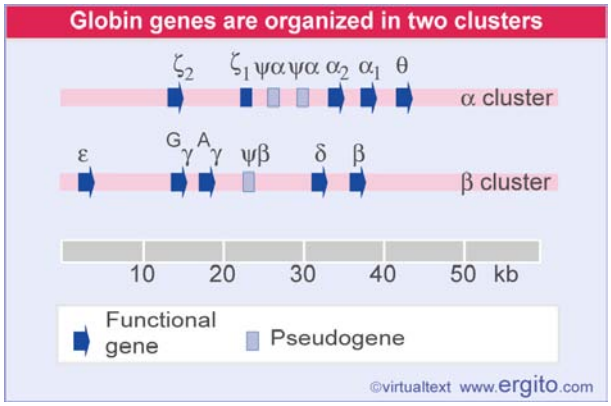


Figure 4.3 Each of the α -like and β -like globin gene families is organized into a single cluster that includes functional genes and pseudogenes (ψ).

Stretching over 50 kb, the β cluster contains five functional genes (ε , two γ , δ , and β) and one nonfunctional gene (ψ β). The two γ genes differ in their coding sequence in only one amino acid; the G variant has glycine at position 136, where the A variant has alanine.

The more compact α cluster extends over 28 kb and includes one active ζ gene, one ζ nonfunctional gene, two α genes, two α nonfunctional genes, and the θ gene of unknown function. The two α genes code for the same protein. Two (or more) identical genes present on the same chromosome are described as **nonallelic** copies.

The details of the relationship between embryonic and adult hemoglobins vary with the organism. The human pathway has three stages: embryonic, fetal, and adult. The distinction between embryonic and adult is common to mammals, but the number of pre-adult stages varies. In Man, zeta and alpha are the two α-like chains. Epsilon, gamma, delta, and beta are the β-like chains. **Figure 4.4** shows how yhe chains are expressed at different stages of development.

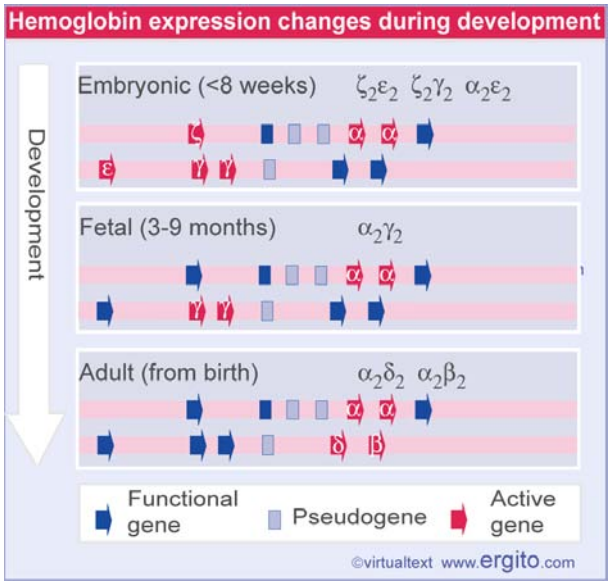


Figure 4.4 Different hemoglobin genes are expressed during embyonic, fetal, and adult periods of human development.

In the human pathway, ζ is the first α-like chain to be expressed, but is soon replaced by α. In the β-pathway, ε and γ are expressed first, with δ and β replacing them later. In adults, the $\alpha_2\ \beta_2$ form provides 97% of the hemoglobin, $\alpha_2\ \delta_2$ is ~2%, and ~1% is provided by persistence of the fetal form $\alpha_2\ \gamma_2$.

What is the significance of the differences between embryonic and adult globins? The embryonic and fetal forms have a higher affinity for oxygen. This is necessary in order to obtain oxygen from the mother's blood. This explains why there is no equivalent in (for example) chicken, where the embryonic stages occur outside the body (that is, within the egg).

Functional genes are defined by their expression in RNA, and ultimately by the proteins for which they code. Nonfunctional genes are defined as such by their inability to code for proteins; the reasons for inactivity vary, and the deficiencies may be in transcription or translation (or both). They are called **pseudogenes** and given the symbol ψ.

A similar general organization is found in other vertebrate globin gene clusters, but details of the types, numbers, and order of genes all vary, as illustrated in **Figure 4.5**. Each cluster contains both embryonic and adult genes. The total lengths of the clusters vary widely. The longest is found in the goat, where a basic cluster of 4 genes has been duplicated twice. The distribution of active genes and pseudogenes differs in each case, illustrating the random nature of the conversion of one copy of a duplicated gene into the inactive state.

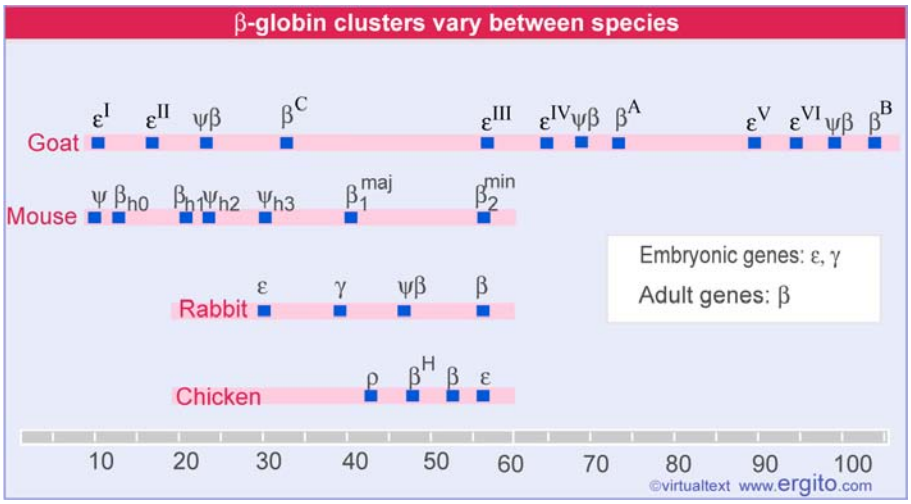


Figure 4.5 Clusters of β -globin genes and pseudogenes are found in vertebrates. Seven mouse genes include 2 early embryonic, 1 late embryonic, 2 adult genes, and 2 pseudogenes. Rabbit and chick each have four genes.

The characterization of these gene clusters makes an important general point. *There may be more members of a gene family, both functional and nonfunctional, than we would suspect on the basis of protein analysis.* The extra functional genes may represent duplicates that code for identical polypeptides; or they may be related to known proteins, although different from them (and presumably expressed only briefly or in low amounts).

With regard to the question of how much DNA is needed to code for a particular function, we see that coding for the β-like globins requires a range of 20-120 kb in different mammals. This is much greater than we would expect just from scrutinizing the known β-globin proteins or even considering the individual genes. However, clusters of this type are not common; most genes are found as individual loci.

From the organization of globin genes in a variety of species, we should be able to trace the evolution of present globin gene clusters from a single ancestral globin gene. Our present view of the evolutionary descent is pictured in **Figure 4.6** (for review see 3041).

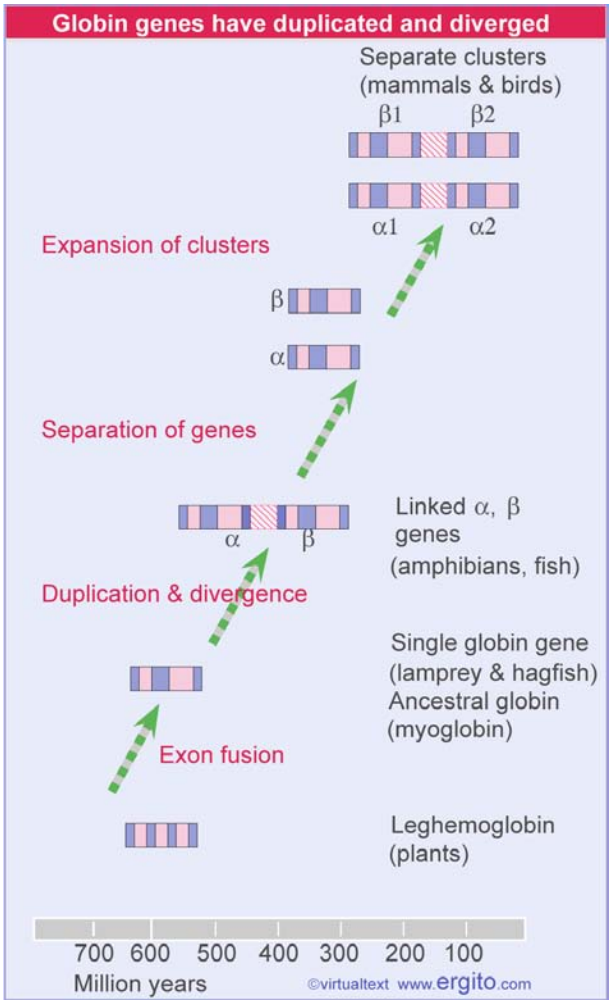


Figure 4.6 All globin genes have evolved by a series of duplications, transpositions, and mutations from a single ancestral gene.

The leghemoglobin gene of plants, which is related to the globin genes, may represent the ancestral form. The furthest back that we can trace a globin gene in modern form is provided by the sequence of the single chain of mammalian myoglobin, which diverged from the globin line of descent ~800 million years ago. The myoglobin gene has the same organization as globin genes, so we may take the three-exon structure to represent their common ancestor.

Some "primitive fish" have only a single type of globin chain, so they must have diverged from the line of evolution before the ancestral globin gene was duplicated to give rise to the α and β variants. This appears to have occurred ~500 million years ago, during the evolution of the bony fish.

The next stage of evolution is represented by the state of the globin genes in the frog *X. laevis*, which has two globin clusters. However, each cluster contains *both* α and β genes, of both larval and adult types. The cluster must therefore have evolved by duplication of a linked α- β pair, followed by divergence between the individual copies. Later the entire cluster was duplicated.

The amphibians separated from the mammalian/avian line ~350 million years ago, so the separation of the α- and β-globin genes must have resulted from a transposition in the mammalian/avian forerunner after this time. This probably occurred in the period of early vertebrate evolution. Since there are separate clusters for α and β globins in both birds and mammals, the α and β genes must have been physically separated before the mammals and birds diverged from their common ancestor, an event that occurred probably ~270 million years ago.

Changes have occurred within the separate α and β clusters in more recent times, as we see from the description of the divergence of the individual genes in *Molecular Biology 1.4.4 Sequence divergence is the basis for the evolutionary clock.*

Last updated on 7-18-2002

Reviews

3041. Hardison, R. (1998). *Hemoglobins from bacteria to man: evolution of different patterns of gene expression.* J. Exp. Biol. 201, 1099-1117.

CLUSTERS AND REPEATS

1.4.4 Sequence divergence is the basis for the evolutionary clock

Key Terms

A **neutral** mutation has no significant effect on evolutionary fitness and usually has no effect on the phenotype.

Random drift describes the chance fluctuation (without selective pressure) of the levels of two alleles in a population.

Fixation is the process by which a new allele replaces the allele that was previously predominant in a population.

Divergence is the percent difference in nucleotide sequence between two related DNA sequences or in amino acid sequences between two proteins.

Replacement sites in a gene are those at which mutations alter the amino acid that is coded.

A **silent site** in a coding region is one where mutation does not change the sequence of the protein.

The **evolutionary clock** is defined by the rate at which mutations accumulate in a given gene.

Key Concepts

- The sequences of homologous genes in different species vary at replacement sites (where mutation causes amino acid substitutions) and silent sites (where mutation does not affect the protein sequence).
- Mutations accumulate at silent sites ~10× faster than at replacement sites.
- The evolutionary divergence between two proteins is measured by the per cent of positions at which the corresponding amino acids are different.
- Mutations accumulate at a more or less even speed after genes separate, so that the divergence between any pair of globin sequences is proportional to the time since their genes separated.

Most changes in protein sequences occur by small mutations that accumulate slowly with time. Point mutations and small insertions and deletions occur by chance, probably with more or less equal probability in all regions of the genome, except for hotspots at which mutations occur much more frequently. Most mutations that change the amino acid sequence are deleterious and will be eliminated by natural selection.

Few mutations are advantageous, but when a rare one occurs, it is likely to spread through the population, eventually replacing the former sequence. When a new variant replaces the previous version of the gene, it is said to have become *fixed* in

the population.

A contentious issue is what proportion of mutational changes in an amino acid sequence are **neutral**, that is, without any effect on the function of the protein, and able therefore to accrue as the result of **random drift** and **fixation**.

The rate at which mutational changes accumulate is a characteristic of each protein, presumably depending at least in part on its flexibility with regard to change. Within a species, a protein evolves by mutational substitution, followed by elimination or fixation within the single breeding pool. Remember that when we scrutinize the gene pool of a species, we see only the variants that have survived. When multiple variants are present, they may be stable (because neither has any selective advantage) or one may in fact be transient because it is in process of being displaced.

When a species separates into two new species, each now constitutes an independent pool for evolution. By comparing the corresponding proteins in two species, we see the differences that have accumulated between them *since the time when their ancestors ceased to interbreed.* Some proteins are highly conserved, showing little or no change from species to species. This indicates that almost any change is deleterious and therefore selected against.

The difference between two proteins is expressed as their **divergence**, the percent of positions at which the amino acids are different. The divergence between proteins can be different from the divergence between the corresponding nucleic acid sequences. The source of this difference is the representation of each amino acid in a three-base codon, in which often the third base has no effect on the meaning.

We may divide the nucleotide sequence of a coding region into potential **replacement sites** and **silent sites**:

- At replacement sites, a mutation alters the amino acid that is coded. The effect of the mutation (deleterious, neutral, or advantageous) depends on the result of the amino acid replacement.
- At silent sites, mutation only substitutes one synonym codon for another, so there is no change in the protein. Usually the replacement sites account for 75% of a coding sequence and the silent sites provide 25%.

In addition to the coding sequence, a gene contains nontranslated regions. Here again, mutations are potentially neutral, apart from their effects on either secondary structure or (usually rather short) regulatory signals.

