THE CONTENT OF THE GENOME

1.3.1 Introduction

Key Terms

The **genome** is the complete set of sequences in the genetic material of an organism. It includes the sequence of each chromosome plus any DNA in organelles.

The **transcriptome** is the complete set of RNAs present in a cell, tissue, or organism. Its complexity is due mostly to mRNAs, but it also includes noncoding RNAs.

The **proteome** is the complete set of proteins that is expressed by the entire genome. Because some genes code for multiple proteins, the size of the proteome is greater than the number of genes. Sometimes the term is used to describe complement of proteins expressed by a cell at any one time.

The key question about the genome is how many genes it contains. We can think about the total number of genes at four levels, corresponding to successive stages in gene expression:

- The **genome** is the complete set of genes of an organism. Ultimately it is defined by the complete DNA sequence, although as a practical matter it may not be possible to identify every gene unequivocally solely on the basis of sequence.
- The **transcriptome** is the complete set of genes expressed under particular conditions. It is defined in terms of the set of RNA molecules that is present, and can refer to a single cell type or to any more complex assembly of cells up to the complete organism. Because some genes generate multiple mRNAs, the transcriptome is likely to be larger than the number of genes defined directly in the genome. The transcriptome includes noncoding RNAs as well as mRNAs.
- The **proteome** is the complete set of proteins. It should correspond to the mRNAs in the transcriptome, although there can be differences of detail reflecting changes in the relative abundance or stabilities of mRNAs and proteins. It can be used to refer to the set of proteins coded by the whole genome or produced in any particular cell or tissue.
- Proteins may function independently or as part of multiprotein assemblies. If we could identify all protein-protein interactions, we could define the total number of independent assemblies of proteins.

The number of genes in the genome can be identified directly by defining open reading frames. Large scale mapping of this nature is complicated by the fact that interrupted genes may consist of many separated open reading frames. Since we do not necessarily have information about the functions of the protein products, or indeed proof that they are expressed at all, this approach is restricted to defining the *potential* of the genome. However, a strong presumption exists that any conserved open reading frame is likely to be expressed.

Another approach is to define the number of genes directly in terms of the transcriptome (by directly identifying all the mRNAs) or proteome (by directly identifying all the proteins). This gives an assurance that we are dealing with *bona fide* genes that are expressed under known circumstances. It allows us to ask how many genes are expressed in a particular tissue or cell type, what variation exists in the relative levels of expression, and how many of the genes expressed in one particular cell are unique to that cell or are also expressed elsewhere.

Concerning the types of genes, we may ask whether a particular gene is essential: what happens to a null mutant? If a null mutation is lethal, or the organism has a visible defect, we may conclude that the gene is essential or at least conveys a selective advantage. But some genes can be deleted without apparent effect on the phenotype. Are these genes really dispensable, or does a selective disadvantage result from the absence of the gene, perhaps in other circumstances, or over longer periods of time?

This content is available online at http://www.ergito.com/main.jsp?bcs=MBIO.1.3.1

THE CONTENT OF THE GENOME

1.3.2 Genomes can be mapped by linkage, restriction cleavage, or DNA sequence

Defining the contents of a genome essentially means making a map. We can think about mapping genes and genomes at several levels of resolution:

- A genetic (or linkage) map identifies the distance between mutations in terms of recombination frequencies. It is limited by its reliance on the occurrence of mutations that affect the phenotype. Because recombination frequencies can be distorted relative to the physical distance between sites, it does not accurately represent physical distances along the genetic material.

- A linkage map can also be constructed by measuring recombination between sites in genomic DNA. These sites have sequence variations that generate differences in the susceptibility to cleavage by certain (restriction) enzymes. Because such variations are common, such a map can be prepared for any organism irrespective of the occurrence of mutants. It has the same disadvantage as any linkage map that the relative distances are based on recombination.

- A restriction map is constructed by cleaving DNA into fragments with restriction enzymes and measuring the distances between the sites of cleavage. This represents distances in terms of the length of DNA, so it provides a physical map of the genetic material. A restriction map does not intrinsically identify sites of genetic interest. For it to be related to the genetic map, mutations have to be characterized in terms of their effects upon the restriction sites. Large changes in the genome can be recognized because they affect the sizes or numbers of restriction fragments. Point mutations are more difficult to detect.

- The ultimate map is to determine the sequence of the DNA. From the sequence, we can identify genes and the distances between them. By analyzing the protein-coding potential of a sequence of the DNA, we can deduce whether it represents a protein. The basic assumption here is that natural selection prevents the accumulation of damaging mutations in sequences that code for proteins. Reversing the argument, we may assume that an intact coding sequence is likely to be used to generate a protein.

By comparing the sequence of a wild-type DNA with that of a mutant allele, we can determine the nature of a mutation and its exact site of occurrence. This defines the relationship between the genetic map (based entirely on sites of mutation) and the physical map (based on or even comprising the sequence of DNA).

Similar techniques are used to identify and sequence genes and to map the genome, although there is of course a difference of scale. In each case, the principle is to obtain a series of overlapping fragments of DNA, which can be connected into a continuous map. The crucial feature is that each segment is related to the next segment on the map by characterizing the overlap between them, so that we can be sure no segments are missing. This principle is applied both at the level of ordering large fragments into a map, and in connecting the sequences that make up the fragments.

This content is available online at http://www.ergito.com/main.jsp?bcs=MBIO.1.3.2

THE CONTENT OF THE GENOME

1.3.3 Individual genomes show extensive variation

Key Terms

Polymorphism (more fully genetic polymorphism) refers to the simultaneous occurrence in the population of genomes showing variations at a given position. The original definition applied to alleles producing different phenotypes. Now it is also used to describe changes in DNA affecting the restriction pattern or even the sequence. For practical purposes, to be considered as an example of a polymorphism, an allele should be found at a frequency > 1% in the population.

Single nucleotide polymorphism (SNP) describes a polymorphism (variation in sequence between individuals) caused by a change in a single nucleotide. This is responsible for most of the genetic variation between individuals.

Restriction fragment length polymorphism (RFLP) refers to inherited differences in sites for restriction enzymes (for example, caused by base changes in the target site) that result in differences in the lengths of the fragments produced by cleavage with the relevant restriction enzyme. RFLPs are used for genetic mapping to link the genome directly to a conventional genetic marker.

Key Concepts

- Polymorphism may be detected at the phenotypic level when a sequence affects gene function, at the restriction fragment level when it affects a restriction enzyme target site, and at the sequence level by direct analysis of DNA.

- The alleles of a gene show extensive polymorphism at the sequence level, but many sequence changes do not affect function.

The original Mendelian view of the genome classified alleles as either wild-type or mutant. Subsequently we recognized the existence of multiple alleles, each with a different effect on the phenotype. In some cases it may not even be appropriate to define any one allele as "wild-type".

The coexistence of multiple alleles at a locus is called genetic **polymorphism**. Any site at which multiple alleles exist as stable components of the population is by definition polymorphic. An allele is usually defined as polymorphic if it is present at a frequency of >1% in the population.

What is the basis for the polymorphism among the mutant alleles? They possess different mutations that alter the protein function, thus producing changes in phenotype. If we compare the restriction maps or the DNA sequences of these alleles, they too will be polymorphic in the sense that each map or sequence will be different from the others.

Although not evident from the phenotype, the wild type may itself be polymorphic. Multiple versions of the wild-type allele may be distinguished by differences in sequence that do not affect their function, and which therefore do not produce phenotypic variants. A population may have extensive polymorphism at the level of genotype. Many different sequence variants may exist at a given locus; some of them are evident because they affect the phenotype, but others are hidden because they have no visible effect.

So there may be a continuum of changes at a locus, including those that change DNA sequence but do not change protein sequence, those that change protein sequence without changing function, those that create proteins with different activities, and those that create mutant proteins that are nonfunctional.

A change in a single nucleotide when alleles are compared is called a **single nucleotide polymorphism** (**SNP**). One occurs every ~1330 bases in the human genome. Defined by their SNPs, every human being is unique. SNPs can be detected by various means, ranging from direct comparisons of sequence to mass spectroscopy or biochemical methods that produce differences based on sequence variations in a defined region (for examples of SNP maps see 1187; 1188).

One aim of genetic mapping is to obtain a catalog of common variants. The observed frequency of SNPs per genome predicts that, over the human population as a whole (taking the sum of all human genomes of all living individuals), there should be >10 million SNPs that occur at a frequency of >1%. Already >1 million have been identified.

Some polymorphisms in the genome can be detected by comparing the restriction maps of different individuals. The criterion is a change in the pattern of fragments produced by cleavage with a restriction enzyme. **Figure 3.1** shows that when a target site is present in the genome of one individual and absent from another, the extra cleavage in the first genome will generate two fragments corresponding to the single fragment in the second genome.

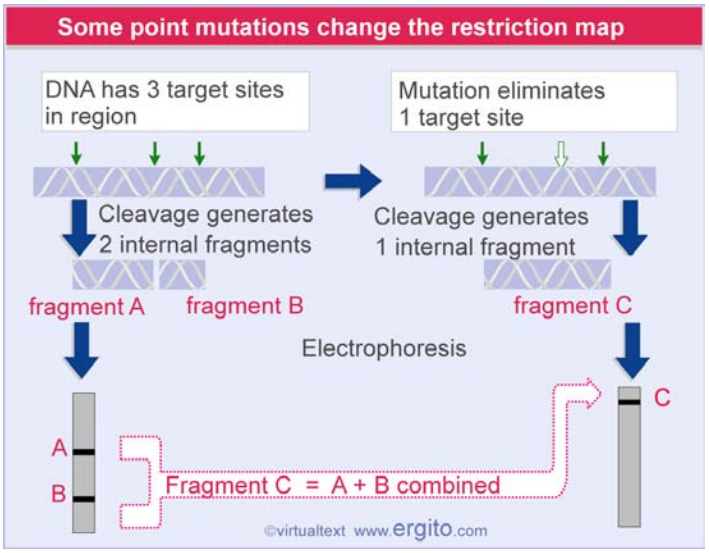

Some point mutations change the restriction map
DNA has 3 target sites in region
Mutation eliminates 1 target site
Cleavage generates 2 internal fragments
Cleavage generates 1 internal fragment
fragment A
fragment B
fragment C
Electrophoresis
A
B
C
Fragment C = A + B combined
©virtualtext www.ergito.com

Figure 3.1 A point mutation that affects a restriction site is detected by a difference in restriction fragments.

Because the restriction map is independent of gene function, a polymorphism at this level can be detected *irrespective of whether the sequence change affects the phenotype*. Probably very few of the restriction site polymorphisms in a genome actually affect the phenotype. Most involve sequence changes that have no effect on the production of proteins (for example, because they lie between genes).

A difference in restriction maps between two individuals is called a **restriction fragment length polymorphism** (**RFLP**). Basically a RFLP is a SNP that is located in the target site for a restriction enzyme. It can be used as a genetic marker in exactly the same way as any other marker. Instead of examining some feature of the phenotype, we directly assess the genotype, as revealed by the restriction map. **Figure 3.2** shows a pedigree of a restriction polymorphism followed through three generations. It displays Mendelian segregation at the level of DNA marker fragments.

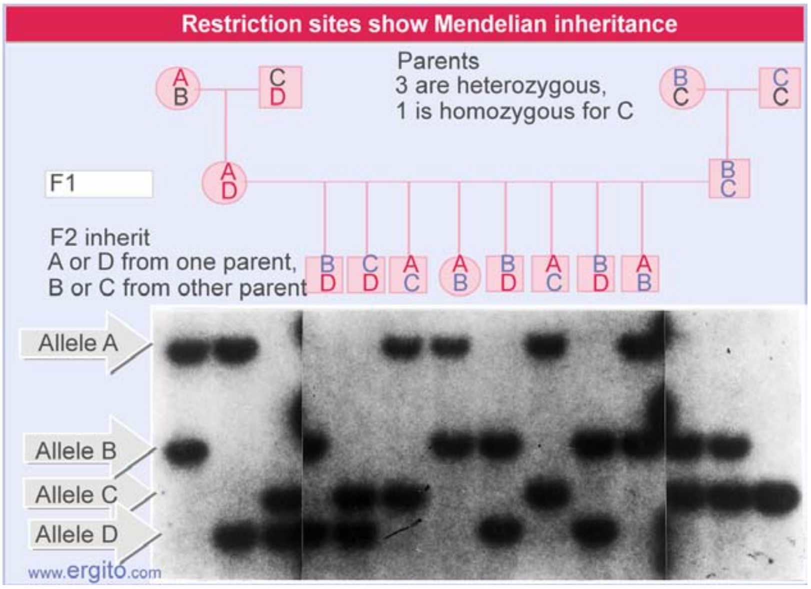


Figure 3.2 Restriction site polymorphisms are inherited according to Mendelian rules. Four alleles for a restriction marker are found in all possible pairwise combinations, and segregate independently at each generation. Photograph kindly provided by Ray White.

Last updated on 8-15-2002

References

1187. Altshuler, D., Pollara, V. J., Cowles, C. R., Van Etten, W. J., Baldwin, J., Linton, L., and Lander, E. S. (2000). *An SNP map of the human genome generated by reduced representation shotgun sequencing*. Nature 407, 513-516.

1188. Mullikin, J. C., Hunt, S. E., Cole, C. G., Mortimore, B. J., Rice, C. M., Burton, J., Matthews, L. H., Pavitt, R., Plumb, R. W., Sims, S. K., Ainscough, R. M., Attwood, J., Bailey, J. M., Barlow, K., Bruskiewich, R. M., Butcher, P. N., Carter, N. P., Chen, Y., and Clee, C. M. (2000). *An SNP map of human chromosome 22.* Nature 407, 516-520.