Although silent mutations are neutral with regard to the protein, they could affect gene expression via the sequence change in RNA. For example, a change in secondary structure might influence transcription, processing, or translation. Another possibility is that a change in synonym codons calls for a different tRNA to respond, influencing the efficiency of translation.

The mutations in replacement sites should correspond with the amino acid divergence (determined by the percent of changes in the protein sequence). A nucleic

acid divergence of 0.45% at replacement sites corresponds to an amino acid divergence of 1% (assuming that the average number of replacement sites per codon is 2.25). Actually, the measured divergence underestimates the differences that have occurred during evolution, because of the occurrence of multiple events at one codon. Usually a correction is made for this.

To take the example of the human β- and δ-globin chains, there are 10 differences in 146 residues, a divergence of 6.9%. The DNA sequence has 31 changes in 441 residues. However, these changes are distributed very differently in the replacement and silent sites. There are 11 changes in the 330 replacement sites, but 20 changes in only 111 silent sites. This gives (corrected) rates of divergence of 3.7% in the replacement sites and 32% in the silent sites, almost an order of magnitude in difference.

The striking difference in the divergence of replacement and silent sites demonstrates the existence of much greater constraints on nucleotide positions that influence protein constitution relative to those that do not. So probably very few of the amino acid changes are neutral.

Suppose we take the rate of mutation at silent sites to indicate the underlying rate of mutational fixation (this assumes that there is no selection at all at the silent sites). Then over the period since the β and δ genes diverged, there should have been changes at 32% of the 330 replacement sites, a total of 105. All but 11 of them have been eliminated, which means that ~90% of the mutations did not survive.

The divergence between any pair of globin sequences is (more or less) proportional to the time since they separated. This provides an **evolutionary clock** that measures the accumulation of mutations at an apparently even rate during the evolution of a given protein.

The rate of divergence can be measured as the percent difference per million years, or as its reciprocal, the unit evolutionary period (UEP), the time in millions of years that it takes for 1% divergence to develop. Once the clock has been established by pairwise comparisons between species (remembering the practical difficulties in establishing the actual time of speciation), it can be applied to related genes *within* a species. From their divergence, we can calculate how much time has passed since the duplication that generated them.

By comparing the sequences of homologous genes in different species, the rate of divergence at both replacement and silent sites can be determined, as plotted in **Figure 4.7**.

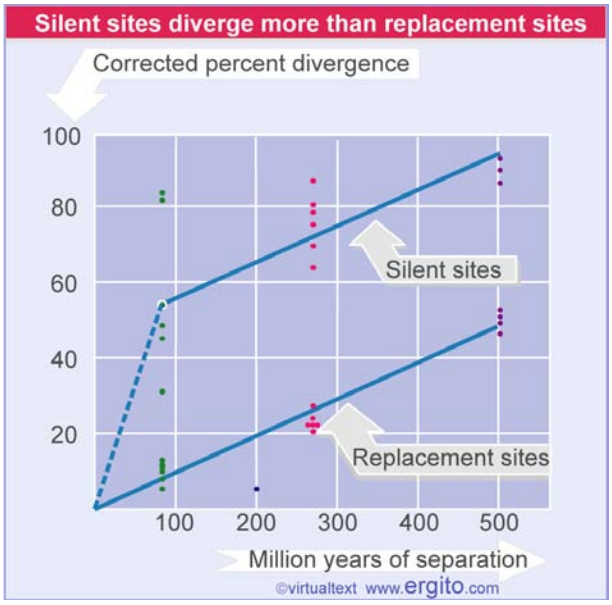


Figure 4.7 Divergence of DNA sequences depends on evolutionary separation. Each point on the graph represents a pairwise comparison.

In pairwise comparisons, there is an average divergence of 10% in the replacement sites of either the α- or β-globin genes of mammals that have been separated since the mammalian radiation occurred ~85 million years ago. This corresponds to a replacement divergence rate of 0.12% per million years.

The rate is steady when the comparison is extended to genes that diverged in the more distant past. For example, the average replacement divergence between corresponding mammalian and chicken globin genes is 23%. Relative to a separation ~270 million years ago, this gives a rate of 0.09% per million years.

Going further back, we can compare the α- with the β-globin genes within a species. They have been diverging since the individual gene types separated ≥ 500 million years ago (see **Figure 4.6**). They have an average replacement divergence of ~50%, which gives a rate of 0.1% per million years.

The summary of these data in **Figure 4.7** shows that replacement divergence in the globin genes has an average rate of ~0.096% per million years (or a UEP of 10.4). Considering the uncertainties in estimating the times at which the species diverged, the results lend good support to the idea that there is a linear clock.

The data on silent site divergence are much less clear. In every case, it is evident that the silent site divergence is much greater than the replacement site divergence, by a factor that varies from 2 to 10. But the spread of silent site divergences in pairwise comparisons is too great to show whether a clock is applicable (so we must base temporal comparisons on the replacement sites).

From **Figure 4.7**, it is clear that the rate at silent sites is not linear with regard to time. *If we assume that there must be zero divergence at zero years of separation,* we

see that the rate of silent site divergence is much greater for the first ~100 million years of separation. One interpretation is that a fraction of roughly half of the silent sites is rapidly (within 100 million years) saturated by mutations; this fraction behaves as neutral sites. The other fraction accumulates mutations more slowly, at a rate approximately the same as that of the replacement sites; this fraction identifies sites that are silent with regard to the protein, but that come under selective pressure for some other reason.

Now we can reverse the calculation of divergence rates to estimate the times since genes within a species have been apart. The difference between the human β and δ genes is 3.7% for replacement sites. At a UEP of 10.4, these genes must have diverged 10.4 × 3.7 = 40 million years ago – about the time of the separation of the lines leading to New World monkeys, Old World monkeys, great apes, and man. All of these higher primates have both β and δ genes, which suggests that the gene divergence commenced just before this point in evolution.

Proceeding further back, the divergence between the replacement sites of γ and ε genes is 10%, which corresponds to a time of separation ~100 million years ago. The separation between embryonic and fetal globin genes therefore may have just preceded or accompanied the mammalian radiation.

An evolutionary tree for the human globin genes is constructed in **Figure 4.8**. Features that evolved before the mammalian radiation – such as the separation of β/δ from γ – should be found in all mammals. Features that evolved afterward – such as the separation of β- and δ-globin genes – should be found in individual lines of mammals.

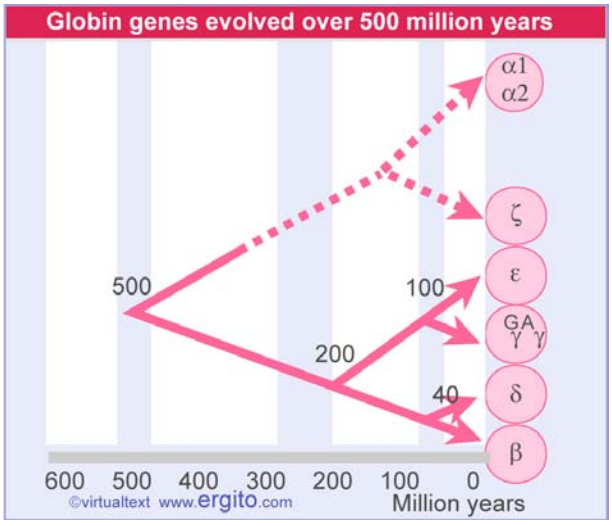


Figure 4.8 Replacement site divergences between pairs of β-globin genes allow the history of the human cluster to be reconstructed. This tree accounts for the separation of classes of globin genes.

In each species, there have been comparatively recent changes in the structures of the clusters, since we see differences in gene number (one adult β-globin gene in man, two in mouse) or in type (most often concerning whether there are separate

embryonic and fetal genes).

When sufficient data have been collected on the sequences of a particular gene, the arguments can be reversed, and comparisons between genes in different species can be used to assess taxonomic relationships.

CLUSTERS AND REPEATS

1.4.5 The rate of neutral substitution can be measured from divergence of repeated sequences

Key Concepts

- The rate of substitution per year at neutral sites is greater in the mouse than in the human genome.

We can make the best estimate of the rate of substitution at neutral sites by examining sequences that do not code for protein. (We use the term neutral here rather than silent, because there is no coding potential). An informative comparison can be made by comparing the members of common repetitive family in the human and mouse genomes (3203).

The principle of the analysis is summarized in **Figure 4.9**. We start with a family of related sequences that have evolved by duplication and substitution from an original family member. We assume that the common ancestral sequence can be deduced by taking the base that is most common at each position. Then we can calculate the divergence of each individual family member as the proportion of bases that differ from the deduced ancestral sequence. In this example, individual members vary from 0.13 - 0.18 divergence, and the average is 0.16.

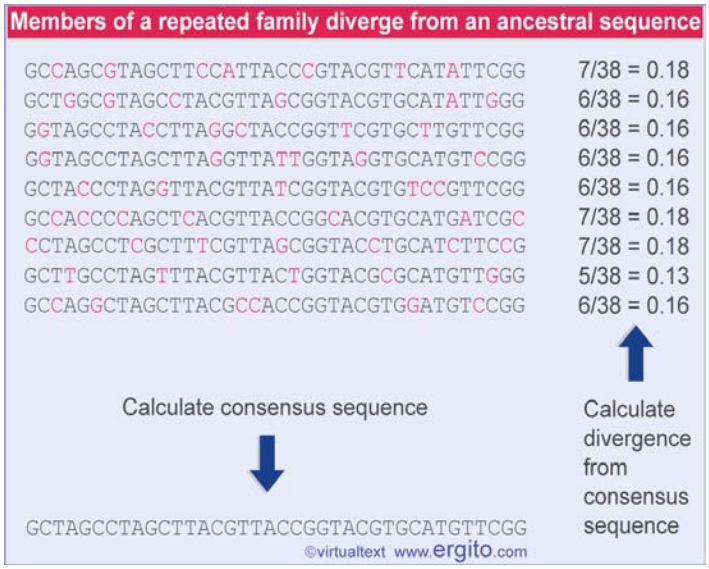


Figure 4.9 An ancestral consensus sequence for a family is calculated by taking the most common base at each position. The divergence of each existing current member of the family is calculated as the proportion of bases at which it differs from the ancestral sequence.

One family used for this analysis in the human and mouse genomes derives from a sequence that is thought to have ceased to be active at about the time of the

divergence between Man and rodents (the LINES family; see *Molecular Biology 4.17.9 Retroposons fall into three classes*). This means that it has been diverging without any selective pressure for the same length of time in both species. Its average divergence in Man is ~0.17 substitutions per site, corresponding to a rate of 2.2×10^{-9} substitutions per base per year over the 75 million years since the separation. In the mouse genome, however, neutral substitutions have occurred at twice this rate, corresponding to 0.34 substitutions per site in the family, or a rate of 4.5×10^{-9} .However, note that if we calculated the rate per generation instead of per year, it would be greater in Man than in mouse ($\sim 2.2 \times 10^{-8}$ as opposed to $\sim 10^{-9}$).

These figures probably underestimate the rate of substitution in the mouse, because at the time of divergence the rates in both species would have been the same, and the difference must have evolved since then. The current rate of neutral substitution per year in the mouse is probably 2-3× greater than the historical average. These rates reflect the balance between the occurrence of mutations and the ability of the genetic system of the organism to correct them. The difference between the species demonstrates that each species has systems that operate with a characteristic efficiency.

Comparing the mouse and human genomes allows us to assess whether syntenic (corresponding) sequences show signs of conservation or have differed at the rate expected from accumulation of neutral substitutions. The proportion of sites that show signs of selection is ~5%. This is much higher than the proportion that codes for protein or RNA (~1%). It implies that the genome includes many more stretches whose sequence is important for non-coding functions than for coding functions. Known regulatory elements are likely to comprise only a small part of this proportion. This number also suggests that most (i.e., the rest) of the genome sequences do not have any function that depends on the exact sequence.

Last updated on 12-20-2002

References

3203. Waterston et al. (2002). *Initial sequencing and comparative analysis of the mouse genome.* Nature 420, 520-562.

CLUSTERS AND REPEATS

1.4.6 Pseudogenes are dead ends of evolution

Key Concepts

- Pseudogenes have no coding function, but they can be recognized by sequence similarities with existing functional genes. They arise by the accumulation of mutations in (formerly) functional genes.

Pseudogenes (Ψ) are defined by their possession of sequences that are related to those of the functional genes, but that cannot be translated into a functional protein.

Some pseudogenes have the same general structure as functional genes, with sequences corresponding to exons and introns in the usual locations. They may have been rendered inactive by mutations that prevent any or all of the stages of gene expression. The changes can take the form of abolishing the signals for initiating transcription, preventing splicing at the exon-intron junctions, or prematurely terminating translation.

Usually a pseudogene has several deleterious mutations. Presumably once it ceased to be active, there was no impediment to the accumulation of further mutations. Pseudogenes that represent inactive versions of currently active genes have been found in many systems, including globin, immunoglobulins, and histocompatibility antigens, where they are located in the vicinity of the gene cluster, often interspersed with the active genes.

A typical example is the rabbit pseudogene, Ψ β 2, which has the usual organization of exons and introns, and is related most closely to the functional globin gene β 1. But it is not functional. **Figure 4.10** summarizes the many changes that have occurred in the pseudogene. The deletion of a base pair at codon 20 of Ψ β 2 has caused a frameshift that would lead to termination shortly after. Several point mutations have changed later codons representing amino acids that are highly conserved in the β globins. Neither of the two introns any longer possesses recognizable boundaries with the exons, so probably the introns could not be spliced out even if the gene were transcribed. However, there are no transcripts corresponding to the gene, possibly because there have been changes in the 5′ flanking region.

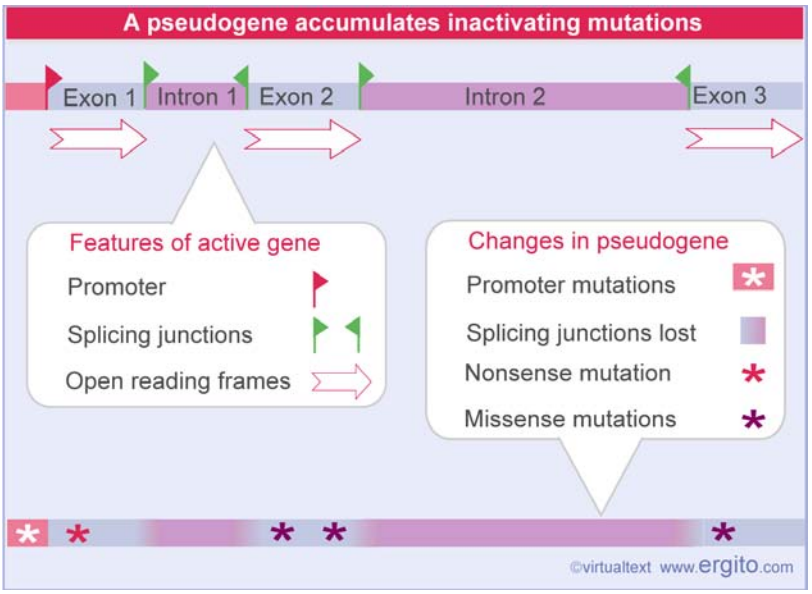


Figure 4.10 Many changes have occurred in a β globin gene since it became a pseudogene.

Since this list of defects includes mutations potentially preventing each stage of gene expression, we have no means of telling which event originally inactivated this gene. However, from the divergence between the pseudogene and the functional gene, we can estimate when the pseudogene originated and when its mutations started to accumulate.

If the pseudogene had become inactive as soon as it was generated by duplication from β 1, we should expect both replacement site and silent site divergence rates to be the same. (They will be different only if the gene is translated to create selective pressure on the replacement sites.) But actually there are fewer replacement site substitutions than silent site substitutions. This suggests that at first (while the gene was expressed) there was selection against replacement site substitution. From the relative extents of substitution in the two types of site, we can calculate that Ψ β 2 diverged from β 1 ~55 million years ago, remained a functional gene for 22 million years, but has been a pseudogene for the last 33 million years.

Similar calculations can be made for other pseudogenes. Some appear to have been active for some time before becoming pseudogenes, but others appear to have been inactive from the very time of their original generation. The general point made by the structures of these pseudogenes is that each has evolved independently during the development of the globin gene cluster in each species. This reinforces the conclusion that the creation of new genes, followed by their acceptance as functional duplicates, variation to become new functional genes, or inactivation as pseudogenes, is a continuing process in the gene cluster. Most gene families have members that are pseudogenes. Usually the pseudogenes represent a small minority of the total gene number.

The mouse Ψ α 3 globin gene has an interesting property: it precisely lacks both introns. Its sequence can be aligned (allowing for accumulated mutations) with the α -globin mRNA. The apparent time of inactivation coincides with the original

duplication, which suggests that the original inactivating event was associated with the loss of introns.

Inactive genomic sequences that resemble the RNA transcript are called processed pseudogenes. They originate by insertion at some random site of a product derived from the RNA, following a retrotransposition event, as discussed in Retroviruses and retroposons . Their characteristic features are summarized in **Figure 17.19**.

If pseudogenes are evolutionary dead ends, simply an unwanted accompaniment to the rearrangement of functional genes, why are they still present in the genome? Do they fulfill any function or are they entirely without purpose, in which case there should be no selective pressure for their retention?

We should remember that we see those genes that have survived in present populations. In past times, any number of other pseudogenes may have been eliminated. This elimination could occur by deletion of the sequence as a sudden event or by the accretion of mutations to the point where the pseudogene can no longer be recognized as a member of its original sequence family (probably the ultimate fate of any pseudogene that is not suddenly eliminated).

Even relics of evolution can be duplicated. In the β -globin genes of the goat, there are three adult species, β A , β B , and β C (see **Figure 4.5**). Each of these has a pseudogene a few kb upstream of it. The pseudogenes are better related to each other than to the adult β -globin genes; in particular, they share several inactivating mutations. Also, the adult β -globin genes are better related to each other than to the pseudogenes. This implies that an original Ψ β - β structure was itself duplicated, giving functional β genes (which diverged further) and two nonfunctional genes (which diverged into the current pseudogenes).

The mechanisms responsible for gene duplication, deletion, and rearrangement act on all sequences that are recognized as members of the cluster, whether or not they are functional. It is left to selection to discriminate among the products.

By definition, pseudogenes do not code for proteins, and usually they have no function at all, but in at least one exceptional case, a pseudogene has a regulatory function. Transcription of a pseudogene inhibits degradation of the mRNA produced by its homologous active gene (3996). Probably there is a protein responsible for this degradation that binds a specific sequence in the mRNA. If this sequence is also present in the RNA transcribed from the pseudogene, the effect of the protein will be diluted when the pseudogene is transcribed. It is not clear how common such effects may be, but as a general rule, we might expect dilution effects of this type to be possible whenever pseudogenes are transcribed.

Last updated on 7-16-2003

References

3996. Hirotsune, S., Yoshida, N., Chen, A., Garrett, L., Sugiyama, F., Takahashi, S., Yagami, K., Wynshaw-Boris, A., and Yoshiki, A. (2003). *An expressed pseudogene regulates the messenger-RNA stability of its homologous coding gene.* Nature 423, 91-96.

CLUSTERS AND REPEATS

1.4.7 Unequal crossing-over rearranges gene clusters

Key Terms

Nonreciprocal recombination (unequal crossing-over) results from an error in pairing and crossing-over in which nonequivalent sites are involved in a recombination event. It produces one recombinant with a deletion of material and one with a duplication.

Thalassemia is disease of red blood cells resulting from lack of either α or β globin.

HbH disease results from a condition in which there is a disproportionate amount of the abnormal tetramer β_4 relative to the amount of normal hemoglobin (β_2).

Hydrops fetalis is a fatal disease resulting from the absence of the hemoglobin α gene.

Hb Lepore is an unusual globin protein that results from unequal crossing-over between the β and δ genes. The genes become fused together to produce a single β -like chain that consists of the N-terminal sequence of δ joined to the C-terminal sequence of β .

Hb anti-Lepore is a fusion gene produced by unequal crossing-over that has the N-terminal part of β globin and the C-terminal part of δ globin.

Hb Kenya is a fusion gene produced by unequal crossing-over between the between $^{A}\gamma$ and β globin genes.

Key Concepts

- When a genome contains a cluster of genes with related sequences, mispairing between nonallelic genes can cause unequal crossing-over. This produces a deletion in one recombinant chromosome and a corresponding duplication in the other.

- Different thalassemias are caused by various deletions that eliminate α- or β-globin genes. The severity of the disease depends on the individual deletion.

There are frequent opportunities for rearrangement in a cluster of related or identical genes. We can see the results by comparing the mammalian β clusters included in **Figure 4.5**. Although the clusters serve the same function, and all have the same general organization, each is different in size, there is variation in the total number and types of β-globin genes, and the numbers and structures of pseudogenes are different. All of these changes must have occurred since the mammalian radiation, ~85 million years ago (the last point in evolution common to all the mammals).

The comparison makes the general point that gene duplication, rearrangement, and variation is as important a factor in evolution as the slow accumulation of point mutations in individual genes. What types of mechanisms are responsible for gene reorganization?

Unequal crossing-over (also known as **nonreciprocal recombination**) can occur as the result of pairing between two sites that are *not* homologous. Usually, recombination involves corresponding sequences of DNA held in exact alignment between the two homologous chromosomes. However, when there are two copies of a gene on each chromosome, an occasional misalignment allows pairing between them. (This requires some of the adjacent regions to go unpaired.) This can happen in a region of short repeats (see **Figure 4.1**) or in a gene cluster. **Figure 4.11** shows that unequal crossing-over in a gene cluster can have two consequences, quantitative and qualitative:

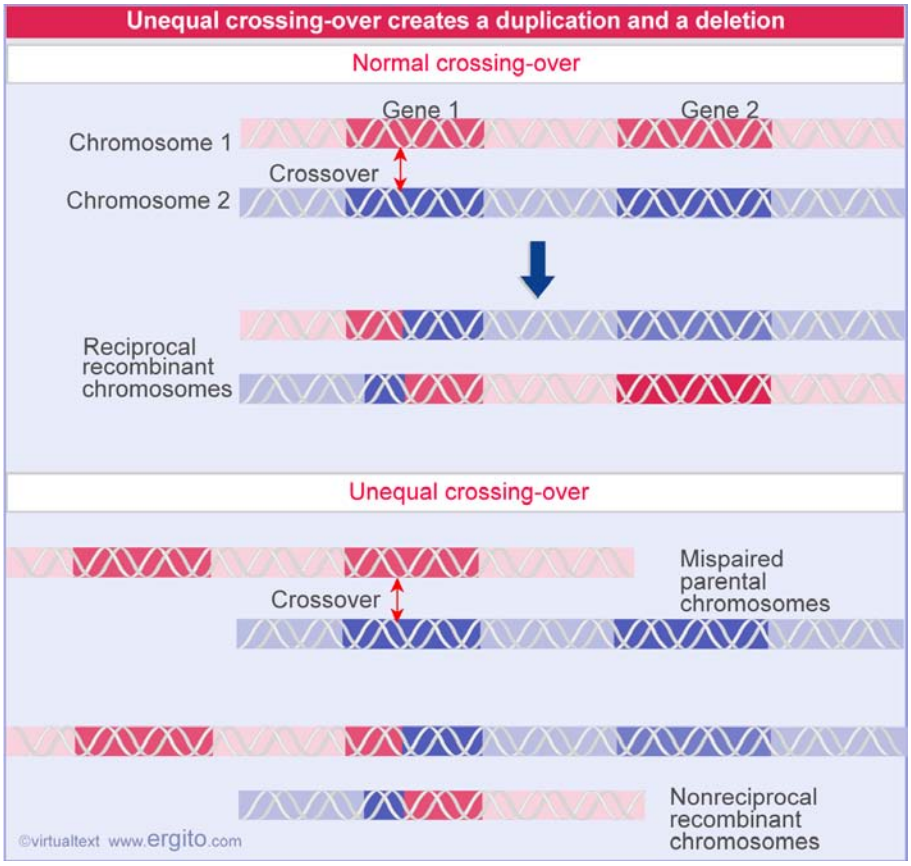


Figure 4.11 Gene number can be changed by unequal crossing-over. If gene 1 of one chromosome pairs with gene 2 of the other chromosome, the other gene copies are excluded from pairing. Recombination between the mispaired genes produces one chromosome with a single (recombinant) copy of the gene and one chromosome with three copies of the gene (one from each parent and one recombinant).

- The number of repeats increases in one chromosome and decreases in the other. In effect, one recombinant chromosome has a deletion and the other has an insertion. This happens irrespective of the exact location of the crossover. In the figure, the first recombinant has an increase in the number of gene copies from 2 to 3, while the second has a decrease from 2 to 1.

- If the recombination event occurs within a gene (as opposed to between genes), the result depends on whether the recombining genes are identical or only related. If the noncorresponding gene copies 1 and 2 are entirely homologous,

there is no change in the sequence of either gene. However, unequal crossing-over also can occur when the adjacent genes are well related (although the probability is less than when they are identical). In this case, each of the recombinant genes has a sequence that is different from either parent.

Whether the chromosome has a selective advantage or disadvantage will depend on the consequence of any change in the sequence of the gene product as well as on the change in the number of gene copies.

An obstacle to unequal crossing-over is presented by the interrupted structure of the genes. In a case such as the globins, the corresponding exons of adjacent gene copies are likely to be well enough related to support pairing; but the sequences of the introns have diverged appreciably. The restriction of pairing to the exons considerably reduces the continuous length of DNA that can be involved. This lowers the chance of unequal crossing-over. So divergence between introns could enhance the stability of gene clusters by hindering the occurrence of unequal crossing-over.

Thalassemias result from mutations that reduce or prevent synthesis of either α or β globin. The occurrence of unequal crossing-over in the human globin gene clusters is revealed by the nature of certain thalassemias.

Many of the most severe thalassemias result from deletions of part of a cluster. In at least some cases, the ends of the deletion lie in regions that are homologous, which is exactly what would be expected if it had been generated by unequal crossing-over.

Figure 4.12 summarizes the deletions that cause the α-thalassemias. α-thal-1 deletions are long, varying in the location of the left end, with the positions of the right ends located beyond the known genes. They eliminate both the α genes. The α-thal-2 deletions are short and eliminate only one of the two α genes. The L deletion removes 4.2 kb of DNA, including the α2 gene. It probably results from unequal crossing-over, because the ends of the deletion lie in homologous regions, just to the right of the ψ α and α2 genes, respectively. The R deletion results from the removal of exactly 3.7 kb of DNA, the precise distance between the α1 and α2 genes. It appears to have been generated by unequal crossing-over between the α1 and α2 genes themselves. This is precisely the situation depicted in **Figure 4.11**.

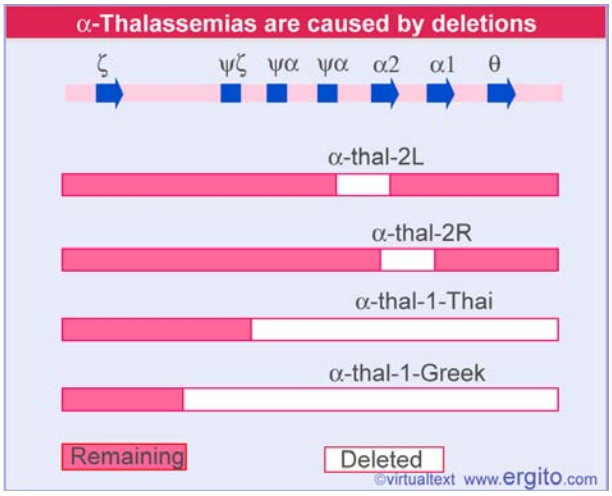


Figure 4.12 Thalassemias result from various deletions in the α -globin gene cluster.