This content is available online at http://www.ergito.com/main.jsp?bcs=MBIO.1.3.3

THE CONTENT OF THE GENOME

1.3.4 RFLPs and SNPs can be used for genetic mapping

Key Terms

The **haplotype** is the particular combination of alleles in a defined region of some chromosome, in effect the genotype in miniature. Originally used to described combinations of MHC alleles, it now may be used to describe particular combinations of RFLPs, SNPs, or other markers.

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

- RFLPs and SNPs can be the basis for linkage maps and are useful for establishing parent-progeny relationships.

Recombination frequency can be measured between a restriction marker and a visible phenotypic marker as illustrated in **Figure 3.3**. So a genetic map can include both genotypic and phenotypic markers.

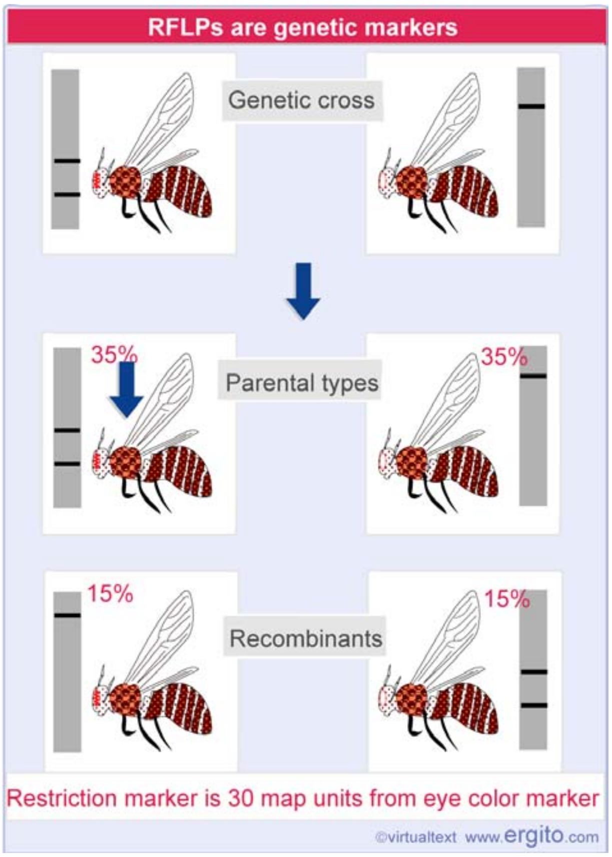

RFLPs are genetic markers
Genetic cross
35%
Parental types
35%
15%
Recombinants
15%
Restriction marker is 30 map units from eye color marker
©virtualtext www.ergito.com

Figure 3.3 A restriction polymorphism can be used as a genetic marker to measure recombination distance from a phenotypic marker (such as eye color). The figure simplifies the situation by showing only the DNA bands corresponding to the allele of one genome in a diploid.

Because restriction markers are not restricted to those genome changes that affect the phenotype, they provide the basis for an extremely powerful technique for identifying genetic loci at the molecular level. A typical problem concerns a mutation with known effects on the phenotype, where the relevant genetic locus can be placed on a genetic map, but for which we have no knowledge about the corresponding gene or protein. Many damaging or fatal human diseases fall into this category. For example cystic fibrosis shows Mendelian inheritance, but the molecular nature of the mutant function was unknown until it could be identified as a result of characterizing the gene.

If restriction polymorphisms occur at random in the genome, some should occur near any particular target gene. We can identify such restriction markers by virtue of their tight linkage to the mutant phenotype. If we compare the restriction map of DNA from patients suffering from a disease with the DNA of normal people, we may find that a particular restriction site is always present (or always absent) from the patients.

A hypothetical example is shown in **Figure 3.4**. This situation corresponds to finding 100% linkage between the restriction marker and the phenotype. It would imply that the restriction marker lies so close to the mutant gene that it is never separated from it by recombination.

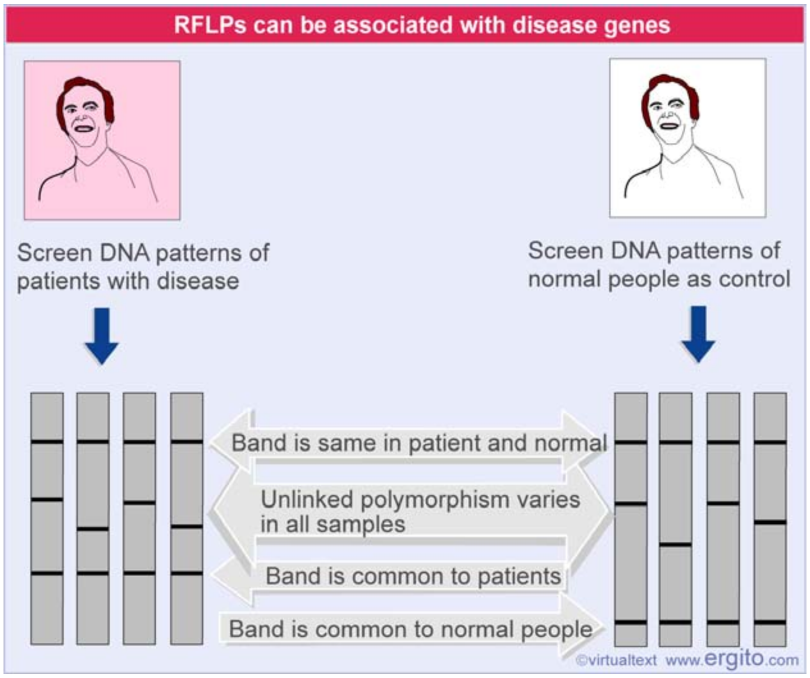

RFLPs can be associated with disease genes
Screen DNA patterns of patients with disease
Screen DNA patterns of normal people as control
Band is same in patient and normal
Unlinked polymorphism varies in all samples
Band is common to patients
Band is common to normal people
©virtualtext www.ergito.com

Figure 3.4 If a restriction marker is associated with a phenotypic characteristic, the restriction site must be located near the gene responsible for the phenotype. The mutation changing the band that is common in normal people into the band that is common in patients is very closely linked to the disease gene.

The identification of such a marker has two important consequences:

- It may offer a diagnostic procedure for detecting the disease. Some of the human diseases that are genetically well characterized but ill defined in molecular terms cannot be easily diagnosed. If a restriction marker is reliably linked to the phenotype, then its presence can be used to diagnose the disease.
- It may lead to isolation of the gene. The restriction marker must lie relatively near the gene on the genetic map if the two loci rarely or never recombine. Although "relatively near" in genetic terms can be a substantial distance in terms of base pairs of DNA, nonetheless it provides a starting point from which we can proceed along the DNA to the gene itself.

The frequent occurrence of SNPs in the human genome makes them useful for genetic mapping. From the 1.4×10^6 SNPs that have already been identified, there is on average an SNP every 1-2 kb. This should allow rapid localization of new disease genes by locating them between the nearest SNPs (1442).

On the same principle, RFLP mapping has been in use for some time. Once an RFLP has been assigned to a linkage group, it can be placed on the genetic map. RFLP mapping in man and mouse has led to the construction of linkage maps for both genomes. Any unknown site can be tested for linkage to these sites and by this means rapidly placed on to the map (393; 394; 395). Because there are fewer RFLPs than SNPs, the resolution of the RFLP map is in principle more limited.

The frequency of polymorphism means that every individual has a unique constellation of SNPs or RFLPs. The particular combination of sites found in a specific region is called a **haplotype**, a genotype in miniature. Haplotype was originally introduced as a concept to describe the genetic constitution of the major histocompatibility locus, a region specifying proteins of importance in the immune system (see *Molecular Biology 5.25 Immune diversity*). The concept now has been extended to describe the particular combination of alleles or restriction sites (or any other genetic marker) present in some defined area of the genome.

The existence of RFLPs provides the basis for a technique to establish unequivocal parent-progeny relationships. In cases where parentage is in doubt, a comparison of the RFLP map in a suitable chromosome region between potential parents and child allows absolute assignment of the relationship. The use of DNA restriction analysis to identify individuals has been called **DNA fingerprinting**. Analysis of especially variable "minisatellite" sequences is used mapping in the human genome (for review see 11; 13) (see *Molecular Biology 1.4.14 Minisatellites are useful for genetic mapping*).

Last updated on 2-16-2001

Reviews

11. White, R. et al. (1985). *Construction of linkage maps with DNA markers for human chromosomes.* Nature 313, 101-105.

13. Gusella, J. F. (1986). *DNA polymorphism and human disease.* Annu. Rev. Biochem. 55, 831-854.

References

393. Donis-Keller, J. et al. (1987). *A genetic linkage map of the human genome*. Cell 51, 319-337.

394. Dietrich, W. F. et al. (1996). *A comprehensive genetic map of the mouse genome*. Nature 380, 149-152.

395. Dib, C. et al. (1996). *A comprehensive genetic map of the human genome based on 5,264 microsatellites*. Nature 380, 152-154.

1442. Sachidanandam, R. et al. (2001). *A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. The International SNP Map Working Group.* Nature 409, 928-933.

This content is available online at http://www.ergito.com/main.jsp?bcs=MBIO.1.3.4

THE CONTENT OF THE GENOME

1.3.5 Why are genomes so large?

Key Terms

The **C-value** is the total amount of DNA in the genome (per haploid set of chromosomes).

The **C-value paradox** describes the lack of relationship between the DNA content (C-value) of an organism and its coding potential.

Key Concepts

- There is no good correlation between genome size and genetic complexity.
- There is an increase in the minimum genome size required to make organisms of increasing complexity.
- There are wide variations in the genome sizes of organisms within many phyla.

The total amount of DNA in the (haploid) genome is a characteristic of each living species known as its **C-value**. There is enormous variation in the range of C-values, from $<10^6$ bp for a mycoplasma to $>10^{11}$ bp for some plants and amphibians.

Figure 3.5 summarizes the range of C-values found in different evolutionary phyla. There is an increase in the minimum genome size found in each group as the complexity increases. But as absolute amounts of DNA increase in the higher eukaryotes, we see some wide variations in the genome sizes within some phyla.

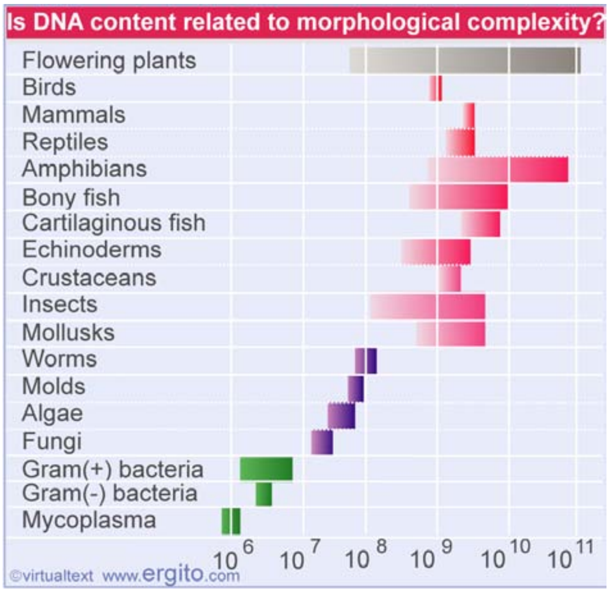


Figure 3.5 DNA content of the haploid genome is related to the morphological complexity of lower eukaryotes, but varies extensively among the higher eukaryotes. The range of DNA values within a phylum is indicated by the shaded area.

Plotting the *minimum* amount of DNA required for a member of each group suggests in **Figure 3.6** that an increase in genome size is required to make more complex prokaryotes and lower eukaryotes.

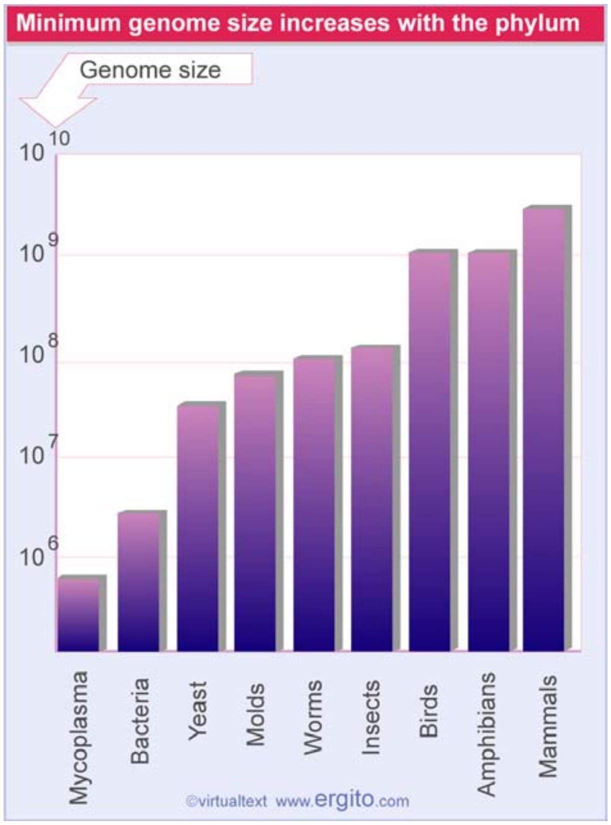
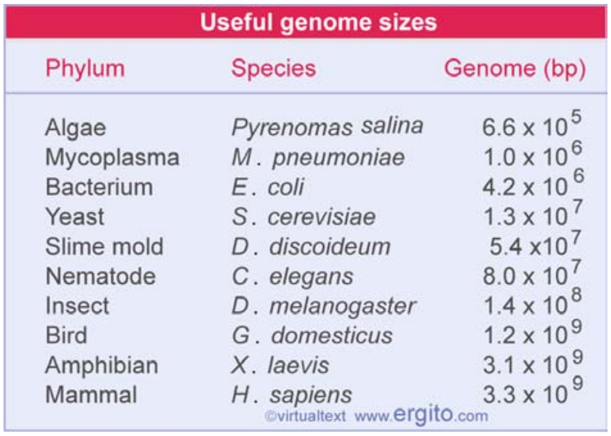


Figure 3.6 The minimum genome size found in each phylum increases from prokaryotes to mammals.