Depending on the diploid combination of thalassemic chromosomes, an affected individual may have any number of α chains from zero to three. There are few differences from the wild type (four α genes) in individuals with three or two α genes. But with only one α gene, the excess β chains form the unusual tetramer β_4, which causes **HbH** disease. The complete absence of α genes results in **hydrops fetalis**, which is fatal at or before birth.

The same unequal crossing-over that generated the thalassemic chromosome should also have generated a chromosome with three α genes. Individuals with such chromosomes have been identified in several populations. In some populations, the frequency of the triple α locus is about the same as that of the single α locus; in others, the triple α genes are much *less* common than single αgenes. This suggests that (unknown) selective factors operate in different populations to adjust the gene levels.

Variations in the number of α genes are found relatively frequently, which argues that unequal crossing-over in the cluster must be fairly common. It occurs more often in the α cluster than in the β cluster, possibly because the introns in α genes are much shorter, and therefore present less impediment to mispairing between nonhomologous genes.

The deletions that cause β-thalassemias are summarized in **Figure 4.13**. In some (rare) cases, only the β gene is affected. These have a deletion of 600 bp, extending from the second intron through the 3′ flanking regions. In the other cases, more than one gene of the cluster is affected. Many of the deletions are very long, extending from the 5′ end indicated on the map for >50 kb toward the right.

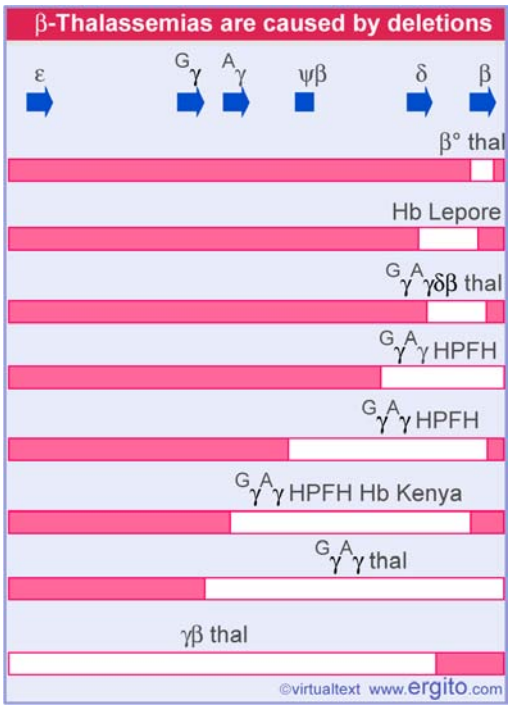


Figure 4.13 Deletions in the β -globin gene cluster cause several types of thalassemia.

The **Hb Lepore** type provided the classic evidence that deletion can result from unequal crossing-over between linked genes. The β and δ genes differ only ~7% in sequence. Unequal recombination deletes the material between the genes, thus fusing them together (see **Figure 4.11**). The fused gene produces a single β-like chain that consists of the N-terminal sequence of δ joined to the C-terminal sequence of β.

Several types of Hb Lepore now are known, the difference between them lying in the point of transition from δ to β sequences. So when the δ and β genes pair for unequal crossing-over, the exact point of recombination determines the position at which the switch from δ to β sequence occurs in the amino acid chain.

The reciprocal of this event has been found in the form of **Hb anti-Lepore**, which is produced by a gene that has the N-terminal part of β and the C-terminal part of δ . The fusion gene lies between normal δ and β genes.

Evidence that unequal crossing-over can occur between more distantly related genes is provided by the identification of **Hb Kenya**, another fused hemoglobin. This contains the N-terminal sequence of the $^{A}\gamma$ gene and the C-terminal sequence of the β gene. The fusion must have resulted from unequal crossing-over between $^{A}\gamma$ and β, which differ ~20% in sequence.

From the differences between the globin gene clusters of various mammals, we see that duplication followed (sometimes) by variation has been an important feature in the evolution of each cluster. The human thalassemic deletions demonstrate that unequal crossing-over continues to occur in both globin gene clusters. Each such event generates a duplication as well as the deletion, and we must account for the

fate of both recombinant loci in the population. Deletions can also occur (in principle) by recombination between homologous sequences lying on the *same* chromosome. This does not generate a corresponding duplication.

It is difficult to estimate the natural frequency of these events, because selective forces rapidly adjust the levels of the variant clusters in the population. Generally a contraction in gene number is likely to be deleterious and selected against. However, in some populations, there may be a balancing advantage that maintains the deleted form at a low frequency.

The structures of the present human clusters show several duplications that attest to the importance of such mechanisms. The *functional* sequences include two α genes coding the same protein, fairly well related β and δ genes, and two almost identical γ genes. These comparatively recent independent duplications have survived in the population, not to mention the more distant duplications that originally generated the various types of globin genes. Other duplications may have given rise to pseudogenes or have been lost. We expect continual duplication and deletion to be a feature of all gene clusters.

CLUSTERS AND REPEATS

1.4.8 Genes for rRNA form tandem repeats

Key Terms

Ribosomal DNA (rDNA) is usually a tandemly repeated series of genes coding for a precursor to the two large rRNAs.

The **nucleolus (nucleoli)** is a discrete region of the nucleus where ribosomes are produced.

The **nucleolar organizer** is the region of a chromosome carrying genes coding for rRNA.

The **nontranscribed spacer** is the region between transcription units in a tandem gene cluster.

Key Concepts

- Ribosomal RNA is coded by a large number of identical genes that are tandemly repeated to form one or more clusters.
- Each rDNA cluster is organized so that transcription units giving a joint precursor to the major rRNAs alternate with nontranscribed spacers.

In the cases we have discussed so far, there are differences between the individual members of a gene cluster that allow selective pressure to act independently upon each gene. A contrast is provided by two cases of large gene clusters that contain many identical copies of the same gene or genes. Most organisms contain multiple copies of the genes for the histone proteins that are a major component of the chromosomes; and there are almost always multiple copies of the genes that code for the ribosomal RNAs. These situations pose some interesting evolutionary questions.

Ribosomal RNA is the predominant product of transcription, constituting some 80-90% of the total mass of cellular RNA in both eukaryotes and prokaryotes. The number of major rRNA genes varies from 7 in *E. coli*, 100-200 in lower eukaryotes, to several hundred in higher eukaryotes. The genes for the large and small rRNA (found respectively in the large and small subunits of the ribosome) usually form a tandem pair. (The sole exception is the yeast mitochondrion.)

The lack of any detectable variation in the sequences of the rRNA molecules implies that all the copies of each gene must be identical, or at least must have differences below the level of detection in rRNA (~1%). A point of major interest is what mechanism(s) are used to prevent variations from accruing in the individual sequences.

In bacteria, the multiple rRNA gene pairs are dispersed. In most eukaryotic nuclei, the rRNA genes are contained in a tandem cluster or clusters. Sometimes these regions are called **rDNA**. (In some cases, the proportion of rDNA in the total DNA, together with its atypical base composition, is great enough to allow its isolation as a

separate fraction directly from sheared genomic DNA.) An important diagnostic feature of a tandem cluster is that it generates a circular restriction map, as shown in **Figure 4.14**.

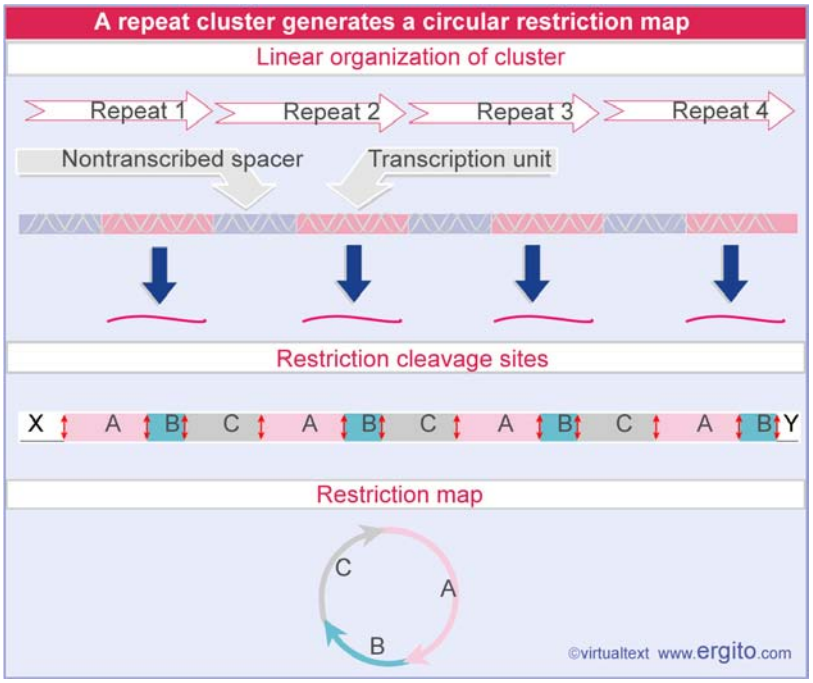


Figure 4.14 A tandem gene cluster has an alternation of transcription unit and nontranscribed spacer and generates a circular restriction map.

Suppose that each repeat unit has 3 restriction sites. In the example shown in the figure, fragments A and B are contained entirely within a repeat unit, and fragment C contains the end of one repeat and the beginning of the next. When we map these fragments by conventional means, we find that A is next to B, which is next to C, which is next to A, generating the circular map. If the cluster is large, the internal fragments (A, B, C) will be present in much greater quantities than the terminal fragments (X, Y) which connect the cluster to adjacent DNA. In a cluster of 100 repeats, X and Y would be present at 1% of the level of A, B, C. This can make it difficult to obtain the ends of a gene cluster for mapping purposes.

The region of the nucleus where rRNA synthesis occurs has a characteristic appearance, with a core of fibrillar nature surrounded by a granular cortex. The fibrillar core is where the rRNA is transcribed from the DNA template; and the granular cortex is formed by the ribonucleoprotein particles into which the rRNA is assembled. The whole area is called the **nucleolus**. Its characteristic morphology is evident in **Figure 4.15**.

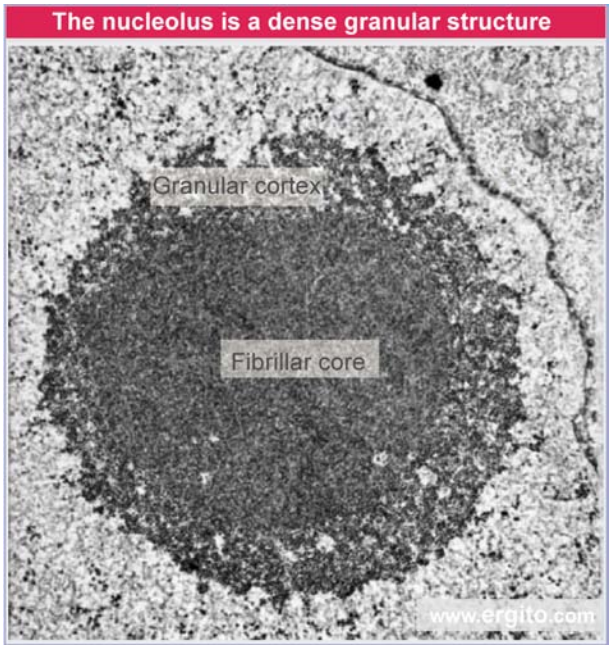


Figure 4.15 The nucleolar core identifies rDNA under transcription, and the surrounding granular cortex consists of assembling ribosomal subunits. This thin section shows the nucleolus of the newt *Notopthalmus viridescens*. Photograph kindly provided by Oscar Miller.

The particular chromosomal regions associated with a nucleolus are called **nucleolar organizers**. Each nucleolar organizer corresponds to a cluster of tandemly repeated rRNA genes on one chromosome. The concentration of the tandemly repeated rRNA genes, together with their very intensive transcription, is responsible for creating the characteristic morphology of the nucleoli.

The pair of major rRNAs is transcribed as a single precursor in both bacteria and eukaryotic nuclei. Following transcription, the precursor is cleaved to release the individual rRNA molecules. The transcription unit is shortest in bacteria and is longest in mammals (where it is known as 45S RNA, according to its rate of sedimentation). An rDNA cluster contains many transcription units, each separated from the next by a **nontranscribed spacer**. The alternation of transcription unit and nontranscribed spacer can be seen directly in electron micrographs. The example shown in **Figure 4.16** is taken from the newt *N. viridescens*, in which each transcription unit is intensively expressed, so that many RNA polymerases are simultaneously engaged in transcription on one repeating unit. The polymerases are so closely packed that the RNA transcripts form a characteristic matrix displaying increasing length along the transcription unit.

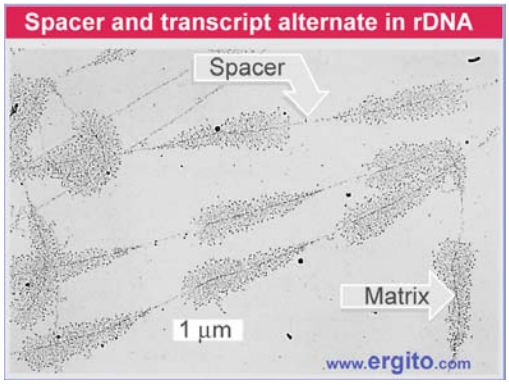


Figure 4.16 Transcription of rDNA clusters generates a series of matrices, each corresponding to one transcription unit and separated from the next by the nontranscribed spacer. Photograph kindly provided by Oscar Miller.

This content is available online at http://www.ergito.com/main.jsp?bcs=MBIO.1.4.8

CLUSTERS AND REPEATS

1.4.9 The repeated genes for rRNA maintain constant sequence

Key Terms

Bam islands are a series of short, repeated sequences found in the nontranscribed spacer of *Xenopus* rDNA genes. The name reflects their isolation by use of the BamI restriction enzyme.

Key Concepts

- The genes in an rDNA cluster all have an identical sequence.
- The nontranscribed spacers consist of shorter repeating units whose number varies so that the lengths of individual spacers are different.

The nontranscribed spacer varies widely in length between and (sometimes) within species. In yeast there is a short nontranscribed spacer, relatively constant in length. In *D. melanogaster,* there is almost a twofold variation in the length of the nontranscribed spacer between different copies of the repeating unit. A similar situation is seen in *X. laevis.* In each of these cases, all of the repeating units are present as a single tandem cluster on one particular chromosome. (In the example of *D. melanogaster,* this happens to be the sex chromosome. The cluster on the X chromosome is larger than that on the Y chromosome, so female flies have more copies of the rRNA genes than male flies.)

In mammals the repeating unit is very much larger, comprising the transcription unit of ~13 kb and a nontranscribed spacer of ~30 kb. Usually, the genes lie in several dispersed clusters – in the case of man and mouse residing on five and six chromosomes, respectively. One interesting (but unanswered) question is how the corrective mechanisms that presumably function within a single cluster to ensure constancy of rRNA sequence are able to work when there are several clusters.

The variation in length of the nontranscribed spacer in a single gene cluster contrasts with the conservation of sequence of the transcription unit. In spite of this variation, the sequences of longer nontranscribed spacers remain homologous with those of the shorter nontranscribed spacers. This implies that each nontranscribed spacer is *internally repetitious,* so that the variation in length results from changes in the number of repeats of some subunit.

The general nature of the nontranscribed spacer is illustrated by the example of *X. laevis.***Figure 4.17** illustrates the situation. Regions that are fixed in length alternate with regions that vary. Each of the three repetitious regions comprises a variable number of repeats of a rather short sequence. One type of repetitious region has repeats of a 97 bp sequence; the other, which occurs in two locations, has a repeating unit found in two forms, 60 bp and 81 bp long. The variation in the number of repeating units in the repetitious regions accounts for the overall variation in spacer

length. The repetitious regions are separated by shorter constant sequences called **Bam islands**. (This description takes its name from their isolation via the use of the BamHI restriction enzyme.) From this type of organization, we see that the cluster has evolved by duplications involving the promoter region.

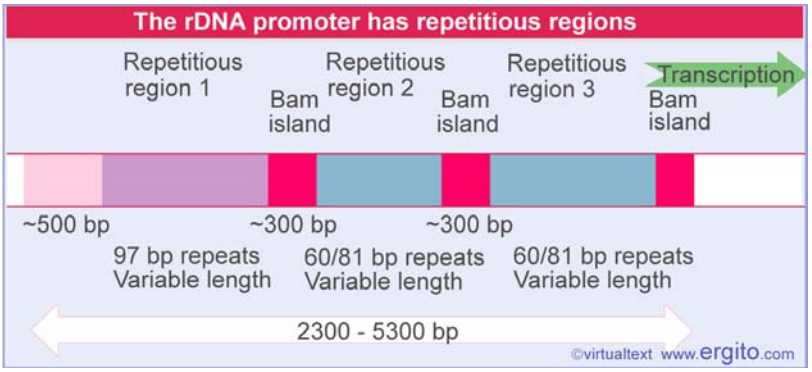


Figure 4.17 The nontranscribed spacer of *X. laevis* rDNA has an internally repetitious structure that is responsible for its variation in length. The Bam islands are short constant sequences that separate the repetitious regions.

We need to explain the lack of variation in the expressed copies of the repeated genes. One model would suppose that there is a quantitative demand for a certain number of "good" sequences. But this would enable mutated sequences to accumulate up to a point at which their proportion of the cluster is great enough for selective pressure to be exerted. We can exclude such models because of the lack of such variation in the cluster.

The lack of variation implies the existence of selective pressure in some form that is sensitive to individual variations. One model would suppose that the entire cluster is regenerated periodically from one or from a very few members. As a practical matter any mechanism would need to involve regeneration every generation. We can exclude such models because a regenerated cluster would not show variation in the nontranscribed regions of the individual repeats.

We are left with a dilemma. Variation in the nontranscribed regions suggests that there is frequent unequal crossing over. This will change the size of the cluster, but will not otherwise change the properties of the individual repeats. So how are mutations prevented from accumulating? We see in *Molecular Biology 1.4.10 Crossover fixation could maintain identical repeats* that continuous contraction and expansion of a cluster may provide a mechanism for homogenizing its copies.

This content is available online at http://www.ergito.com/main.jsp?bcs=MBIO.1.4.9

CLUSTERS AND REPEATS

1.4.10 Crossover fixation could maintain identical repeats

Key Terms

Concerted evolution describes the ability of two related genes to evolve together as though constituting a single locus.

Coincidental evolution (Coevolution) describes a situation in which two genes evolve together as a single unit.

Gene conversion is the alteration of one strand of a heteroduplex DNA to make it complementary with the other strand at any position(s) where there were mispaired bases.

Crossover fixation refers to a possible consequence of unequal crossing-over that allows a mutation in one member of a tandem cluster to spread through the whole cluster (or to be eliminated).

Key Concepts

- Unequal crossing-over changes the size of a cluster of tandem repeats.
- Individual repeating units can be eliminated or can spread through the cluster.

The same problem is encountered whenever a gene has been duplicated. How can selection be imposed to prevent the accumulation of deleterious mutations?

The duplication of a gene is likely to result in an immediate relaxation of the evolutionary pressure on its sequence. Now that there are two identical copies, a change in the sequence of either one will not deprive the organism of a functional protein, since the original amino acid sequence continues to be coded by the other copy. Then the selective pressure on the two genes is diffused, until one of them mutates sufficiently away from its original function to refocus all the selective pressure on the other.

Immediately following a gene duplication, changes might accumulate more rapidly in one of the copies, leading eventually to a new function (or to its disuse in the form of a pseudogene). If a new function develops, the gene then evolves at the same, slower rate characteristic of the original function. Probably this is the sort of mechanism responsible for the separation of functions between embryonic and adult globin genes.

Yet there are instances where duplicated genes retain the same function, coding for the identical or nearly identical proteins. Identical proteins are coded by the two human α-globin genes, and there is only a single amino acid difference between the two γ-globin proteins. How is selective pressure exerted to maintain their sequence identity?

The most obvious possibility is that the two genes do not actually have identical functions, but differ in some (undetected) property, such as time or place of expression. Another possibility is that the need for two copies is quantitative, because neither by itself produces a sufficient amount of protein.

In more extreme cases of repetition, however, it is impossible to avoid the conclusion that no single copy of the gene is essential. When there are many copies of a gene, the immediate effects of mutation in any one copy must be very slight. The consequences of an individual mutation are diluted by the large number of copies of the gene that retain the wild-type sequence. Many mutant copies could accumulate before a lethal effect is generated.

Lethality becomes quantitative, a conclusion reinforced by the observation that half of the units of the rDNA cluster of *X. laevis* or *D. melanogaster* can be deleted without ill effect. So how are these units prevented from gradually accumulating deleterious mutations? And what chance is there for the rare favorable mutation to display its advantages in the cluster?

The basic principle of models to explain the maintenance of identity among repeated copies is to suppose that nonallelic genes are not independently inherited, but must be continually regenerated from *one* of the copies of a preceding generation. In the simplest case of two identical genes, when a mutation occurs in one copy, either it is by chance eliminated (because the sequence of the other copy takes over), or it is spread to both duplicates (because the mutant copy becomes the dominant version). Spreading exposes a mutation to selection. The result is that the two genes evolve together as though only a single locus existed. This is called **coincidental evolution** or **concerted evolution** (occasionally **coevolution**). It can be applied to a pair of identical genes or (with further assumptions) to a cluster containing many genes.

One mechanism supposes that the sequences of the nonallelic genes are directly compared with one another and homogenized by enzymes that recognize any differences. This can be done by exchanging single strands between them, to form genes one of whose strands derives from one copy, one from the other copy. Any differences show as improperly paired bases, which attract attention from enzymes able to excise and replace a base, so that only A·T and G·C pairs survive. This type of event is called **gene conversion** and is associated with genetic recombination as described in *Molecular Biology 4.15 Recombination and repair*.

We should be able to ascertain the scope of such events by comparing the sequences of duplicate genes. If they are subject to concerted evolution, we should not see the accumulation of silent site substitutions between them (because the homogenization process applies to these as well as to the replacement sites). We know that the extent of the maintenance mechanism need not extend beyond the gene itself, since there are cases of duplicate genes whose flanking sequences are entirely different. Indeed, we may see abrupt boundaries that mark the ends of the sequences that were homogenized.

We must remember that the existence of such mechanisms can invalidate the determination of the history of such genes via their divergence, *because the divergence reflects only the time since the last homogenization/regeneration event, not the original duplication*.

The **crossover fixation** model supposes that an entire cluster is subject to continual rearrangement by the mechanism of unequal crossing-over. Such events can explain the concerted evolution of multiple genes if unequal crossing-over causes all the copies to be regenerated physically from one copy.

Following the sort of event depicted in , for example, the chromosome carrying a triple locus could suffer deletion of one of the genes. Of the two remaining genes, 1½ represent the sequence of one of the original copies; only ½ of the sequence of the other original copy has survived. Any mutation in the first region now exists in both genes and is subject to selective pressure.

Tandem clustering provides frequent opportunities for "mispairing" of genes whose sequences are the same, but that lie in different positions in their clusters. By continually expanding and contracting the number of units via unequal crossing-over, it is possible for all the units in one cluster to be derived from rather a small proportion of those in an ancestral cluster. The variable lengths of the spacers are consistent with the idea that unequal crossing-over events take place in spacers that are internally mispaired. This can explain the homogeneity of the genes compared with the variability of the spacers. The genes are exposed to selection when individual repeating units are amplified within the cluster; but the spacers are irrelevant and can accumulate changes.

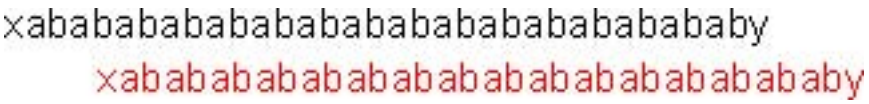
In a region of nonrepetitive DNA, recombination occurs between precisely matching points on the two homologous chromosomes, generating reciprocal recombinants. The basis for this precision is the ability of two duplex DNA sequences to align exactly. We know that unequal recombination can occur when there are multiple copies of genes whose exons are related, even though their flanking and intervening sequences may differ. This happens because of the mispairing between corresponding exons in *nonallelic* genes.

Imagine how much more frequently misalignment must occur in a tandem cluster of identical or nearly identical repeats. Except at the very ends of the cluster, the close relationship between successive repeats makes it impossible even to define the exactly corresponding repeats! This has two consequences: there is continual adjustment of the size of the cluster; and there is homogenization of the repeating unit.

Consider a sequence consisting of a repeating unit "ab" with ends "x" and "y." If we represent one chromosome in black and the other in color, the exact alignment between "allelic" sequences would be:

xababababababababababababababababy
xababababababababababababababababy

But probably *any* sequence *ab* in one chromosome could pair with *any* sequence *ab* in the other chromosome. In a misalignment such as:



the region of pairing is no less stable than in the perfectly aligned pair, although it is shorter. We do not know very much about how pairing is initiated prior to recombination, but very likely it starts between short corresponding regions and then spreads. If it starts within satellite DNA, it is more likely than not to involve repeating units that do not have exactly corresponding locations in their clusters.

Now suppose that a recombination event occurs within the unevenly paired region. The recombinants will have different numbers of repeating units. In one case, the cluster has become longer; in the other, it has become shorter,

where "× " indicates the site of the crossover.

If this type of event is common, clusters of tandem repeats will undergo continual expansion and contraction. This can cause a particular repeating unit to spread through the cluster, as illustrated in **Figure 4.18**. Suppose that the cluster consists initially of a sequence *abcde*, where each letter represents a repeating unit. The different repeating units are closely enough related to one another to mispair for recombination. Then by a series of unequal recombination events, the size of the repetitive region increases or decreases, and also one unit spreads to replace all the others.

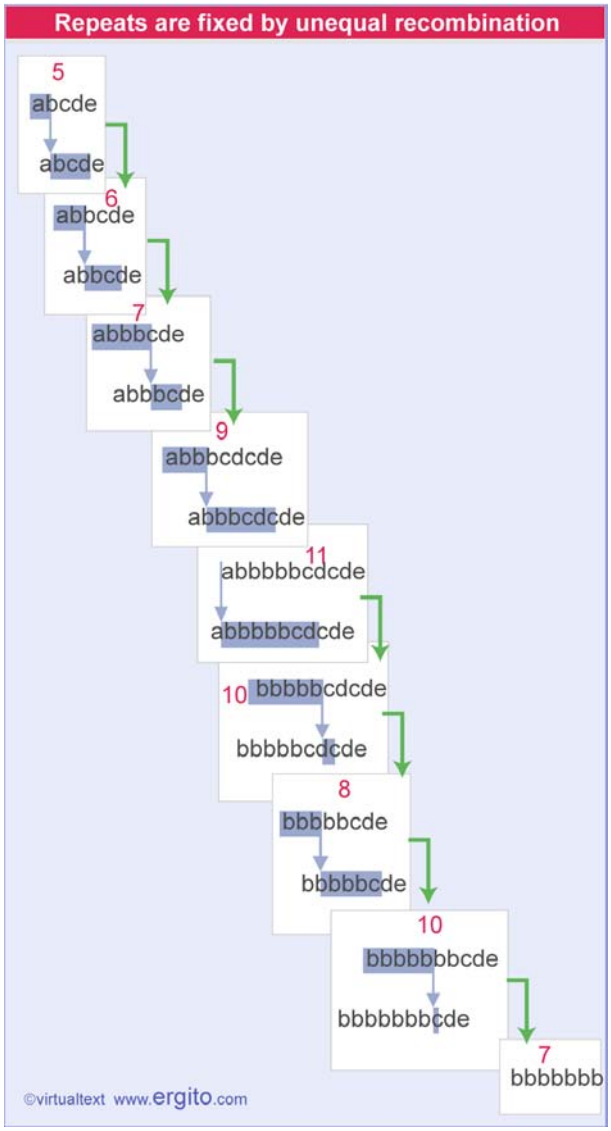


Figure 4.18 Unequal recombination allows one particular repeating unit to occupy the entire cluster. The numbers indicate the length of the repeating unit at each stage.