Mycoplasma are the smallest prokaryotes, and have genomes only ~3× the size of a large bacteriophage. Bacteria start at ~2 × 10^6 bp. Unicellular eukaryotes (whose life-styles may resemble the prokaryotic) get by with genomes that are also small, although larger than those of the bacteria. Being eukaryotic *per se* does not imply a vast increase in genome size; a yeast may have a genome size of ~1.3 × 10^7 bp, only about twice the size of the largest bacterial genomes.

A further twofold increase in genome size is adequate to support the slime mold *D. discoideum*, able to live in either unicellular or multicellular modes. Another increase in complexity is necessary to produce the first fully multicellular organisms; the nematode worm *C. elegans* has a DNA content of 8 × 10^7 bp.

We can also see the steady increase in genome size with complexity in the listing in **Figure 3.7** of some of the most commonly analyzed organisms. It is necessary to increase the genome size in order to make insects, birds or amphibians, and mammals. However, after this point there is no good relationship between genome size and morphological complexity of the organism.



Useful genome sizes

Phylum	Species	Genome (bp)
Algae	*Pyrenomas salina*	6.6×10^5
Mycoplasma	*M. pneumoniae*	1.0×10^6
Bacterium	*E. coli*	4.2×10^6
Yeast	*S. cerevisiae*	1.3×10^7
Slime mold	*D. discoideum*	5.4×10^7
Nematode	*C. elegans*	8.0×10^7
Insect	*D. melanogaster*	1.4×10^8
Bird	*G. domesticus*	1.2×10^9
Amphibian	*X. laevis*	3.1×10^9
Mammal	*H. sapiens*	3.3×10^9

©virtualtext www.ergito.com

Figure 3.7 The genome sizes of some common experimental animals.

We know that genes are much larger than the sequences needed to code for proteins, because exons (coding regions) may comprise only a small part of the total length of a gene). This explains why there is much more DNA than is needed to provide reading frames for all the proteins of the organism. Large parts of an interrupted gene may not be concerned with coding for protein. And there may also be significant lengths of DNA between genes. So it is not possible to deduce from the overall size of the genome anything about the number of genes.

The **C-value paradox** refers to the lack of correlation between genome size and genetic complexity (3040; 3039). There are some extremely curious variations in relative genome size. The toad *Xenopus* and man have genomes of essentially the same size. But we assume that man is more complex in terms of genetic development! And in some phyla there are extremely large variations in DNA content between organisms that do not vary much in complexity (see **Figure 3.5**). (This is especially marked in insects, amphibians, and plants, but does not occur in birds, reptiles, and mammals, which all show little variation within the group, with an ~2× range of genome sizes.) A cricket has a genome 11× the size of a fruit fly. In amphibians, the smallest genomes are $<10^9$ bp, while the largest are $\sim 10^{11}$ bp. There is unlikely to be a large difference in the number of genes needed to specify these amphibians. We do not understand why natural selection allows this variation and whether it has evolutionary consequences.

Reviews

3039. Gregory, T. R. (2001). *Coincidence, coevolution, or causation? DNA content, cell size, and the C-value enigma.* Biol. Rev. Camb. Philos. Soc. 76, 65-101.

3040. Gall, J. G. (1981). *Chromosome structure and the C-value paradox.* J. Cell Biol. 91, 3s-14s.

This content is available online at http://www.ergito.com/main.jsp?bcs=MBIO.1.3.5

THE CONTENT OF THE GENOME

1.3.6 Eukaryotic genomes contain both nonrepetitive and repetitive DNA sequences

Key Terms

Nonrepetitive DNA shows reassociation kinetics expected of unique sequences.

Repetitive DNA behaves in a reassociation reaction as though many (related or identical) sequences are present in a component, allowing any pair of complementary sequences to reassociate.

A **transposon (transposable element)** is a DNA sequence able to insert itself (or a copy of itself) at a new location in the genome, without having any sequence relationship with the target locus.

Selfish DNA describes sequences that do not contribute to the genotype of the organism but have self-perpetuation within the genome as their sole function.

Key Concepts

- The kinetics of DNA reassociation after a genome has been denatured distinguish sequences by their frequency of repetition in the genome.
- Genes are generally coded by sequences in nonrepetitive DNA.
- Larger genomes within a phylum do not contain more genes, but have large amounts of repetitive DNA.
- A large part of repetitive DNA may be made up of transposons.

The general nature of the eukaryotic genome can be assessed by the kinetics of reassociation of denatured DNA. This technique was used extensively before large scale DNA sequencing became possible (for review see *Molecular Biology Supplement 32.1 DNA reassociation kinetics*).

Reassociation kinetics identify two general types of genomic sequences (15; 16):

- **Nonrepetitive DNA** consists of sequences that are unique: there is only one copy in a haploid genome.
- **Repetitive DNA** describes sequences that are present in more than one copy in each genome.

Repetitive DNA is often divided into two general types:

- Moderately repetitive DNA consists of relatively short sequences that are repeated typically 10-1000× in the genome. The sequences are dispersed throughout the genome, and are responsible for the high degree of secondary structure formation in pre-mRNA, when (inverted) repeats in the introns pair to form duplex regions.

- Highly repetitive DNA consists of very short sequences (typically <100 bp) that are present many thousands of times in the genome, often organized as long tandem repeats (see *Molecular Biology 1.4.11 Satellite DNAs often lie in heterochromatin*). Neither class represents protein.

The proportion of the genome occupied by nonrepetitive DNA varies widely. **Figure 3.8** summarizes the genome organization of some representative organisms. Prokaryotes contain only nonrepetitive DNA. For lower eukaryotes, most of the DNA is nonrepetitive; <20% falls into one or more moderately repetitive components. In animal cells, up to half of the DNA often is occupied by moderately and highly repetitive components. In plants and amphibians, the moderately and highly repetitive components may account for up to 80% of the genome, so that the nonrepetitive DNA is reduced to a minority component.

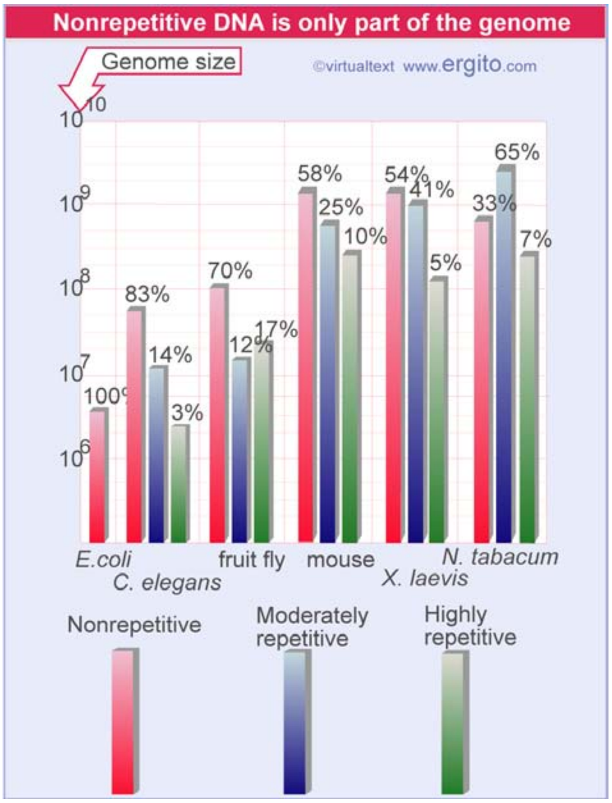


Figure 3.8 The proportions of different sequence components vary in eukaryotic genomes. The absolute content of nonrepetitive DNA increases with genome size, but reaches a plateau at ~2 × 10^9 bp.

A significant part of the moderately repetitive DNA consists of **transposons**, short sequences of DNA (~1 kb) that have the ability to move to new locations in the genome and/or to make additional copies of themselves (see *Molecular*

Biology 4.16 Transposons and *Molecular Biology 4.17 Retroviruses and retroposons*). In some higher eukaryotic genomes they may even occupy more than half of the genome (see *Molecular Biology 1.3.11 The human genome has fewer genes than expected*).

Transposons are sometimes viewed as fitting the concept of **selfish DNA**, defined as sequences that propagate themselves within a genome, without contributing to the development of the organism. Transposons may sponsor genome rearrangements, and these could confer selective advantages, but it is fair to say that we do not really understand why selective forces do not act against transposons becoming such a large proportion of the genome. Another term that is sometimes used to describe the apparent excess of DNA is *junk DNA*, meaning genomic sequences without any apparent function. Of course, it is likely that there is a balance in the genome between the generation of new sequences and the elimination of unwanted sequences, and some proportion of DNA that apparently lacks function may be in the process of being eliminated.

The length of the nonrepetitive DNA component tends to increase with overall genome size, as we proceed up to a total genome size ~3×10^9 (characteristic of mammals). Further increase in genome size, however, generally reflects an increase in the amount and proportion of the repetitive components, so that it is rare for an organism to have a nonrepetitive DNA component >2×10^9. The nonrepetitive DNA content of genomes therefore accords better with our sense of the relative complexity of the organism. *E. coli* has 4.2×10^6 bp, *C. elegans* increases an order of magnitude to 6.6×10^7 bp, *D. melanogaster* increases further to ~10^8 bp, and mammals increase another order of magnitude to ~2×10^9 bp.

What type of DNA corresponds to protein-coding genes? Reassociation kinetics typically show that mRNA is derived from nonrepetitive DNA. The amount of nonrepetitive DNA is therefore a better indication that the total DNA of the coding potential. (However, more detailed analysis based on genomic sequences shows that many exons have related sequences in other exons [see *Molecular Biology 1.2.5 Exon sequences are conserved but introns vary*]. Such exons evolve by a duplication to give copies that initially are identical, but which then diverge in sequence during evolution.)

Last updated on 2-29-2000

Reviews

15. Britten, R. J. and Davidson, E. H. (1971). *Repetitive and nonrepetitive DNA sequences and a speculation on the origins of evolutionary novelty*. Q. Rev. Biol. 46, 111-133.

16. Davidson, E. H. and Britten, R. J. (1973). *Organization, transcription, and regulation in the animal genome.* Q. Rev. Biol. 48, 565-613.

This content is available online at http://www.ergito.com/main.jsp?bcs=MBIO.1.3.6

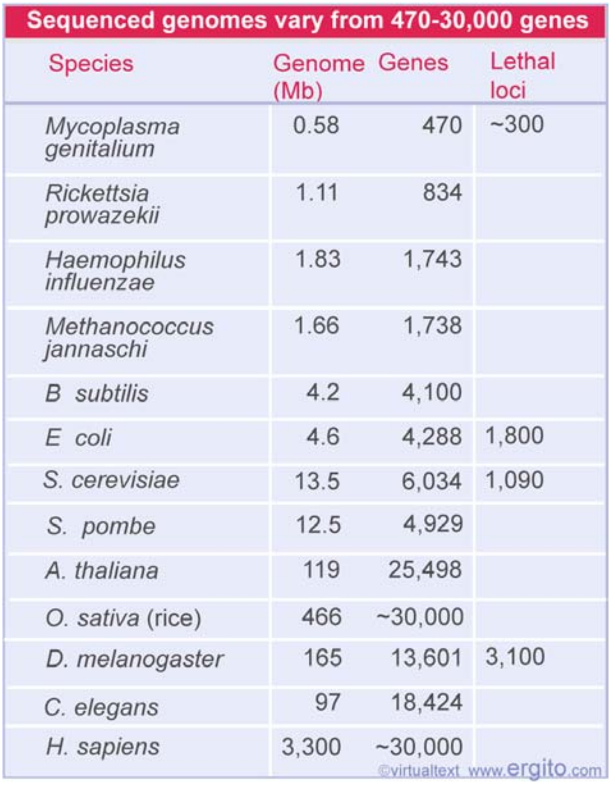
THE CONTENT OF THE GENOME

1.3.7 Bacterial gene numbers range over an order of magnitude

Key Concepts

- Genome sequences show that there are 500-1200 genes in parasitic bacteria, 1500-7500 genes in free-living bacteria, and 1500-2700 genes in archaea.

Large-scale efforts have now led to the sequencing of many genomes. A range is summarized in **Figure 3.9**. They extend from the 0.6×10^6 bp of a mycoplasma to the 3.3×10^9 bp of the human genome, and include several important experimental animals, including yeasts, the fruit fly, and a nematode worm. (Web sites with summaries of genome sequences are listed at the end of this section).



Sequenced genomes vary from 470-30,000 genes

Species	Genome (Mb)	Genes	Lethal loci
Mycoplasma genitalium	0.58	470	~300
Rickettsia prowazekii	1.11	834	
Haemophilus influenzae	1.83	1,743	
Methanococcus jannaschi	1.66	1,738	
B subtilis	4.2	4,100	
E coli	4.6	4,288	1,800
S. cerevisiae	13.5	6,034	1,090
S. pombe	12.5	4,929	
A. thaliana	119	25,498	
O. sativa (rice)	466	~30,000	
D. melanogaster	165	13,601	3,100
C. elegans	97	18,424	
H. sapiens	3,300	~30,000	

Figure 3.9 Genome sizes and gene numbers are known from complete sequences for several organisms. Lethal loci are estimated from genetic data.

Figure 3.10 summarizes the minimum number of genes found in each class of organism; of course, many species may have more than the minimum number required for their type.