The crossover fixation model predicts that *any sequence of DNA that is not under selective pressure will be taken over by a series of identical tandem repeats generated in this way* (for review see 3042). The critical assumption is that the process of crossover fixation is fairly rapid relative to mutation, so that new mutations either are eliminated (their repeats are lost) or come to take over the entire cluster. In the case of the rDNA cluster, of course, a further factor is imposed by selection for an effective transcribed sequence.

Reviews

3042. Charlesworth, B., Sniegowski, P., and Stephan, W. (1994). *The evolutionary dynamics of repetitive DNA in eukaryotes.* Nature 371, 215-220.

CLUSTERS AND REPEATS

1.4.11 Satellite DNAs often lie in heterochromatin

Key Terms

Highly repetitive DNA (Simple sequence DNA) is the first component to reassociate and is equated with satellite DNA.

Satellite DNA (Simple-sequence DNA) consists of many tandem repeats (identical or related) of a short basic repeating unit.

A **density gradient** is used to separate macromolecules on the basis of differences in their density. It is prepared from a heavy soluble compound such as CsCl.

A **cryptic satellite** is a satellite DNA sequence not identified as such by a separate peak on a density gradient; that is, it remains present in main-band DNA.

***In situ* hybridization (Cytological hybridization)** is performed by denaturing the DNA of cells squashed on a microscope slide so that reaction is possible with an added single-stranded RNA or DNA; the added preparation is radioactively labeled and its hybridization is followed by autoradiography.

Heterochromatin describes regions of the genome that are highly condensed, are not transcribed, and are late-replicating. Heterochromatin is divided into two types, which are called constitutive and facultative.

Euchromatin comprises all of the genome in the interphase nucleus except for the heterochromatin. The euchromatin is less tightly coiled than heterochromatin, and contains the active or potentially active genes.

Key Concepts

- Highly repetitive DNA has a very short repeating sequence and no coding function.
- It occurs in large blocks that can have distinct physical properties.
- It is often the major constituent of centromeric heterochromatin.

Repetitive DNA is defined by its (relatively) rapid rate of renaturation. The component that renatures most rapidly in a eukaryotic genome is called *highly repetitive* DNA, and consists of very short sequences repeated many times in tandem in large clusters. Because of its short repeating unit, it is sometimes described as **simple sequence DNA**. This type of component is present in almost all higher eukaryotic genomes, but its overall amount is extremely variable. In mammalian genomes it is typically <10%, but in (for example) *Drosophila virilis*, it amounts to ~50%. In addition to the large clusters in which this type of sequence was originally discovered, there are smaller clusters interspersed with nonrepetitive DNA. It typically consists of short sequences that are repeated in identical or related copies in the genome.

The tandem repetition of a short sequence often creates a fraction with distinctive physical properties that can be used to isolate it. In some cases, the repetitive sequence has a base composition distinct from the genome average, which allows it

to form a separate fraction by virtue of its distinct buoyant density. A fraction of this sort is called **satellite DNA**. The term satellite DNA is essentially synonymous with simple sequence DNA. Consistent with its simple sequence, this DNA is not transcribed or translated.

Tandemly repeated sequences are especially liable to undergo misalignments during chromosome pairing, and thus the sizes of tandem clusters tend to be highly polymorphic, with wide variations between individuals. In fact, the smaller clusters of such sequences can be used to characterize individual genomes in the technique of "DNA fingerprinting" (see *Molecular Biology 1.4.14 Minisatellites are useful for genetic mapping*).

The buoyant density of a duplex DNA depends on its G·C content according to the empirical formula

$$\rho = 1.660 + 0.00098\ (\%G{\cdot}C)\ \text{g-cm}^{-3}$$

Buoyant density usually is determined by centrifuging DNA through a **density gradient** of CsCl. The DNA forms a band at the position corresponding to its own density. Fractions of DNA differing in G·C content by >5% can usually be separated on a density gradient.

When eukaryotic DNA is centrifuged on a density gradient, two types of material may be distinguished:

- Most of the genome forms a continuum of fragments that appear as a rather broad peak centered on the buoyant density corresponding to the average G·C content of the genome. This is called the main band.
- Sometimes an additional, smaller peak (or peaks) is seen at a different value. This material is the satellite DNA.

Satellites are present in many eukaryotic genomes. They may be either heavier or lighter than the main band; but it is uncommon for them to represent >5% of the total DNA. A clear example is provided by mouse DNA, shown in **Figure 4.19**. The graph is a quantitative scan of the bands formed when mouse DNA is centrifuged through a CsCl density gradient. The main band contains 92% of the genome and is centered on a buoyant density of 1.701 g-cm^{-3} (corresponding to its average G·C of 42%, typical for a mammal). The smaller peak represents 8% of the genome and has a distinct buoyant density of 1.690 g-cm^{-3}. It contains the mouse satellite DNA, whose G·C content (30%) is much lower than any other part of the genome.

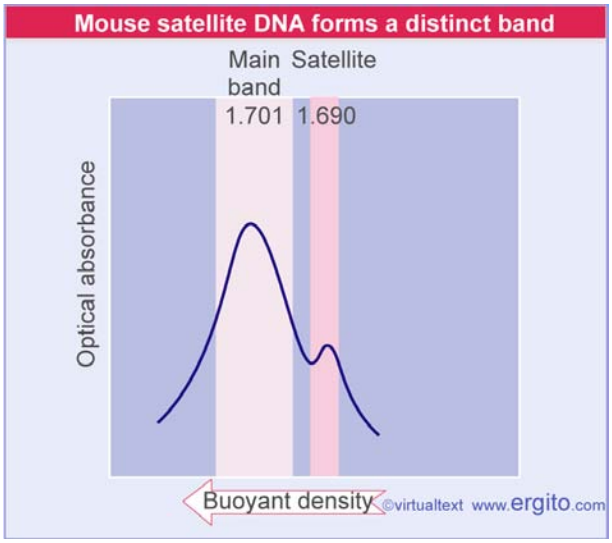


Figure 4.19 Mouse DNA is separated into a main band and a satellite by centrifugation through a density gradient of CsCl.

The behavior of satellite DNA on density gradients is often anomalous. When the actual base composition of a satellite is determined, it is different from the prediction based on its buoyant density. The reason is that ρ is a function not just of base composition, but of the constitution in terms of nearest neighbor pairs. For simple sequences, these are likely to deviate from the random pairwise relationships needed to obey the equation for buoyant density. Also, satellite DNA may be methylated, which changes its density.

Often most of the highly repetitive DNA of a genome can be isolated in the form of satellites. When a highly repetitive DNA component does not separate as a satellite, on isolation its properties often prove to be similar to those of satellite DNA. That is to say that it consists of multiple tandem repeats with anomalous centrifugation. Material isolated in this manner is sometimes referred to as a **cryptic satellite**. Together the cryptic and apparent satellites usually account for all the large tandemly repeated blocks of highly repetitive DNA. When a genome has more than one type of highly repetitive DNA, each exists in its own satellite block (although sometimes different blocks are adjacent).

Where in the genome are the blocks of highly repetitive DNA located? An extension of nucleic acid hybridization techniques allows the location of satellite sequences to be determined directly in the chromosome complement. In the technique of ***in situ* hybridization**, the chromosomal DNA is denatured by treating cells that have been squashed on a cover slip. Then a solution containing a radioactively labeled DNA or RNA probe is added. The probe hybridizes with its complements in the denatured genome. The location of the sites of hybridization can be determined by autoradiography (see **Figure 19.19**).

Satellite DNAs are found in regions of **heterochromatin**. Heterochromatin is the term used to describe regions of chromosomes that are permanently tightly coiled up and inert, in contrast with the **euchromatin** that represents most of the genome (see *Molecular Biology 5.19.7 Chromatin is divided into euchromatin and heterochromatin*). Heterochromatin is commonly found at centromeres (the regions

where the kinetochores are formed at mitosis and meiosis for controlling chromosome movement). The centromeric location of satellite DNA suggests that it has some structural function in the chromosome. This function could be connected with the process of chromosome segregation.

An example of the localization of satellite DNA for the mouse chromosomal complement is shown in **Figure 4.20**. In this case, one end of each chromosome is labeled, because this is where the centromeres are located in *M. musculus* chromosomes.

Figure 4.20 Cytological hybridization shows that mouse satellite DNA is located at the centromeres. Photograph kindly provided by Mary Lou Pardue and Joe Gall.

This content is available online at http://www.ergito.com/main.jsp?bcs=MBIO.1.4.11

CLUSTERS AND REPEATS

1.4.12 Arthropod satellites have very short identical repeats

Key Terms

Heavy strands and light strands of a DNA duplex refer to the density differences that result when there is an asymmetry between base representation in the two strands such that one strand is rich in T and G bases and the other is rich in C and A bases. This occurs in some satellite and mitochondrial DNAs.

Key Concepts

- The repeating units of arthropod satellite DNAs are only a few nucleotides long. Most of the copies of the sequence are identical.

In the arthropods, as typified by insects and crabs, each satellite DNA appears to be rather homogeneous. Usually, a single very short repeating unit accounts for >90% of the satellite. This makes it relatively straightforward to determine the sequence.

Drosophila virilis has three major satellites and also a cryptic satellite, together representing >40% of the genome. The sequences of the satellites are summarized in **Figure 4.21**. The three major satellites have closely related sequences. A single base substitution is sufficient to generate either satellite II or III from the sequence of satellite I.

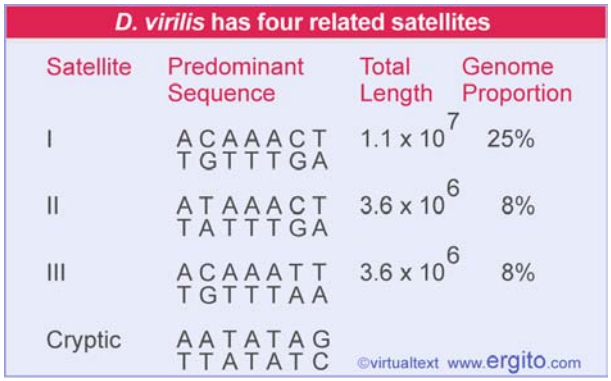


Figure 4.21 Satellite DNAs of *D. virilis* are related. More than 95% of each satellite consists of a tandem repetition of the predominant sequence.

The satellite I sequence is present in other species of *Drosophila* related to *virilis*, and so may have preceded speciation. The sequences of satellites II and III seem to be specific to *D. virilis*, and so may have evolved from satellite I after speciation.

The main feature of these satellites is their very short repeating unit: only 7 bp. Similar satellites are found in other species. *D. melanogaster* has a variety of

satellites, several of which have very short repeating units (5, 7, 10, or 12 bp). Comparable satellites are found in the crabs.

The close sequence relationship found among the *D. virilis* satellites is not necessarily a feature of other genomes, where the satellites may have unrelated sequences. *Each satellite has arisen by a lateral amplification of a very short sequence.* This sequence may represent a variant of a previously existing satellite (as in *D. virilis*), or could have some other origin.

Satellites are continually generated and lost from genomes. This makes it difficult to ascertain evolutionary relationships, since a current satellite could have evolved from some previous satellite that has since been lost. The important feature of these satellites is that *they represent very long stretches of DNA of very low sequence complexity, within which constancy of sequence can be maintained*

One feature of many of these satellites is a pronounced asymmetry in the orientation of base pairs on the two strands. In the example of the *D. virilis* satellites shown in **Figure 4.20**, in each of the major satellites one of the strands is much richer in T and G bases. This increases its buoyant density, so that upon denaturation this **heavy strand** (H) can be separated from the complementary light strand (L). This can be useful in sequencing the satellite.

CLUSTERS AND REPEATS

1.4.13 Mammalian satellites consist of hierarchical repeats

Key Concepts

- Mouse satellite DNA has evolved by duplication and mutation of a short repeating unit to give a basic repeating unit of 234 bp in which the original half, quarter, and eighth repeats can be recognized.

In the mammals, as typified by various rodents, the sequences comprising each satellite show appreciable divergence between tandem repeats. Common short sequences can be recognized by their preponderance among the oligonucleotide fragments released by chemical or enzymatic treatment. However, the predominant short sequence usually accounts for only a small minority of the copies. The other short sequences are related to the predominant sequence by a variety of substitutions, deletions, and insertions.

But a series of these variants of the short unit can constitute a longer repeating unit that is itself repeated in tandem with some variation. So mammalian satellite DNAs are constructed from a hierarchy of repeating units. These longer repeating units constitute the sequences that renature in reassociation analysis. They can also be recognized by digestion with restriction enzymes.

When any satellite DNA is digested with an enzyme that has a recognition site in its repeating unit, one fragment will be obtained for every repeating unit in which the site occurs. In fact, when the DNA of a eukaryotic genome is digested with a restriction enzyme, most of it gives a general smear, due to the random distribution of cleavage sites. But satellite DNA generates sharp bands, because a large number of fragments of identical or almost identical size are created by cleavage at restriction sites that lie a regular distance apart.

Determining the sequence of satellite DNA can be difficult. Using the discrete bands generated by restriction cleavage, we can attempt to obtain a sequence directly. However, if there is appreciable divergence between individual repeating units, different nucleotides will be present at the same position in different repeats, so the sequencing gels will be obscure. If the divergence is not too great – say, within ~2% – it may be possible to determine an average repeating sequence.

Individual segments of the satellite can be inserted into plasmids for cloning. A difficulty is that the satellite sequences tend to be excised from the chimeric plasmid by recombination in the bacterial host. However, when the cloning succeeds, it is possible to determine the sequence of the cloned segment unambiguously. While this gives the actual sequence of a repeating unit or units, we should need to have many individual such sequences to reconstruct the type of divergence typical of the satellite as a whole.

By either sequencing approach, the information we can gain is limited to the distance that can be analyzed on one set of sequence gels. The repetition of divergent tandem copies makes it impossible to reconstruct longer sequences by obtaining overlaps between individual restriction fragments.