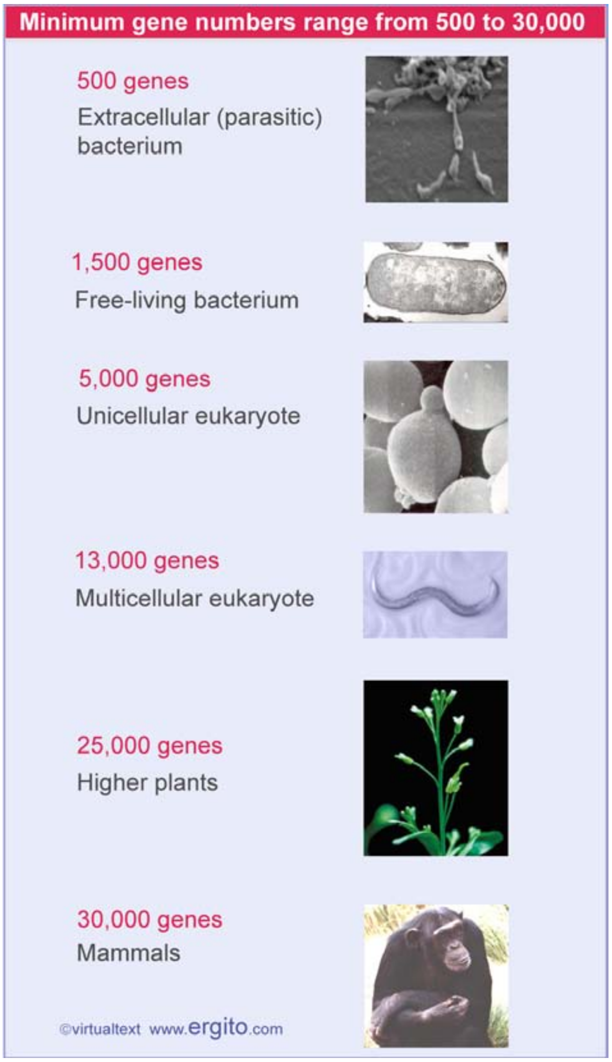


Figure 3.10 The minimum gene number required for any type of organism increases with its complexity. Photograph of mycoplasma kindly provided by A. Albay, K. Frantz, and K. Bott. Photograph of bacterium kindly provided by Jonathan King.

The sequences of the genomes of bacteria and archaea show that virtually all of the DNA (typically 85-90%) codes for RNA or protein. **Figure 3.11** shows that the range of genome sizes is about an order of magnitude, and that the genome size is proportional to the number of genes. The typical gene is about 1000 bp in length.

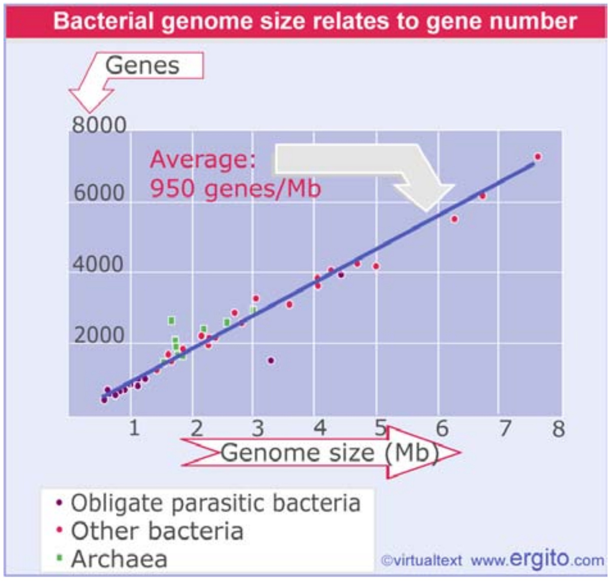


Figure 3.11 The number of genes in bacterial and archaeal genomes is proportional to genome size.

All of the bacteria with genome sizes below 1.5 Mb are obligate intracellular parasites – they live within a eukaryotic host that provides them with small molecules. Their genomes identify the minimum number of functions required to construct a cell. All classes of genes are reduced in number compared with bacteria with larger genomes, but the most significant reduction is in loci coding for enzymes concerned with metabolic functions (which are largely provided by the host cell) and with regulation of gene expression. *Mycoplasma genitalium* has the smallest genome, ~470 genes.

The archaea have biological properties that are intermediate between the prokaryotes and eukaryotes, but their genome sizes and gene numbers fall in the same range as bacteria. Their genome sizes vary from 1.5 - 3 Mb, corresponding to 1500 - 2700 genes. *M. jannaschii* is a methane-producing species that lives under high pressure and temperature. Its total gene number is similar to that of *H. influenzae*, but fewer of its genes can be identified on the basis of comparison with genes known in other organisms. Its apparatus for gene expression resembles eukaryotes more than prokaryotes, but its apparatus for cell division better resembles prokaryotes.

The archaea and the smallest free-living bacteria identify the minimum number of genes required to make a cell able to function independently in the environment. The smallest archaeal genome has ~1500 genes. The free-living bacterium with the smallest known genome is the thermophile *Aquifex aeolicus*, with 1.5 Mb and 1512 genes (2373). A "typical" gram-negative bacterium, *H. influenzae,* has 1,743 genes each of ~900 bp. So we can conclude that ~1500 genes are required to make a free-living organism.

Bacterial genome sizes extend over almost an order of magnitude to <8 Mb. The larger genomes have more genes. The bacteria with the largest genomes, *S. meliloti* and *M. loti*, are nitrogen-fixing bacteria that live on plant roots. Their genome sizes (~7 Mb) and total gene numbers (>6000) are similar to those of yeasts (2031).

The size of the genome of *E. coli* is in the middle of the range. The common laboratory strain has 4,288 genes, with an average length ~950 bp, and an average separation between genes of 118 bp (406). But there can be quite significant differences between strains. The known extremes of *E. coli* are from the smallest strain that has 4.6 Mb with 4249 genes to the largest strain that has 5.5 Mb bp with 5361 genes

We still do not know the functions of all the genes. In most of these genomes, ~60% of the genes can be identified on the basis of homology with known genes in other species. These genes fall approximately equally into classes whose products are concerned with metabolism, cell structure or transport of components, and gene expression and its regulation. In virtually every genome, >25% of the genes cannot be ascribed any function. Many of these genes can be found in related organisms, which implies that they have a conserved function.

There has been some emphasis on sequencing the genomes of pathogenic bacteria, given their medical importance. An important insight into the nature of pathogenicity has been provided by the demonstration that "pathogenicity islands" are a characteristic feature of their genomes (for review see 2491). These are large regions, ~10-200 kb, that are present in the genome of a pathogenic species, but absent from the genomes of nonpathogenic variants of the same or related species. Their G-C content often differs from that of the rest of the genome, and it is likely that they migrate between bacteria by a process of horizontal transfer. For example, the bacterium that causes anthrax (*B. anthracis*) has two large plasmids (extrachromosomal DNA), one of which has a pathogenicity island that includes the gene coding for the anthrax toxin.

Useful Websites

Direct access to genome sequences
http://www.ncbi.nlm.nih.gov/Genomes/index.html

Reviews

2491. Hacker, J. and Kaper, J. B. (2000). *Pathogenicity islands and the evolution of microbes.* Annu. Rev. Microbiol. 54, 641-679.

References

406. Blattner, F. R. et al. (1997). *The complete genome sequence of Escherichia coli K-12.* Science 277, 1453-1474.

2031. Galibert, F. et al. (2001). *The composite genome of the legume symbiont Sinorhizobium meliloti.* Science 293, 668-672.

2373. Deckert, G. et al. (1998). *The complete genome of the hyperthermophilic bacterium Aquifex aeolicus.* Nature 392, 353-358.

This content is available online at http://www.ergito.com/main.jsp?bcs=MBIO.1.3.7

THE CONTENT OF THE GENOME

1.3.8 Total gene number is known for several eukaryotes

Key Concepts

- There are 6000 genes in yeast, 18,500 in worm, 13,600 in fly, 25,000 in the small plant *Arabidopsis*, and probably 30,000 in mouse and <40,000 in Man.

As soon as we look at eukaryotic genomes, the relationship between genome size and gene number is lost. The genomes of unicellular eukaryotes fall in the same size range as the largest bacterial genomes. Higher eukaryotes have more genes, but the number does not correlate with genome size, as can be seen from **Figure 3.12**.

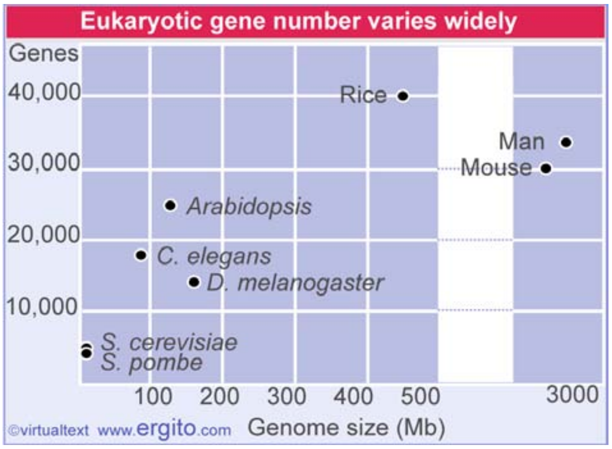


Figure 3.12 The number of genes in a eukaryote varies from 6000 to 40,000 but does not correlate with the genome size or the complexity of the organism.

The most extensive data for lower eukaryotes are available from the sequences of the genomes of the yeasts *S. cerevisiae* and *S. pombe*. **Figure 3.13** summarizes the most important features. The yeast genomes of 13.5 Mb and 12.5 Mb have ~6000 and ~5000 genes, respectively. The average open reading frame is ~1.4 kb, so that ~70% of the genome is occupied by coding regions. The major difference between them is that only 5% of *S. cerevisiae* genes have introns, compared to 43% in *S. pombe*. The density of genes is high; organization is generally similar, although the spaces between genes are a bit shorter in *S. cerevisiae*. About half of the genes identified by sequence were either known previously or related to known genes. The remainder are new, which gives some indication of the number of new types of genes that may be discovered (402, 403, 404, 2372).

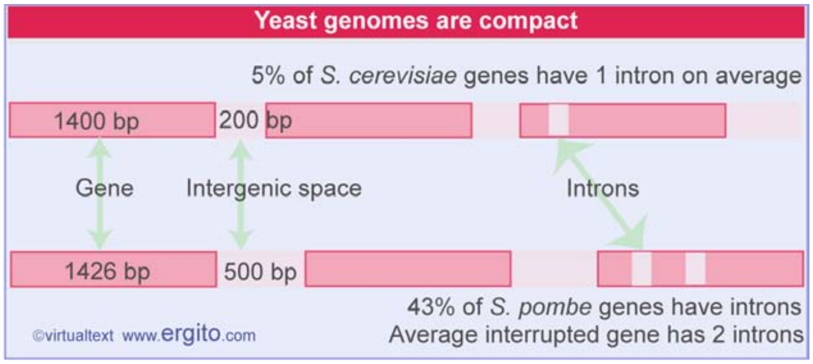

Yeast genomes are compact
5% of S. cerevisiae genes have 1 intron on average
1400 bp
200 bp
Gene
Intergenic space
Introns
1426 bp
500 bp
43% of S. pombe genes have introns
Average interrupted gene has 2 introns
©virtualtext www.ergito.com

Figure 3.13 The *S. cerevisiae* genome of 13.5 Mb has 6000 genes, almost all uninterrupted. The *S. pombe* genome of 12.5 Mb has 5000 genes, almost half having introns. Gene sizes and spacing are fairly similar.

The identification of long reading frames on the basis of sequence is quite accurate. However, ORFs coding for <100 amino acids cannot be identified solely by sequence because of the high occurrence of false positives. Analysis of gene expression suggests that ~300 of 600 such ORFs in *S. cerevisiae* are likely to be genuine genes.

A powerful way to validate gene structure is to compare sequences in closely related species # if a gene is active, it is likely to be conserved. Comparisons between the sequences of four closely related yeast species suggest that 503 of the genes originally identified in *S. cerevisiae* do not have counterparts in the other species, and therefore should be deleted from the catalog. This reduces the total gene number for *S. cerevisiae* to 5726 (3997).

The genome of *C. elegans* DNA varies between regions rich in genes and regions in which genes are more sparsely organized. The total sequence contains ~18,500 genes. Only ~42% of the genes have putative counterparts outside the Nematoda (405, 407).

Although the fly genome is larger than the worm genome, there are fewer genes (13,600) in *D. melanogaster* (949). The number of different transcripts is slightly larger (14,100) as the result of alternative splicing. We do not understand why the fly —a much more complex organism—has only 70% of the number of genes in the worm. This emphasizes forcefully the lack of an exact relationship between gene number and complexity of the organism.

The plant *Arabidopsis thaliana* has a genome size intermediate between the worm and the fly, but has a larger gene number (25,000) than either (1403). This again shows the lack of a clear relationship, and also emphasizes the special quality of plants, which may have more genes (due to ancestral duplications) than animal cells. A majority of the *Arabidopsis* genome is found in duplicated segments, suggesting that there was an ancient doubling of the genome (to give a tetraploid). Only 35% of *Arabidopsis* genes are present as single copies.

The genome of rice (*Oryza sativa*) is ~4 larger than *Arabidopsis*, but the number of genes is only ~50% larger, probably ~40,000 (2429, 2430). Repetitive DNA occupies 42-45% of the genome. More than 80% of the genes found in *Arabidopsis* are represented in rice. Of these common genes, ~8000 are found in *Arabidopsis* and rice but not in any of the bacterial or animal genomes that have been sequenced. These are probably the set of genes that code for plant-specific functions, such as photosynthesis.

From the fly genome, we can form an impression of how many genes are devoted to each type of function. **Figure 3.14** breaks down the functions into different categories. Among the genes that are identified, we find 2500 enzymes, ~750 transcription factors, ~700 transporters and ion channels, and ~700 proteins involved with signal transduction. But just over the half genes code for products of unknown function. ~20% of the proteins reside in membranes.

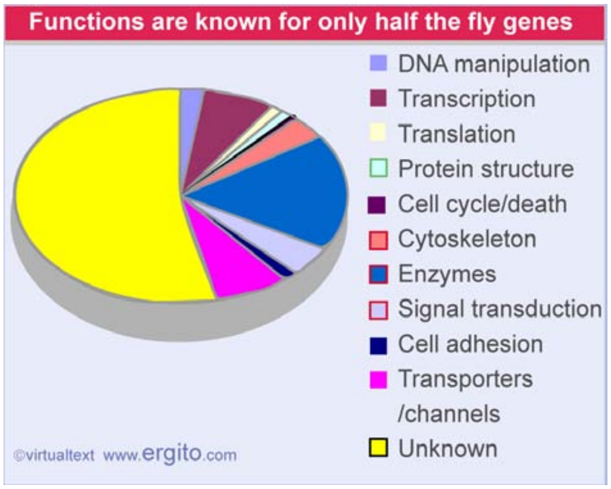


Figure 3.14 ~20% of *Drosophila* genes code for proteins concerned with maintaining or expressing genes, ~20% for enzymes, <10% for proteins concerned with the cell cycle or signal transduction. Half of the genes of *Drosophila* code for products of unknown function.

Protein size increases from prokaryotes and archaea to eukaryotes. The archaea *M. jannaschi* and bacterium *E. coli* have average protein lengths of 287 and 317 amino acids, respectively; whereas *S. cerevisiae* and *C. elegans* have average lengths of 484 and 442 amino acids, respectively. Large proteins (500 amino acids) are rare in bacteria, but comprise a significant component (~1/3) in eukaryotes. The increase in length is due to the addition of extra domains, with each domain typically constituting 100-300 amino acids. But the increase in protein size is responsible for only a very small part of the increase in genome size.

Another insight into gene number is obtained by counting the number of expressed genes. If we rely upon the estimates of the number of different mRNA species that can be counted in a cell, we would conclude that the average vertebrate cell expresses ~10,000-20,000 genes. The existence of significant overlaps between the messenger populations in different cell types would suggest that the total expressed gene number for the organism should be within a few fold of this. The estimate for the total human genome number of 30,000-40,000 (see *Molecular Biology 1.3.11 The human genome has fewer genes than expected*) would imply that a significant proportion of the total gene number is actually expressed in any given cell.

Eukaryotic genes are transcribed individually, each gene producing a monocistronic messenger. There is only one general exception to this rule; in the genome of *C. elegans*, ~15% of the genes are organized into polycistronic units (which is associated with the use of *trans*-splicing to allow expression of the downstream genes in these units; see *Molecular Biology 5.24.13 trans-splicing reactions use small RNAs*).

Last updated on 7-21-2003

References

402. Oliver, S. G. et al. (1992). *The complete DNA sequence of yeast chromosome III*. Nature 357, 38-46.

403. Dujon, B. et al. (1994). *Complete DNA sequence of yeast chromosome XI*. Nature 369, 371-378.

404. Johnston, M. et al. (1994). *Complete nucleotide sequence of S. cerevisiae chromosome VIII*. Science 265, 2077-2082.

405. Wilson, R. et al. (1994). *22 Mb of contiguous nucleotide sequence from chromosome III of C. elegans*. Nature 368, 32-38.

407. C. elegans sequencing consortium (1998). *Genome sequence of the nematode C. elegans: a platform for investigating biology*. Science 282, 2012-2022.

949. Adams, M. D. et al. (2000). *The genome sequence of D. melanogaster.* Science 287, 2185-2195.

1403. The Arabidopsis Genome Initiative. (2000). *Analysis of the genome sequence of the flowering plant Arabidopsis thaliana.* Nature 408, 796-815.

2372. Wood, V. et al. (2002). *The genome sequence of S pombe.* Nature 415, 871-880.

2429. Duffy, A., and Grof, P. (2001). *Psychiatric diagnoses in the context of genetic studies of bipolar disorder*. Bipolar Disord 3, 270-275.

2430. Goff, S. A. et al. (2002). *A draft sequence of the rice genome (Oryza sativa L. ssp. japonica).* Science 296, 92-114.

3997. Kellis, M., Patterson, N., Endrizzi, M., Birren, B., and Lander, E. S. (2003). *Sequencing and comparison of yeast species to identify genes and regulatory elements.* Nature 423, 241-254.

This content is available online at http://www.ergito.com/main.jsp?bcs=MBIO.1.3.8

THE CONTENT OF THE GENOME

1.3.9 How many different types of genes are there?

Key Terms

The **proteome** is the complete set of proteins that is expressed by the entire genome. Because some genes code for multiple proteins, the size of the proteome is greater than the number of genes. Sometimes the term is used to describe complement of proteins expressed by a cell at any one time.

Orthologs are corresponding proteins in two species as defined by sequence homologies.

Only some genes are unique; others belong to families where the other members are related (but not usually identical).The proportion of unique genes declines with genome size, and the proportion of genes in families increases.The minimum number of gene families required to code a bacterium is >1000, a yeast is >4000, and a higher eukaryote 11,000-14,000.

Because some genes are present in more than one copy or are related to one another, the number of different types of genes is less than the total number of genes. We can divide the total number of genes into sets that have related members, as defined by comparing their exons. (A family of related genes arises by duplication of an ancestral gene followed by accumulation of changes in sequence between the copies. Most often the members of a family are related but not identical.) The number of types of genes is calculated by adding the number of unique genes (where there is no other related gene at all) to the numbers of families that have 2 or more members.

Figure 3.15 compares the total number of genes with the number of distinct families in each of six genomes (950, 1403, 1439). In bacteria, most genes are unique, so the number of distinct families is close to the total gene number. The situation is different even in the lower eukaryote S. cerevisiae, where there is a significant proportion of repeated genes. The most striking effect is that the number of genes increases quite sharply in the higher eukaryotes, but the number of gene families does not change much.

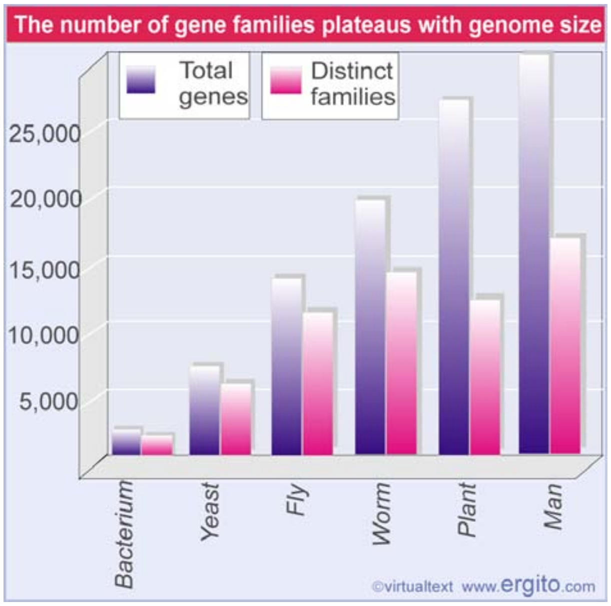
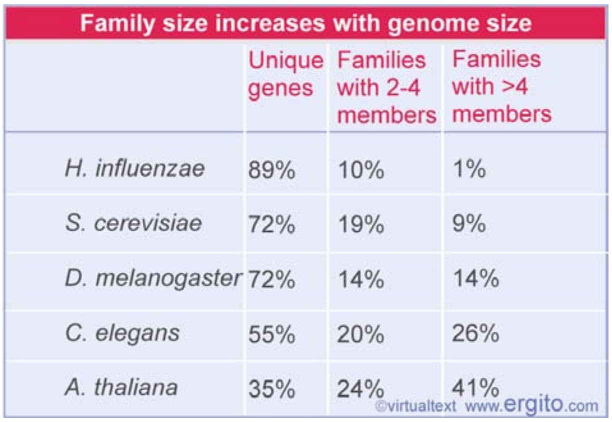


Figure 3.15 Because many genes are duplicated. the number of different gene families is much less than the total number of genes. The histogram compares the total number of genes with the number of distinct gene families.

Figure 3.16 shows that the proportion of unique genes drops sharply with genome size. When genes are present in families, the number of members in a family is small in bacteria and lower eukaryotes, but is large in higher eukaryotes. Much of the extra genome size of Arabidopsis is accounted for by families with >4 members (1403).


Family size increases with genome size

	Unique genes	Families with 2-4 members	Families with >4 members
H. influenzae	89%	10%	1%
S. cerevisiae	72%	19%	9%
D. melanogaster	72%	14%	14%
C. elegans	55%	20%	26%
A. thaliana	35%	24%	41%

©virtualtext www.ergito.com

Figure 3.16 The proportion of genes that are present in multiple copies increases with genome size in higher eukaryotes.

If every gene is expressed, the total number of genes will account the total number of proteins required to make the organism (the **proteome**). However, two effects mean that the proteome is different from the total gene number. Because genes are duplicated, some of them code for the same protein (although it may be expressed in a different time or place) and others may code for related proteins that again play the same role in different times or places. And because some genes can produce more than one protein by means of alternative splicing, the proteome can be larger than the number of genes.

What is the core proteome—the basic number of the different types of proteins in the organism? A minimum estimate is given by the number of gene families, ranging from 1400 in the bacterium, >4000 in the yeast, and a range of 11,000-14,000 for the fly and worm.

What is the distribution of the proteome among types of proteins? The 6000 proteins of the yeast proteome include 5000 soluble proteins and 1000 transmembrane proteins. About half of the proteins are cytoplasmic, a quarter are in the nucleolus, and the remainder are split between the mitochondrion and the ER/Golgi system (2496).

How many genes are common to all organisms (or to groups such as bacteria or higher eukaryotes) and how many are specific for the individual type of organism? **Figure 3.17** summarizes the comparison between yeast, worm, and fly (950). Genes that code for corresponding proteins in different organisms are called **orthologs**. Operationally, we usually reckon that two genes in different organisms can be considered to provide corresponding functions if their sequences are similar over >80% of the length. By this criterion, ~20% of the fly genes have orthologs in both yeast and the worm. These genes are probably required by all eukaryotes. The proportion increases to 30% when fly and worm are compared, probably representing the addition of gene functions that are common to multicellular eukaryotes. This still leaves a major proportion of genes as coding for proteins that are required specifically by either flies or worms, respectively.

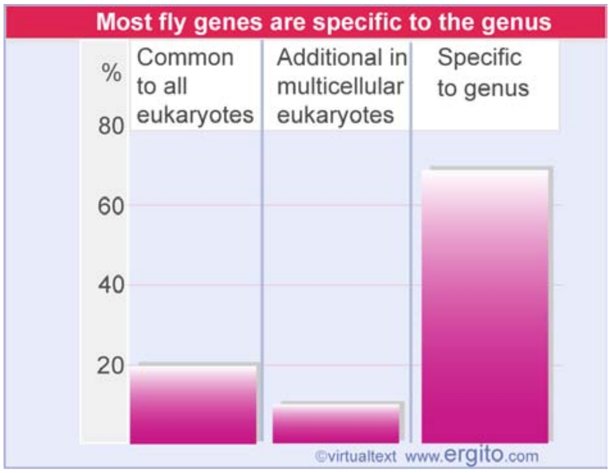


Figure 3.17 The fly genome can be divided into genes that are (probably) present in all eukaryotes, additional genes that are (probably) present in all multicellular eukaryotes, and genes that are more specific to subgroups of species that include flies.

The proteome can be deduced from the number and structures of genes, and can also be directly measured by analyzing the total protein content of a cell or organism. By such approaches, some proteins have been identified that were not suspected on the basis of genome analysis, and that have therefore led to the identification of new genes. Several methods are used for large scale analysis of proteins. Mass spectrometry can be used for separating and identifying proteins in a mixture obtained directly from cells or tissues (for review see 3755). Hybrid proteins bearing tags can be obtained by expression of cDNAs made by linking the sequences of open reading frames to appropriate expression vectors that incorporate the sequences for affinity tags. This allows array analysis to be used to analyze the products (for review see 3756). These methods also can be effective in comparing the proteins of two tissues, for example, a tissue from a normal individual and one from a patient with disease, to pinpoint the differences (for review see 3758).

Once we know the total number of proteins, we can ask how they interact. By definition, proteins in structural multiprotein assemblies must form stable interactions with one another. Proteins in signaling pathways interact with one another transiently. In both cases, such interactions can be detected in test systems where essentially a readout system magnifies the effect of the interaction. One popular such system is the two hybrid assay discussed in Independent domains bind DNA and activate transcription. Such assays cannot detect all interactions: for example, if one enzyme in a metabolic pathway releases a soluble metabolite that then interacts with the next enzyme, the proteins may not interact directly.

As a practical matter, assays of pairwise interactions can give us an indication of the minimum number of independent structures or pathways. An analysis of the ability of all 6000 (predicted) yeast proteins to interact in pairwise combinations shows that ~1000 proteins can bind to at least one other protein (951). Direct analyses of complex formation have identified 1440 different proteins in 232 multiprotein complexes (2260, 2261). This is the beginning of an analysis that will lead to definition of the number of functional assemblies or pathways (for review see 3757).

In addition to functional genes, there are also copies of genes that have become nonfunctional (identified as such by interruptions in their protein-coding sequences). These are called pseudogenes (see *Molecular Biology 1.4.6 Pseudogenes are dead ends of evolution*). The number of pseudogenes can be large. In the mouse and human genomes, the number of pseudogenes is ~10% of the number of (potentially) active genes (see *Molecular Biology 1.3.10 The conservation of genome organization helps to identify genes*).

Besides needing to know the density of genes to estimate the total gene number, we must also ask: is it important in itself? Are there structural constraints that make it necessary for genes to have a certain spacing, and does this contribute to the large size of eukaryotic genomes?