The satellite DNA of the mouse *M. musculus* is cleaved by the enzyme EcoRII into a series of bands, including a predominant monomeric fragment of 234 bp. This sequence must be repeated with few variations throughout the 60-70% of the satellite that is cleaved into the monomeric band. We may analyze this sequence in terms of its successively smaller constituent repeating units.

Figure 4.22 depicts the sequence in terms of two half-repeats. By writing the 234 bp sequence so that the first 117 bp are aligned with the second 117 bp, we see that the two halves are quite well related. They differ at 22 positions, corresponding to 19% divergence. This means that the current 234 bp repeating unit must have been generated at some time in the past by duplicating a 117 bp repeating unit, after which differences accumulated between the duplicates.

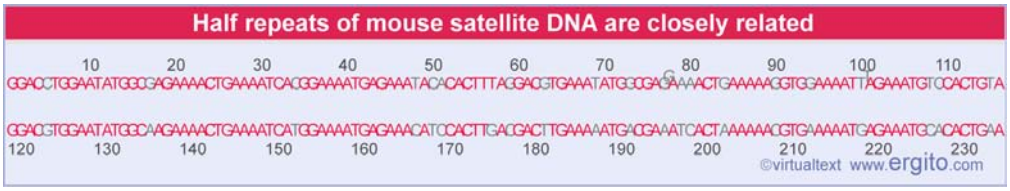


Figure 4.22 The repeating unit of mouse satellite DNA contains two half-repeats, which are aligned to show the identities (in red).

Within the 117 bp unit, we can recognize two further subunits. Each of these is a quarter-repeat relative to the whole satellite. The four quarter-repeats are aligned in **Figure 4.23**. The upper two lines represent the first half-repeat of **Figure 4.22**; the lower two lines represent the second half-repeat. We see that the divergence between the four quarter-repeats has increased to 23 out of 58 positions, or 40%. The first three quarter-repeats are somewhat better related, and a large proportion of the divergence is due to changes in the fourth quarter-repeat.

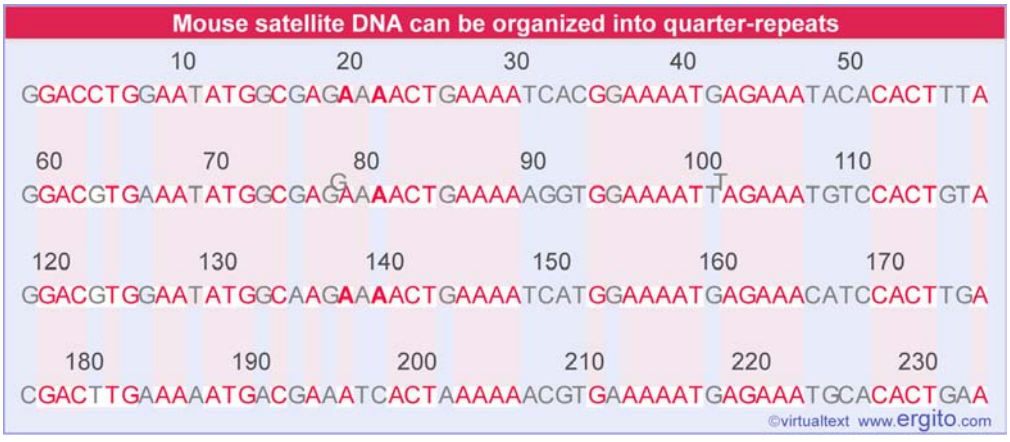


Figure 4.23 The alignment of quarter-repeats identifies homologies between the first and second half of each half-repeat. Positions that are the same in all 4 quarter-repeats are shown in color; identities that extend only through 3 quarter-repeats are indicated by grey letters in the pink area.

Looking within the quarter-repeats, we find that each consists of two related subunits (one-eighth-repeats), shown as the α and β sequences in **Figure 4.24**. The α sequences all have an insertion of a C, and the β sequences all have an insertion of a trinucleotide, relative to a common consensus sequence. This suggests that the quarter-repeat originated by the duplication of a sequence like the consensus sequence, after which changes occurred to generate the components we now see as α and β. Further changes then took place between tandemly repeated α β sequences to generate the individual quarter- and half-repeats that exist today. Among the one-eighth-repeats, the present divergence is 19/31 = 61%.

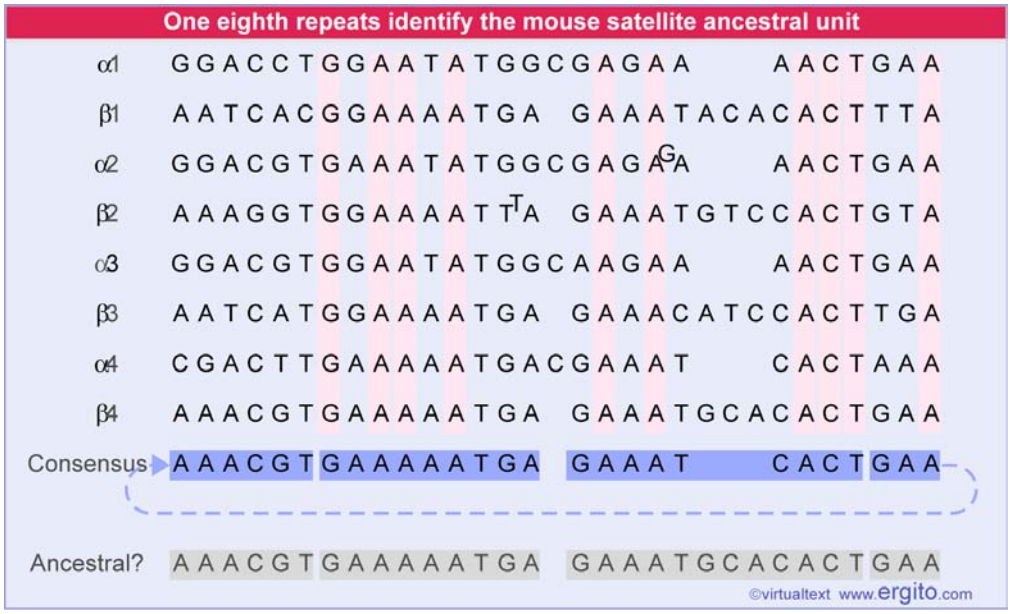


Figure 4.24 The alignment of eighth-repeats shows that each quarter-repeat consists of an α and a β half. The consensus sequence gives the most common base at each position. The "ancestral" sequence shows a sequence very closely related to the consensus sequence, which could have been the predecessor to the α and β units. (The satellite sequence is continuous, so that for the purposes of deducing the consensus sequence, we can treat it as a circular permutation, as indicated by joining the last GAA triplet to the first 6 bp.)

The consensus sequence is analyzed directly in **Figure 4.25**, which demonstrates that the current satellite sequence can be treated as derivatives of a 9 bp sequence. We can recognize three variants of this sequence in the satellite, as indicated at the bottom of **Figure 4.25**. If in one of the repeats we take the next most frequent base at two positions instead of the most frequent, we obtain three well-related 9 bp sequences.

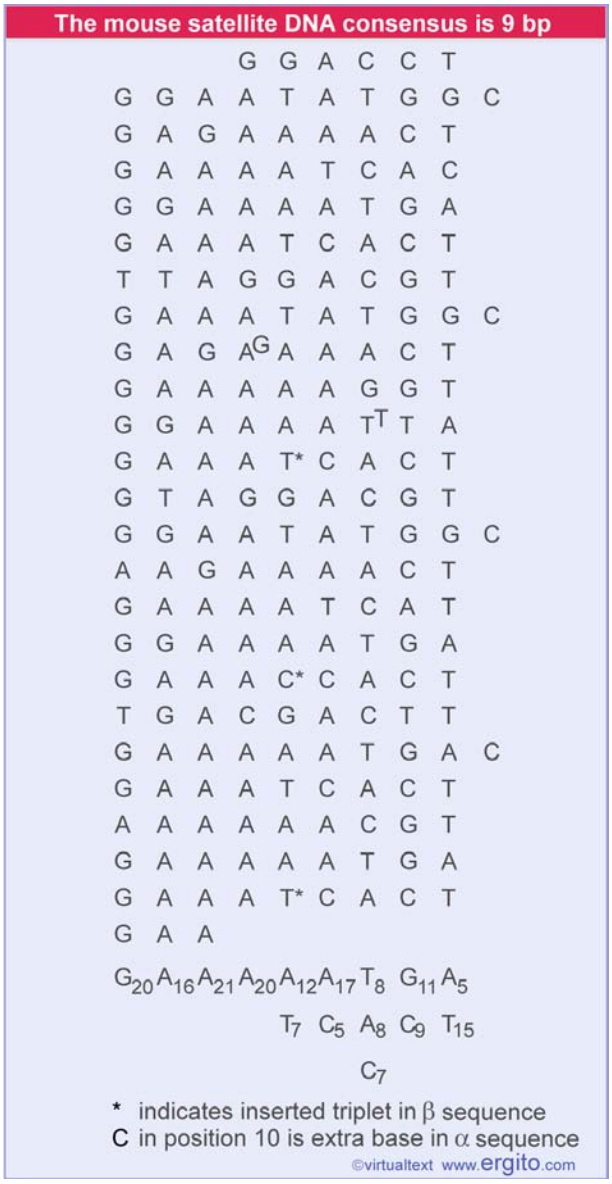


Figure 4.25 The existence of an overall consensus sequence is shown by writing the satellite sequence in terms of a 9 bp repeat.

G A A A A A C G T

G A A A A A T G A

G A A A A A A C T

The origin of the satellite could well lie in an amplification of one of these three nonamers. The overall consensus sequence of the present satellite is GAAAAA$^{AG}_{TC}$T, which is effectively an amalgam of the three 9 bp repeats.

The average sequence of the monomeric fragment of the mouse satellite DNA explains its properties. The longest repeating unit of 234 bp is identified by the restriction cleavage. The unit of reassociation between single strands of denatured satellite DNA is probably the 117 bp half-repeat, because the 234 bp fragments can anneal both in register and in half-register (in the latter case, the first half-repeat of one strand renatures with the second half-repeat of the other).

So far, we have treated the present satellite as though it consisted of identical copies of the 234 bp repeating unit. Although this unit accounts for the majority of the satellite, variants of it also are present. Some of them are scattered at random throughout the satellite; others are clustered.

The existence of variants is implied by our description of the starting material for the sequence analysis as the "monomeric" fragment. When the satellite is digested by an enzyme that has one cleavage site in the 234 bp sequence, it also generates dimers, trimers, and tetramers relative to the 234 bp length. They arise when a repeating unit has lost the enzyme cleavage site as the result of mutation.

The monomeric 234 bp unit is generated when two adjacent repeats each have the recognition site. A dimer occurs when one unit has lost the site, a trimer is generated when two adjacent units have lost the site, and so on. With some restriction enzymes, most of the satellite is cleaved into a member of this repeating series, as shown in the example of **Figure 4.26**. The declining number of dimers, trimers, etc. shows that there is a random distribution of the repeats in which the enzyme's recognition site has been eliminated by mutation.

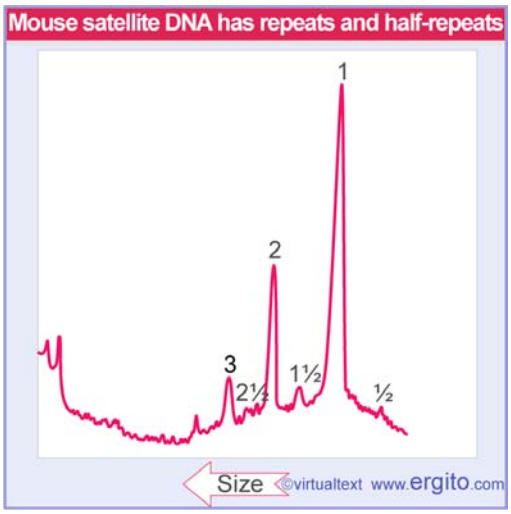


Figure 4.26 Digestion of mouse satellite DNA with the restriction enzyme EcoRII identifies a series of repeating units (1, 2, 3) that are multimers of 234 bp and also a minor series (½,1 ½, 2½) that includes half-repeats (see text later). The band at the far left is a fraction resistant to digestion.

Other restriction enzymes show a different type of behavior with the satellite DNA. They continue to generate the same series of bands. But they cleave only a small proportion of the DNA, say 5-10%. This implies that a certain region of the satellite

contains a concentration of the repeating units with this particular restriction site. Presumably the series of repeats in this domain all are derived from an ancestral variant that possessed this recognition site (although in the usual way, some members since have lost it by mutation).

A satellite DNA suffers unequal recombination. This has additional consequences when there is internal repetition in the repeating unit. Let us return to our cluster consisting of "ab" repeats. Suppose that the "a" and "b" components of the repeating unit are themselves sufficiently well related to pair. Then the two clusters can align in *half-register*, with the "a" sequence of one aligned with the "b" sequence of the other. How frequently this occurs will depend on the closeness of the relationship between the two halves of the repeating unit. In mouse satellite DNA, reassociation between the denatured satellite DNA strands *in vitro* commonly occurs in the half-register.

When a recombination event occurs out of register, it changes the length of the repeating units that are involved in the reaction.

In the upper recombinant cluster, an "ab" unit has been replaced by an "aab" unit. In the lower cluster, the "ab" unit has been replaced by a "b" unit.

This type of event explains a feature of the restriction digest of mouse satellite DNA. **Figure 4.25** shows a fainter series of bands at lengths of ½, 1½, 2½, and 3½ repeating units, in addition to the stronger integral length repeats. Suppose that in the preceding example, "ab" represents the 234 bp repeat of mouse satellite DNA, generated by cleavage at a site in the "b" segment. The "a" and "b" segments correspond to the 117 bp half-repeats.

Then in the upper recombinant cluster, the "aab" unit generates a fragment of 1½ times the usual repeating length. And in the lower recombinant cluster, the "b" unit generates a fragment of half of the usual length. (The multiple fragments in the half-repeat series are generated in the same way as longer fragments in the integral series, when some repeating units have lost the restriction site by mutation.)