Last updated on 7-16-2003

Useful Websites

Yeast protein interactions
http://curatools.curagen.com/extpc/com.curagen.portal.servlet.Yeast

Reviews

3755. Aebersold, R. and Mann, M. (2003). *Mass spectrometry-based proteomics.* Nature 422, 198-207.

3756. Phizicky, E., Bastiaens, P. I., Zhu, H., Snyder, M., and Fields, S. (2003). *Protein analysis on a proteomic scale.* Nature 422, 208-215.

3757. Sali, A., Glaeser, R., Earnest, T., and Baumeister, W. (2003). *From words to literature in structural proteomics.* Nature 422, 216-225.

3758. Hanash, S. (2003). *Disease proteomics.* Nature 422, 226-232.

References

950. Rubin, G. M. et al. (2000). *Comparative genomics of the eukaryotes*. Science 287, 2204-2215.

951. Uetz, P. et al. (2000). *A comprehensive analysis of protein-protein interactions in S. cerevisiae.* Nature 403, 623-630.

1403. The Arabidopsis Genome Initiative. (2000). *Analysis of the genome sequence of the flowering plant Arabidopsis thaliana.* Nature 408, 796-815.

1439. Venter, J. C. et al. (2001). *The sequence of the human genome*. Science 291, 1304-1350.

2260. Gavin, A. C., et al. (2002). *Functional organization of the yeast proteome by systematic analysis of protein complexes.* Nature 415, 141-147.

2261. Ho, Y. et al. (2002). *Systematic identification of protein complexes in S. cerevisiae by mass spectrometry*. Nature 415, 180-183.

2496. Agarwal, S., Heyman, J. A., Matson, S., Heidtman, M., Piccirillo, S., Umansky, L., Drawid, A., Jansen, R., Liu, Y., Miller, P., Gerstein, M., Roeder, G. S., and Snyder, M. (2002). *Subcellular localization of the yeast proteome.* Genes Dev. 16, 707-719.

This content is available online at http://www.ergito.com/main.jsp?bcs=MBIO.1.3.9

THE CONTENT OF THE GENOME

1.3.10 The conservation of genome organization helps to identify genes

Key Terms

Synteny describes a relationship between chromosomal regions of different species where homologous genes occur in the same order.

Key Concepts

- Algorithms for identifying genes are not perfect and many corrections must be made to the initial data set.
- Pseudogenes must be distinguished from active genes.
- Syntenic relationships are extensive between mouse and human genomes, and most active genes are in a syntenic region.

Once we have assembled the sequence of a genome, we still have to identify the genes within it. Coding sequences represent a very small fraction. Exons can be identified as uninterrupted open reading frames flanked by appropriate sequences. What criteria need to be satisfied to identify an active gene from a series of exons?

Figure 3.18 shows that an active gene should consist of a series of exons where the first exon immediately follows a promoter, the internal exons are flanked by appropriate splicing junctions, the last exon is followed by 3′ processing signals, and a single open reading frame starting with an initiation codon and ending with a termination codon can be deduced by joining the exons together. Internal exons can be identified as open reading frames flanked by splicing junctions. In the simplest cases, the first and last exons contain the start and end of the coding region, respectively, (as well as the 5' and 3' untranslated regions), but in more complex cases the first or last exons may have only untranslated regions, and may therefore be more difficult to identify.

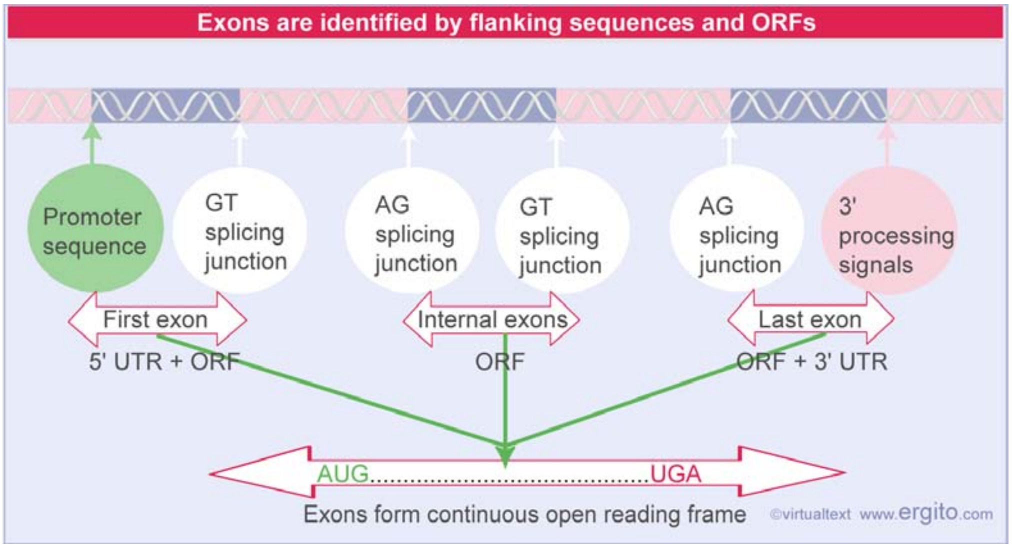


Figure 3.18 Exons of protein-coding genes are identified as coding sequences flanked by appropriate signals (with untranslated regions at both ends). The series of exons must generate an open reading frame with appropriate initiation and termination codons.

The algorithms that are used to connect exons are not completely effective when the genome is very large and the exons may be separated by very large distances. For example, the initial analysis of the human genome mapped 170,000 exons into 32,000 genes. This is unlikely to be correct, because it gives an average of 5.3 exons per gene, whereas the average of individual genes that have been fully characterized is 10.2. Either we have missed many exons, or they should be connected differently into a smaller number of genes in the whole genome sequence.

Even when the organization of a gene is correctly identified, there is the problem of distinguishing active genes from pseudogenes. Many pseudogenes can be recognized by obvious defects in the form of multiple mutations that create an inactive coding sequence. However, pseudogenes that have arisen more recently, and which have not accumulated so many mutations, may be more difficult to recognize. In an extreme example, the mouse has only one active *Gapdh* gene (coding for glyceraldehyde phosphate dehydrogenase), but has ~400 pseudogenes. However, >100 of these pseudogenes initially appeared to be active in the mouse genome sequence. Individual examination was necessary to exclude them from the list of active genes.

Confidence that a gene is active can be increased by comparing regions of the genomes of different species. There has been extensive overall reorganization of sequences between the mouse and human genomes, as seen in the simple fact that there are 23 chromosomes in the human haploid genome and 20 chromosomes in the mouse haploid genome. However, at the local level, the order of genes is generally the same: when pairs of human and mouse homologues are compared, the genes located on either side also tend to be homologues. This relationship is called **synteny**.

Figure 3.19 shows the relationship between mouse chromosome 1 and the human chromosomal set (3203). We can recognize 21 segments in this mouse chromosome that have syntenic counterparts in human chromosomes. The extent of reshuffling that has occurred between the genomes is shown by the fact that the segments are spread among 6 different human chromosome. The same types of relationships are found in all mouse chromosomes, except for the X chromosome, which is syntenic only with the human X chromosome. This is explained by the fact that the X is a special case, subject to dosage compensation to adjust for the difference between males (one copy) and females (two copies) (see *Molecular Biology 5.23.17 X chromosomes undergo global changes*). This may apply selective pressure against the translocation of genes to and from the X chromosome.

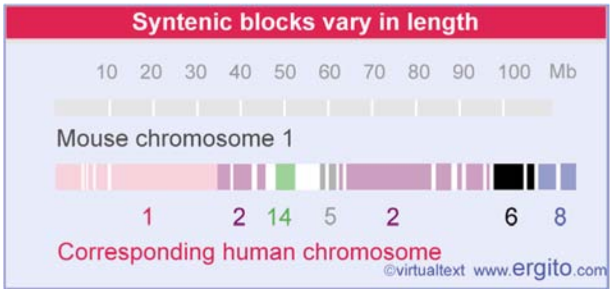

Syntenic blocks vary in length
10
20
30
40
50
60
70
80
90
100
Mb
Mouse chromosome 1
1
2
14
5
2
6
8
Corresponding human chromosome
©virtualtext www.ergito.com

Figure 3.19 Mouse chromosome 1 has 21 segments of 1 - 25 Mb that are syntenic with regions corresponding to parts of 6 human chromosomes.

Comparison of the mouse and human genome sequences shows that >90% of each genome lies in syntenic blocks that range widely in size (from 300 kb to 65 Mb). There is a total of 342 syntenic segments, with an average length of 7 Mb (0.3% of the genome) (3203). 99% of mouse genes have a homologue in the human genome; and for 96% that homologue is in a syntenic region.

Comparing the genomes provides interesting information about the evolution of species. The number of gene families in the mouse and human genomes is the same, and a major difference between the species is the differential expansion of particular families in one of the genomes. This is especially noticeable in genes that affect phenotypic features that are unique to the species. Of 25 families where the size has been expanded in mouse, 14 contain genes specifically involved in rodent reproduction, and 5 contain genes specific to the immune system.

A validation of the importance of syntenic blocks comes from pairwise comparisons of the genes within them. Looking for likely pseudogenes on the basis of sequence comparisons, a gene that is not in a syntenic location (that is, its context is different in the two species) is twice as likely to be a pseudogene. Put another way, translocation away from the original locus tends to be associated with the creation of pseudogenes. The lack of a related gene in a syntenic position is therefore grounds for suspecting that an apparent gene may really be a pseudogene. Overall, >10% of the genes that are initially identified by analysis of the genome are likely to turn out to be pseudogenes.

As a general rule, comparisons between genomes add significantly to the effectiveness of gene prediction. When sequence features indicating active genes are conserved, for example, between Man and mouse, there is an increased probability that they identify active homologues.

Identifying genes coding for RNA is more difficult, because we cannot use the criterion of the open reading frame. It is true here also that comparative genome analysis increased the rigor of the analysis. For example, analysis of either the human or mouse genome alone identifies ~500 genes coding for tRNA in each case, but comparison of features suggests that <350 of these genes are in fact active in each genome.

Last updated on 12-20-2002

Useful Websites

Mouse Genome Sequencing Consortium (Sanger Centre, Washington University, Whitehead Institute.)
Public assembled genome sequence from the C57BL/6J strain, and primary gene annotation
http://www.ensembl.org/Mus_musculus/resources.html

References

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

This content is available online at http://www.ergito.com/main.jsp?bcs=MBIO.1.3.10

THE CONTENT OF THE GENOME

1.3.11 The human genome has fewer genes than expected

Key Concepts

- Only 1% of the human genome consists of coding regions.
- The exons comprise ~5% of each gene, so genes (exons plus introns) comprise ~25% of the genome.
- The human genome has 30,000-40,000 genes.
- ~60% of human genes are alternatively spliced.
- Up to 80% of the alternative splices change protein sequence, so the proteome has ~50,000-60,000 members.

The human genome was the first vertebrate genome to be sequenced (1439, 1440). This massive task has revealed a wealth of information about the genetic makeup of our species, and about the evolution of the genome in general. (Methods used for genome sequencing are reviewed in *Molecular Biology Supplement 32.12 Genome mapping*.) Our understanding is deepened further by the ability to compare the human genome sequence with the more recently sequenced mouse genome (3203).

Mammal and rodent genomes generally fall into a narrow size range, ~ 3×10^9 bp (see *Molecular Biology 1.3.5 Why are genomes so large?*). The mouse genome is ~14% smaller than the human genome, probably because it has had a higher rate of deletion. The genomes contain similar gene families and genes, with most genes having an ortholog in the other genome, but with differences in the number of members of a family, especially in those cases where the functions are specific to the species (see *Molecular Biology 1.3.10 The conservation of genome organization helps to identify genes*). The estimate of 30,000 genes for the mouse genome is at the lower end of the range of estimates for the human genome. **Figure 3.20** plots the distribution of the mouse genes. The 30,000 protein-coding genes are accompanied by ~4000 pseudogenes. There are ~800 genes representing RNAs that do not code for proteins; these are generally small (aside from the rRNAs). Almost half of these genes code for tRNAs, for which a large number of pseudogenes also have been identified.

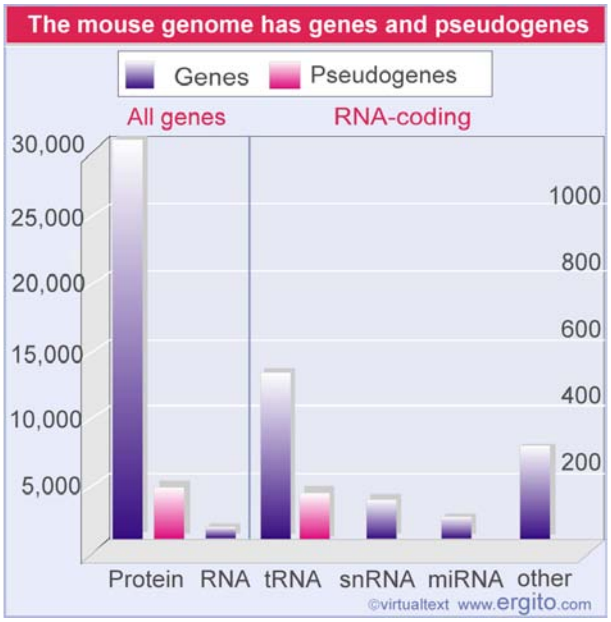


Figure 3.20 The mouse genome has ~30,000 protein-coding genes, which have ~4000 pseudogenes. There are ~800 RNA-coding genes. The data for RNA-coding genes are replotted on the right, at an expanded scale to show that there are ~350 tRNA genes and 150 pseudogenes, and ~450 other noncoding RNA genes, including snRNAs and miRNAs.

The human (haploid) genome contains 22 autosomes plus the X or Y. The chromosomes range in size from 45-279 Mb of DNA, making a total genome content of 3,286 Mb (~3.3 × 10^9 bp). On the basis of chromosome structure, the overall genome can be divided into regions of euchromatin (potentially containing active genes) and heterochromatin (see *Molecular Biology 5.19.7 Chromatin is divided into euchromatin and heterochromatin*). The euchromatin comprises the majority of the genome, ~2.9 × 10^9 bp. The identified genome sequence represents ~90% of the euchromatin. In addition to providing information on the genetic content of the genome, the sequence also identifies features that may be of structural importance (see *Molecular Biology 5.19.8 Chromosomes have banding patterns*).

Figure 3.21 shows that a tiny proportion (~1%) of the human genome is accounted for by the exons that actually code for proteins. The introns that constitute the remaining sequences in the genes bring the total of DNA concerned with producing proteins to ~25%. As shown in **Figure 3.22**, the average human gene is 27 kb long, with 9 exons that include a total coding sequence of 1,340 bp. The average coding sequence is therefore only 5% of the length of the gene.

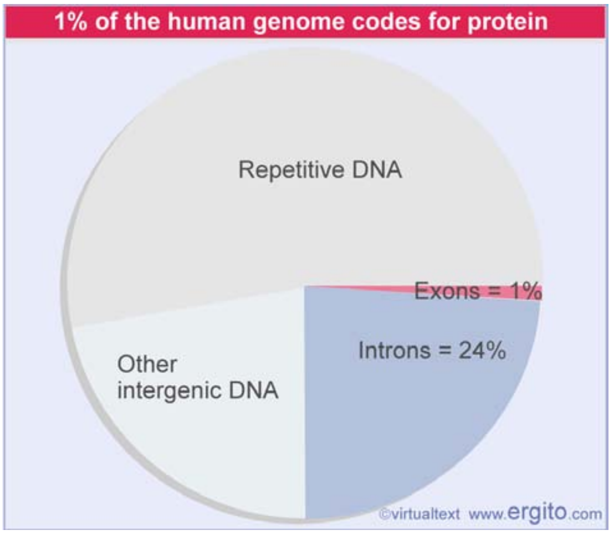


Figure 3.21 Genes occupy 25% of the human genome, but protein-coding sequences are only a tiny part of this fraction.

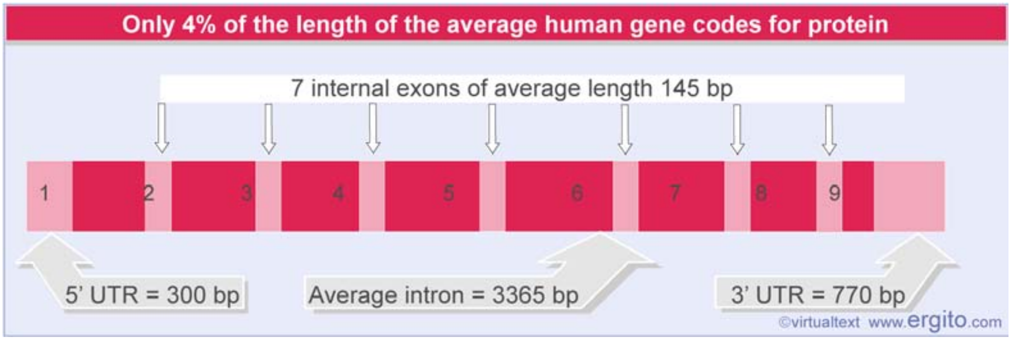

Only 4% of the length of the average human gene codes for protein
7 internal exons of average length 145 bp
1
2
3
4
5
6
7
8
9
5' UTR = 300 bp
Average intron = 3365 bp
3' UTR = 770 bp
©virtualtext www.ergito.com

Figure 3.22 The average human gene is 27 kb long and has 9 exons, usually comprising two longer exons at each end and 7 internal exons. The UTRs in the terminal exons are the untranslated (noncoding) regions at each end of the gene. (This is based on the average. Because some genes are extremely long, the median length is 14 kb with 7 exons.)

Based on comparisons with other species and with known protein-coding genes, there are ~24,000 clearly identifiable genes. Sequence analysis identifies ~12,000 more potential genes. Two independent analyses have produced estimates of ~30,000 and ~40,000 genes, respectively (1439, 1440). One measure of the accuracy of the analyses is whether they identify the same genes. The surprising answer is that the overlap between the two sets of genes is only ~50%, as summarized in **Figure 3.23** (1959). An earlier analysis of the human gene set based on RNA transcripts had identified ~11,000 genes, almost all of which are present in both the large human gene sets, and which account for the major part of the overlap between them. So there is no question about the authenticity of half of each human gene set, but we have yet to establish the relationship between the other half of each set. The discrepancies illustrate the pitfalls of large scale sequence analysis! As the sequence is analyzed further (and as other genomes are sequenced with which it can be compared), the number of valid genes seems to decline, and is now generally thought to be ~30,000.

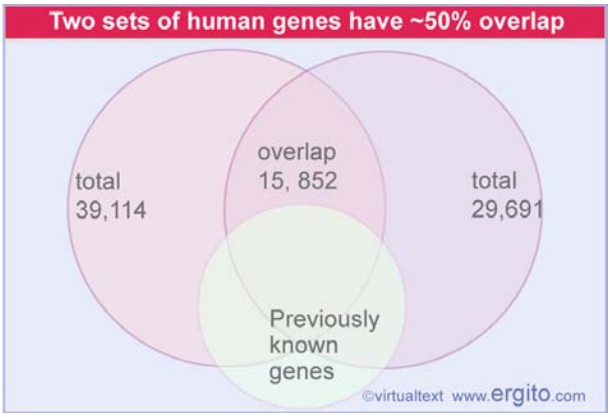

Two sets of human genes have ~50% overlap
total
39,114
overlap
15, 852
total
29,691
Previously
known
genes
©virtualtext www.ergito.com

Figure 3.23 The two sets of genes identified in the human genome overlap only partially, as shown in the two large upper circles. However, they include almost all previously known genes, as shown by the overlap with the smaller, lower circle (see 1959).

By any measure, the total human gene number is much less than we had expected —most previous estimates had been ~100,000. It shows a relatively small increase over flies and worms (13,600 and 18,500, respectively), not to mention the plant *Arabidopsis* (25,000) (see **Figure 3.9**). However, we should not be particularly surprised by the notion that it does not take a great number of additional genes to make a more complex organism. The difference in DNA sequences between man and chimpanzee is extremely small (there is >99% similarity), so it is clear that the functions and interactions between a similar set of genes can produce very different results. The functions of specific groups of genes may be especially important, because detailed comparisons of orthologous genes in man and chimpanzee suggest that there has been accelerated evolution of certain classes of genes, including some involved in early development, olfaction, hearing —all functions that are relatively specific for the species (4514).

The number of genes is less than the number of potential proteins because of alternative splicing. The extent of alternative splicing is greater in Man than in fly or worms; it may affect as many as 60% of the genes, so the increase in size of the human proteome relative to the other eukaryotes may be larger than the increase in the number of genes. A sample of genes from two chromosomes suggests that the proportion of the alternative splices that actually result in changes in the protein sequence may be as high as 80%. This could increase the size of the proteome to 50,000-60,000 members.

In terms of the diversity of the number of gene families, however, the discrepancy between Man and the other eukaryotes may not be so great. Many of the human genes belong to families. An analysis of ~25,000 genes identified 3500 unique genes and 10,300 gene pairs. As can be seen from **Figure 3.15**, this extrapolates to a number of gene families only slightly larger than worm or fly.

Last updated on December 23, 2003

References

1439. Venter, J. C. et al. (2001). *The sequence of the human genome*. Science 291, 1304-1350.

1440. International Human Genome Sequencing Consortium. (2001). *Initial sequencing and analysis of the human genome.* Nature 409, 860-921.

1959. Hogenesch, J. B., Ching, K. A., Batalov, S., Su, A. I., Walker, J. R., Zhou, Y., Kay, S. A., Schultz, P. G., and Cooke, M. P. (2001). *A comparison of the Celera and Ensembl predicted gene sets reveals little overlap in novel genes.* Cell 106, 413-415.

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

4514. Clark, A. G. et al. (2003). *Inferring nonneutral evolution from human-chimp-mouse orthologous gene trios.* Science 302, 1960-1963.

This content is available online at http://www.ergito.com/main.jsp?bcs=MBIO.1.3.11

THE CONTENT OF THE GENOME

1.3.12 How are genes and other sequences distributed in the genome?

Key Concepts

- Repeated sequences (present in more than one copy) account for >50% of the human genome.
- The great bulk of repeated sequences consist of copies of nonfunctional transposons.
- There are many duplications of large chromosome regions.

Are genes uniformly distributed in the genome? Some chromosomes are relatively poor in genes, and have >25% of their sequences as "deserts" – regions longer than 500 kb where there are no genes. Even the most gene-rich chromosomes have >10% of their sequences as deserts. So overall ~20% of the human genome consists of deserts that have no genes.

Repetitive sequences account for >50% of the human genome, as seen in **Figure 3.24**. The repetitive sequences fall into five classes:

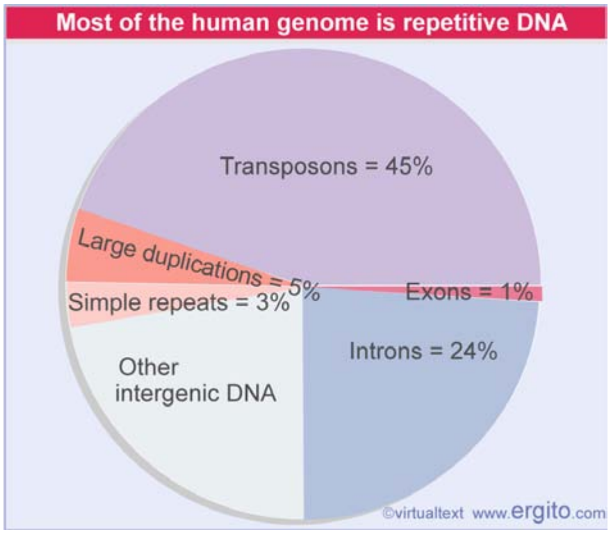


Figure 3.24 The largest component of the human genome consists of transposons. Other repetitive sequences include large duplications and simple repeats.

- Transposons (either active or inactive) account for the vast majority (45% of the genome). All transposons are found in multiple copies.
- Processed pseudogenes (~3000 in all, account for ~0.1% of total DNA). (These

are sequences that arise by insertion of a copy of an mRNA sequence into the genome; see *Molecular Biology 1.4.6 Pseudogenes are dead ends of evolution*).

- Simple sequence repeats (highly repetitive DNA such as $(CA)_n$ account for ~3%).
- Segmental duplications (blocks of 10-300 kb that have been duplicated into a new region) account for ~5%. Only a minority of these duplications are found on the same chromosome; in the other cases, the duplicates are on different chromosomes.
- Tandem repeats form blocks of one type of sequence (especially found at centromeres and telomeres).

The sequence of the human genome emphasizes the importance of transposons. (Transposons have the capacity to replicate themselves and insert into new locations. They may function exclusively as DNA elements [see *Molecular Biology 4.16 Transposons*] or may have an active form that is RNA [see *Molecular Biology 4.17 Retroviruses and retroposons*]. Their distribution in the human genome is summarized in **Figure 17.18**.) Most of the transposons in the human genome are nonfunctional; very few are currently active. However, the high proportion of the genome occupied by these elements indicates that they have played an active role in shaping the genome. One interesting feature is that some present genes originated as transposons, and evolved into their present condition after losing the ability to transpose. Almost 50 genes appear to have originated like this.

Segmental duplication at its simplest involves the tandem duplication of some region within a chromosome (typically because of an aberrant recombination event at meiosis; see *Molecular Biology 1.4.7 Unequal crossing-over rearranges gene clusters*). In many cases, however, the duplicated regions are on different chromosomes, implying that either there was originally a tandem duplication followed by a translocation of one copy to a new site, or that the duplication arose by some different mechanism altogether. The extreme case of a segmental duplication is when a whole genome is duplicated, in which case the diploid genome initially becomes tetraploid. As the duplicated copies develop differences from one another, the genome may gradually become effectively a diploid again, although homologies between the diverged copies leave evidence of the event. This is especially common in plant genomes. The present state of analysis of the human genome identifies many individual duplicated regions, but does not indicate whether there was a whole genome duplication in the vertebrate lineage.

Last updated on 2-16-2001

This content is available online at http://www.ergito.com/main.jsp?bcs=MBIO.1.3.12

THE CONTENT OF THE GENOME

1.3.13 The Y chromosome has several male-specific genes

Key Concepts

- The Y chromosome has ~60 genes that are expressed specifically in testis.
- The male-specific genes are present in multiple copies in repeated chromosomal segments.
- Gene conversion between multiple copies allows the active genes to be maintained during evolution.

The sequence of the human genome has significantly extended our understanding of the role of the sex chromosomes. It is generally thought that the X and Y chromosomes have descended from a common (very ancient) autosome. Their development has involved a process in which the X chromosome has retained most of the original genes, whereas the Y chromosome has lost most of them.

The X chromosome behaves like the autosomes insofar as females have two copies and recombination can take place between them. The density of genes on the X chromosome is comparable to the density of genes on other chromosomes.

The Y chromosome is much smaller than the X chromosome and has many fewer genes. Its unique role results from the fact that only males have the Y chromosome, and there is only one copy, so Y-linked loci are effectively haploid, instead of diploid like all other human genes.

For many years, the Y chromosome was thought to carry almost no genes except for one (or more) sex-determining genes that determine maleness. The vast majority of the Y chromosome (>95% of its sequence) does not undergo crossing-over with the X chromosome, which led to the view that it could not contain active genes, because there would be no means to prevent the accumulation of deleterious mutations. This region is flanked by short pseudoautosomal regions that exchange frequently with the X chromosome during male meiosis. It was originally called the nonrecombining region, but now has been renamed as the male-specific region.