Turning the argument the other way around, the identification of the half-repeat series on the gel shows that the 234 bp repeating unit consists of two half-repeats well enough related to pair sometimes for recombination. Also visible in **Figure 4.26** are some rather faint bands corresponding to ¼- and ¾-spacings. These will be

generated in the same way as the ½-spacings, when recombination occurs between clusters aligned in a quarter-register. The decreased relationship between quarter-repeats compared with half-repeats explains the reduction in frequency of the ¼- and ¾-bands compared with the ½-bands.

CLUSTERS AND REPEATS

1.4.14 Minisatellites are useful for genetic mapping

Key Terms

Microsatellite DNAs consist of repetitions of extremely short (typically <10 bp) units.

Minisatellite DNAs consist of ~10 copies of a short repeating sequence. the length of the repeating unit is measured in 10s of base pairs. The number of repeats varies between individual genomes.

VNTR (variable number tandem repeat) regions describe very short repeated sequences, including microsatellites and minisatellites.

DNA fingerprinting analyzes the differences between individuals of the fragments generated by using restriction enzymes to cleave regions that contain short repeated sequences. Because these are unique to every individual, the presence of a particular subset in any two individuals can be used to define their common inheritance (e.g. a parent-child relationship).

Key Concepts

- The variation between microsatellites or minisatellites in individual genomes can be used to identify heredity unequivocally by showing that 50% of the bands in an individual are derived from a particular parent.

Sequences that resemble satellites in consisting of tandem repeats of a short unit, but that overall are much shorter, consisting of (for example) from 5-50 repeats, are common in mammalian genomes. They were discovered by chance as fragments whose size is extremely variable in genomic libraries of human DNA. The variability is seen when a population contains fragments of many different sizes that represent the same genomic region; when individuals are examined, it turns out that there is extensive polymorphism, and that many different alleles can be found (416).

The name **microsatellite** is usually used when the length of the repeating unit is <10 bp, and the name **minisatellite** is used when the length of the repeating unit is ~10-100 bp, but the terminology is not precisely defined. These types of sequences are also called **VNTR** (variable number tandem repeat) regions.

The cause of the variation between individual genomes at microsatellites or minisatellites is that individual alleles have different numbers of the repeating unit. For example, one minisatellite has a repeat length of 64 bp, and is found in the population with the following distribution:

7% 18 repeats

11% 16 repeats

43% 14 repeats

36% 13 repeats

4% 10 repeats

The rate of genetic exchange at minisatellite sequences is high, ~10^{-4} per kb of DNA. (The frequency of exchanges per actual locus is assumed to be proportional to the length of the minisatellite.) This rate is ~10× greater than the rate of homologous recombination at meiosis, that is, in any random DNA sequence.

The high variability of minisatellites makes them especially useful for genomic mapping, because there is a high probability that individuals will vary in their alleles at such a locus. An example of mapping by minisatellites is illustrated in **Figure 4.27**. This shows an extreme case in which two individuals both are heterozygous at a minisatellite locus, and in fact all four alleles are different. All progeny gain one allele from each parent in the usual way, and it is possible unambiguously to determine the source of every allele in the progeny. In the terminology of human genetics, the meioses described in this figure are highly informative, because of the variation between alleles.

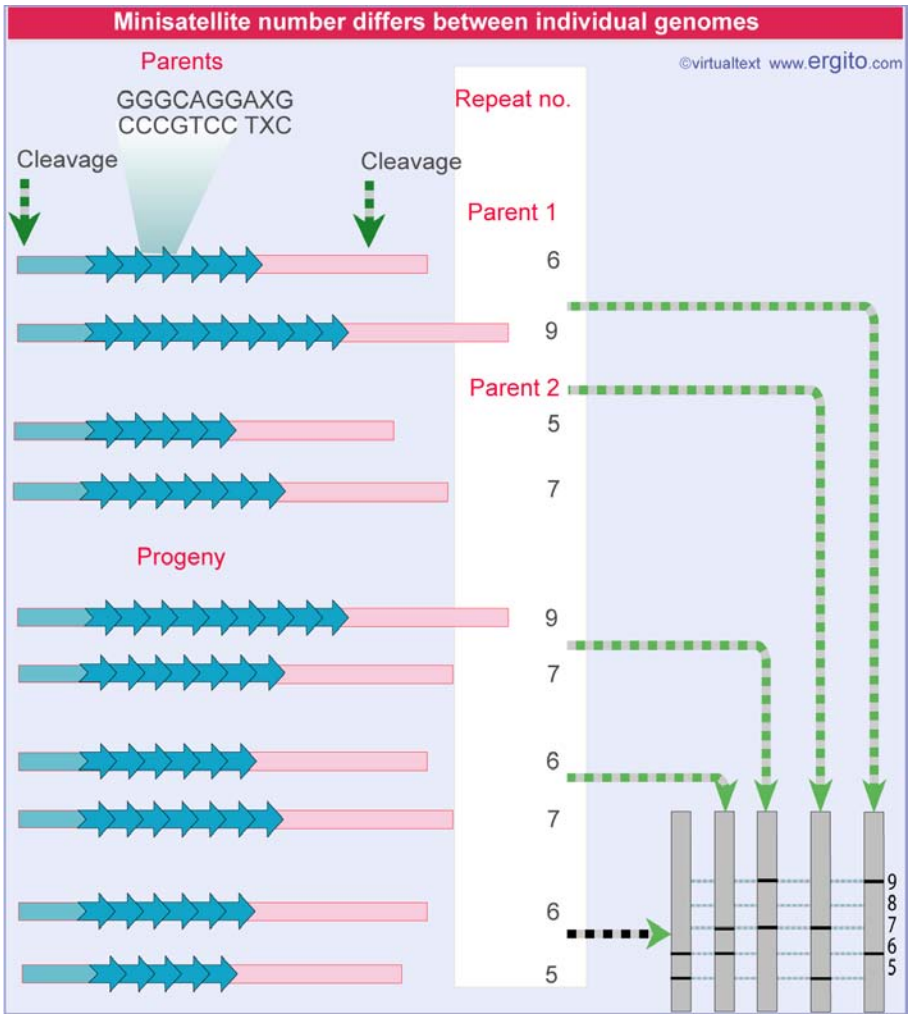


Figure 4.27 Alleles may differ in the number of repeats at a minisatellite locus, so that cleavage on either side generates restriction fragments that differ in length. By using a minisatellite with alleles that differ between parents, the pattern of inheritance can be followed.

One family of minisatellites in the human genome share a common "core" sequence. The core is a G#C-rich sequence of 10-15 bp, showing an asymmetry of purine/pyrimidine distribution on the two strands. Each individual minisatellite has a variant of the core sequence, but ~1000 minisatellites can be detected on Southern blot by a probe consisting of the core sequence.

Consider the situation shown in **Figure 4.27**, but multiplied 1000×. The effect of the variation at individual loci is to create a unique pattern for every individual. This makes it possible to assign heredity unambiguously between parents and progeny, by showing that 50% of the bands in any individual are derived from a particular parent. This is the basis of the technique known as **DNA fingerprinting**.

Both microsatellites and minisatellites are unstable, although for different reasons. Microsatellites undergo intrastrand mispairing, when slippage during replication leads to expansion of the repeat, as shown in **Figure 4.28** (3044). Systems that repair damage to DNA, in particular those that recognize mismatched base pairs, are

important in reversing such changes, as shown by a large increase in frequency when repair genes are inactivated (2255). Because mutations in repair systems are an important contributory factor in the development of cancer, tumor cells often display variations in microsatellite sequences (see *Molecular Biology 6.30.29 Defects in repair systems cause mutations to accumulate in tumors*).

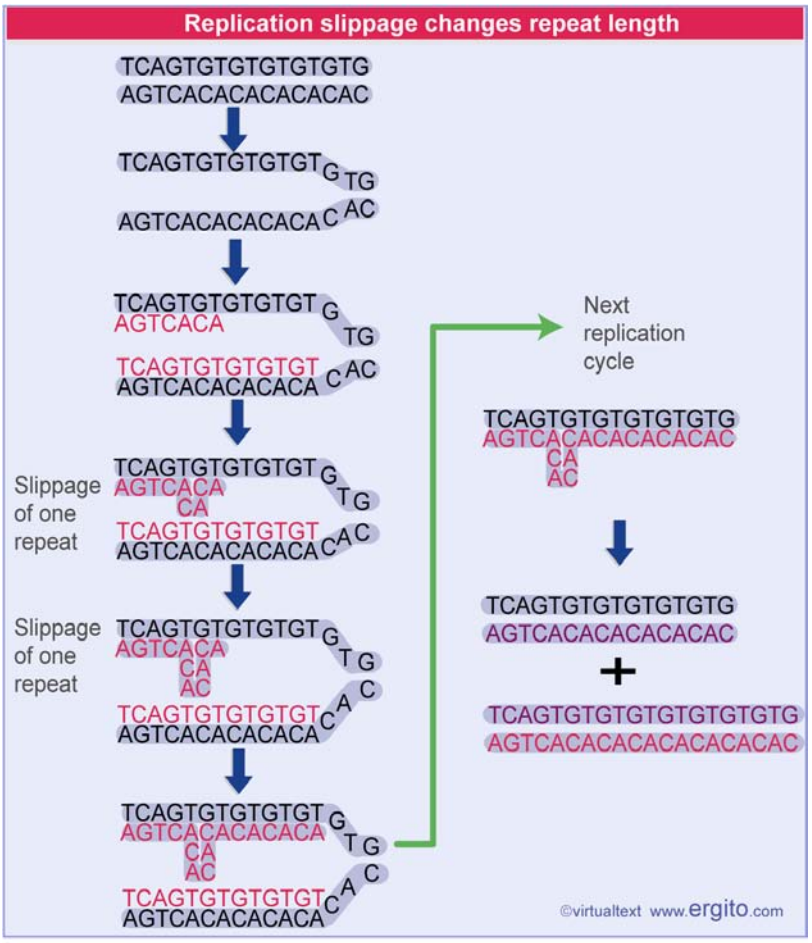


Figure 4.28 Replication slippage occurs when the daughter strand slips back one repeating unit in pairing with the template strand. Each slippage event adds one repeating unit to the daughter strand. The extra repeats are extruded as a single strand loop. Replication of this daughter strand in the next cycle generates a duplex DNA with an increased number of repeats.

Minisatellites undergo the same sort of unequal crossing-over between repeats that we have discussed for satellites (see **Figure 4.1**). One telling case is that increased variation is associated with a meiotic hotspot (3049). The recombination event is not usually associated with recombination between flanking markers, but has a complex form in which the new mutant allele gains information from both the sister chromatid and the other (homologous) chromosome (3048).

It is not clear at what repeating length the cause of the variation shifts from replication slippage to recombination.

Last updated on 10-21-2002

References

416. Jeffreys, A. J., Wilson, V., and Thein, S. L. (1985). *Hypervariable minisatellite regions in human DNA.* Nature 314, 67-73.

2255. Strand, M., Prolla, T. A., Liskay, and Petes, T. D. (1993). *Destabilization of tracts of simple repetitive DNA in yeast by mutations affecting DNA mismatch repair.* Nature 365, 274-276.

3044. Jeffreys, A. J., Jeffreys, A. J., Jeffreys, A. J., Royle, N. J., Wilson, V., and Wong, Z. (1988). *Spontaneous mutation rates to new length alleles at tandem-repetitive hypervariable loci in human DNA.* Nature 332, 278-281.

3048. Jeffreys, A. J., Tamaki, K., MacLeod, A., Monckton, D. G., Neil, D. L., and Armour, J. A. (1994). *Complex gene conversion events in germline mutation at human minisatellites.* Nat. Genet. 6, 136-145.

3049. Jeffreys, A.J., Murray, J., and Neumann, R. (1998). *High-resolution mapping of crossovers in human sperm defines a minisatellite-associated recombination hotspot.* Mol. Cell 2, 267-273.

CLUSTERS AND REPEATS

1.4.15 Summary

Almost all genes belong to families, defined by the possession of related sequences in the exons of individual members. Families evolve by the duplication of a gene (or genes), followed by divergence between the copies. Some copies suffer inactivating mutations and become pseudogenes that no longer have any function. Pseudogenes also may be generated as DNA copies of the mRNA sequences.

An evolving set of genes may remain together in a cluster or may be dispersed to new locations by chromosomal rearrangement. The organization of existing clusters can sometimes be used to infer the series of events that has occurred. These events act with regard to sequence rather than function, and therefore include pseudogenes as well as active genes.

Mutations accumulate more rapidly in silent sites than in replacement sites (which affect the amino acid sequence). The rate of divergence at replacement sites can be used to establish a clock, calibrated in percent divergence per million years. The clock can then be used to calculate the time of divergence between any two members of the family.

A tandem cluster consists of many copies of a repeating unit that includes the transcribed sequence(s) and a nontranscribed spacer(s). rRNA gene clusters code only for a single rRNA precursor. Maintenance of active genes in clusters depends on mechanisms such as gene conversion or unequal crossing-over that cause mutations to spread through the cluster, so that they become exposed to evolutionary pressure.

Satellite DNA consists of very short sequences repeated many times in tandem. Its distinct centrifugation properties reflect its biased base composition. Satellite DNA is concentrated in centromeric heterochromatin, but its function (if any) is unknown. The individual repeating units of arthropod satellites are identical. Those of mammalian satellites are related, and can be organized into a hierarchy reflecting the evolution of the satellite by the amplification and divergence of randomly chosen sequences.

Unequal crossing-over appears to have been a major determinant of satellite DNA organization. Crossover fixation explains the ability of variants to spread through a cluster.

Minisatellites and microsatellites consist of even shorter repeating sequences than satellites, <10 bp for microsatellites and 10-50 bp for minisatellites. The number of repeating units is usually 5-50. There is high variation in the repeat number between individual genomes. Microsatellite repeat number varies as the result of slippage during replication; the frequency is affected by systems that recognize and repair damage in DNA. Minisatellite repeat number varies as the result of recombination-like events. Variations in repeat number can be used to determine hereditary relationships by the technique known as DNA fingerprinting.