Detailed sequencing of the Y chromosome shows that the male-specific region contains three types of regions, as illustrated in **Figure 3.25** (4088):

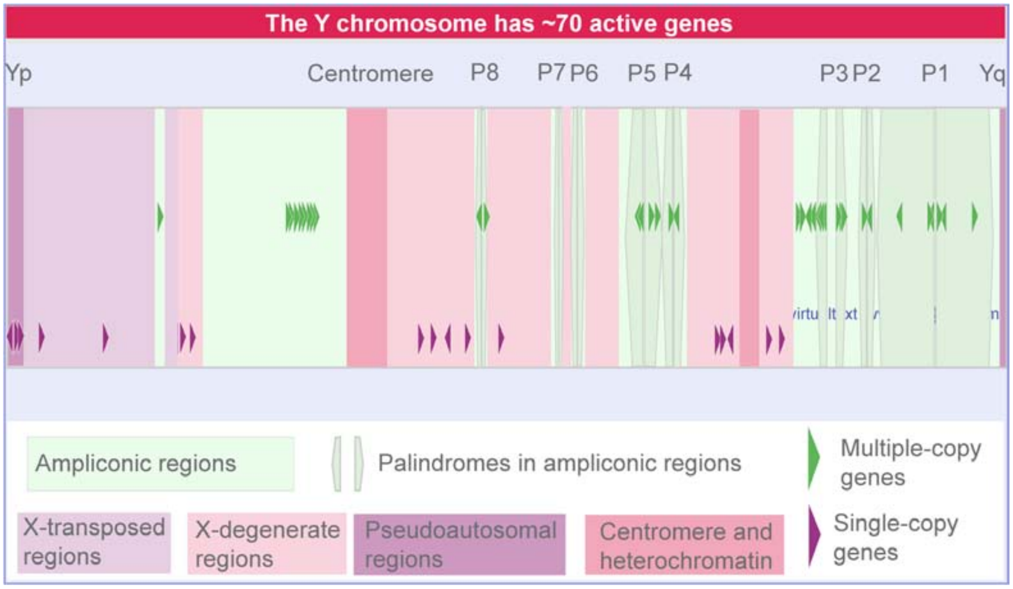


Figure 3.25 The Y chromosome consists of X-transposed regions, X-degenerate regions, and amplicons. The X-transposed X-degenerate regions have 2 and 14 single-copy genes, respectively. The amplicons have eight large palindromes (P1-P8), which contain 9 gene families. Each family contains at least two copies.

- The *X-transposed sequences* consist of a total of 3.4 Mb comprising some large blocks resulting from a transposition from band q21 in the X chromosome about 3-4 million years ago. This is specific to the human lineage. These sequences do not recombine with the X chromosome and have become largely inactive. They now contain only two active genes.

- The *X-degenerate segments* of the Y are sequences that have a common origin with the X chromosome (going back to the common autosome from which both X and Y have descended) and contain genes or pseudogenes related to X-linked genes. There are 14 active genes and 13 pseudogenes. The active genes have in a sense so far defied the trend for genes to be eliminated from chromosomal regions that cannot recombine at meiosis.

- The *ampliconic segments* have a total length of 10.2 Mb and are internally repeated on the Y chromosome. There are 8 large palindromic blocks. They include 9 protein-coding gene families, with copy numbers per family ranging from 2-35. The name "amplicon" reflects the fact that the sequences have been internally amplified on the Y chromosome.

Totaling the genes in these three regions, the Y chromosome contains many more genes than had been expected. There are 156 transcription units, of which half represent protein-coding genes, and half represent pseudogenes.

The presence of the active genes is explained by the fact that the existence of closely related genes copies in the ampliconic segments allows gene conversion between multiple copies of a gene to be used to regenerate active copies. The most common needs for multiple copies of a gene are quantitative (to provide more protein product) or qualitative (to code for proteins with slightly different properties or that are expressed in different times or places), but in this case the essential function is evolutionary. In effect, the existence of multiple copies allows recombination within the Y chromosome itself to substitute for the evolutionary diversity that is usually by provided by recombination between allelic chromosomes.

Most of the protein-coding genes in the ampliconic segments are expressed specifically in testis, and are likely to be involved in male development. If there are ~60 such genes out of a total human gene set of ~30,000, then the genetic difference between man and woman is ~0.2%.

References

4088. Skaletsky, H. et al. (2003). *The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes.* Nature 423, 825-837.

This content is available online at http://www.ergito.com/main.jsp?bcs=MBIO.1.3.13

THE CONTENT OF THE GENOME

1.3.14 More complex species evolve by adding new gene functions

Key Concepts

- Comparisons of different genomes show a steady increase in gene number as additional genes are added to make eukaryotes, make multicellular organisms, make animals, and make vertebrates.
- Most of the genes that are unique to vertebrates are concerned with the immune or nervous systems.

Comparison of the human genome sequence with sequences found in other species is revealing about the process of evolution. **Figure 3.26** analyzes human genes according to the breadth of their distribution in Nature. Starting with the most generally distributed (top right corner of the figure), 21% of genes are common to eukaryotes and prokaryotes. These tend to code for proteins that are essential for all living forms – typically basic metabolism, replication, transcription, and translation. Moving clockwise, another 32% of genes are added in eukaryotes in general – for example, they may be found in yeast. These tend to code for proteins involved in functions that are general to eukaryotic cells but not to bacteria – for example, they may be concerned with specifying organelles or cytoskeletal components. Another 24% of genes are needed to specify animals. These include genes necessary for multicellularity and for development of different tissue types. And 22% of genes are unique to vertebrates. These mostly code for proteins of the immune and nervous systems; they code for very few enzymes, consistent with the idea that enzymes have ancient origins, and that metabolic pathways originated early in evolution. We see, therefore, that the progression from bacteria to vertebrates requires addition of groups of genes representing the necessary new functions at each stage.

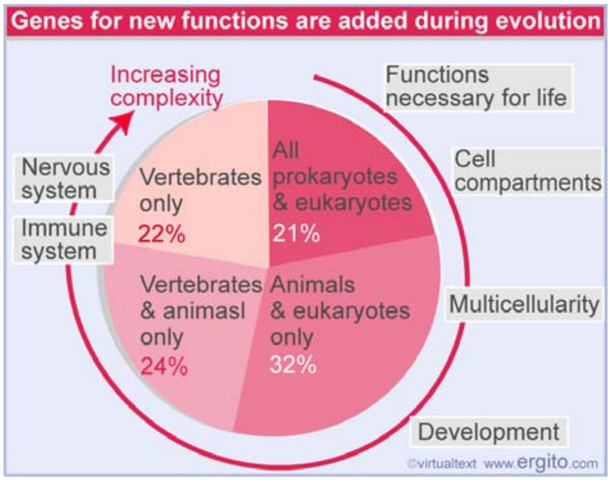


Figure 3.26 Human genes can be classified according to how widely their homologues are distributed in other species.

One way to define commonly needed proteins is to identify the proteins present in all proteomes. Comparing the human proteome in more detail with the proteomes of other organisms, 46% of the yeast proteome, 43% of the worm proteome, and 61% of the fly proteome is represented in the human proteome. A key group of ~1300 proteins is present in all four proteomes. The common proteins are basic housekeeping proteins required for essential functions, falling into the types summarized in **Figure 3.27**. The main functions are concerned with transcription and translation (35%), metabolism (22%), transport (12%), DNA replication and modification (10%), protein folding and degradation (8%), and cellular processes (6%).

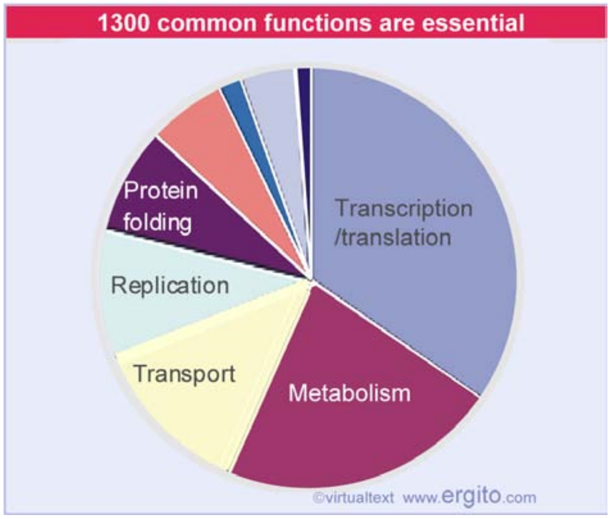


Figure 3.27 Common eukaryotic proteins are concerned with essential cellular functions.

One of the striking features of the human proteome is that it has many new proteins compared with other eukaryotes, but it has relatively few new protein domains. Most protein domains appear to be common to the animal kingdom. However, there are many new protein architectures, defined as new combinations of domains. **Figure 3.28** shows that the greatest increase occurs in transmembrane and extracellular proteins. In yeast, the vast majority of architectures are concerned with intracellular proteins. About twice as many intracellular architectures are found in fly (or worm), but there is a very striking increase in transmembrane and extracellular proteins, as might be expected from the addition of functions required for the interactions between the cells of a multicellular organism. The increase in intracellular architectures required to make a vertebrate (Man) is relatively small, but there is again a large increase in transmembrane and extracellular architectures.

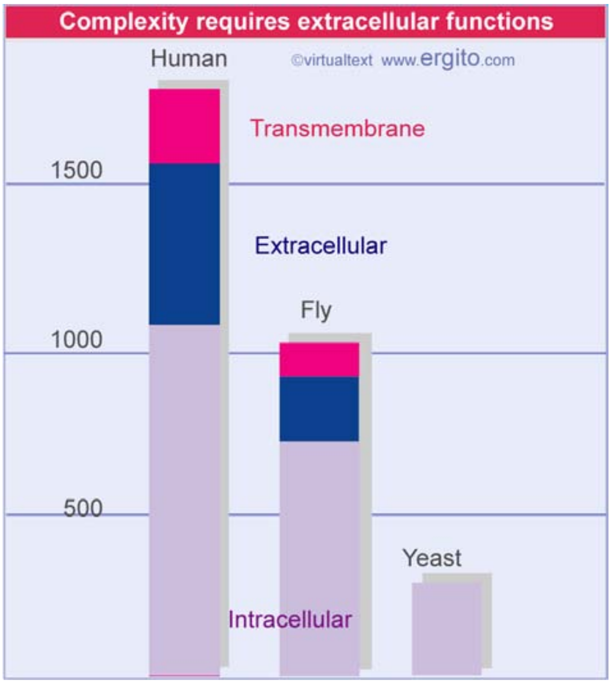


Figure 3.28 Increasing complexity in eukaryotes is accompanied by accumulation of new proteins for transmembrane and extracellular functions

Last updated on 2-16-2001

This content is available online at http://www.ergito.com/main.jsp?bcs=MBIO.1.3.14

THE CONTENT OF THE GENOME

1.3.15 How many genes are essential?

Key Terms

Redundancy describes the concept that two or more genes may fulfill the same function, so that no single one of them is essential.

Synthetic lethality occurs when two mutations that by themselves are viable, cause lethality when combined.

Synthetic genetic array analysis (SGA) is an automated technique in budding yeast whereby a mutant is crossed to an array of approximately 5000 deletion mutants to determine if the mutants interact to cause a synthetic lethal phenotype.

Key Concepts

- Not all genes are essential. In yeast and fly, deletions of <50% of the genes have detectable effects.
- When two or more genes are redundant, a mutation in any one of them may not have detectable effects.
- We do not fully understand the survival in the genome of genes that are apparently dispensable.

Natural selection is the force that ensures that useful genes are retained in the genome. Mutations occur at random, and their most common effect in an open reading frame will be to damage the protein product. An organism with a damaging mutation will be at a disadvantage in evolution, and ultimately the mutation will be eliminated by the competitive failure of organisms carrying it. The frequency of a disadvantageous allele in the population is balanced between the generation of new mutations and the elimination of old mutations. Reversing this argument, whenever we see an intact open reading frame in the genome, we assume that its product plays a useful role in the organism. Natural selection must have prevented mutations from accumulating in the gene. The ultimate fate of a gene that ceases to be useful is to accumulate mutations until it is no longer recognizable.

The maintenance of a gene implies that it confers a selective advantage on the organism. But in the course of evolution, even a small relative advantage may be the subject of natural selection, and a phenotypic defect may not necessarily be immediately detectable as the result of a mutation. However, we should like to know how many genes are actually *essential*. This means that their absence is lethal to the organism. In the case of diploid organisms, it means of course that the homozygous null mutation is lethal.

We might assume that the proportion of essential genes will decline with increase in genome size, given that larger genomes may have multiple, related copies of particular gene functions. So far this expectation has not been borne out by the data (see **Figure 3.9**).

One approach to the issue of gene number is to determine the number of essential genes by mutational analysis. If we saturate some specified region of the chromosome with mutations that are lethal, the mutations should map into a number of complementation groups that corresponds to the number of lethal loci in that region. By extrapolating to the genome as a whole, we may calculate the total essential gene number .

In the organism with the smallest known genome (*Mycoplasma genitalium*), random insertions have detectable effects only in about two thirds of the genes (929). Similarly, fewer than half of the genes of *E. coli* appear to be essential. The proportion is even lower in the yeast *S. cerevisiae.* When insertions were introduced at random into the genome in one early analysis, only 12% were lethal, and another 14% impeded growth. The majority (70%) of the insertions had no effect (401). A more systematic survey based on completely deleting each of 5,916 genes (>96% of the identified genes) shows that only 18.7% are essential for growth on a rich medium (that is, when nutrients are fully provided) (2866). **Figure 3.29** shows that these include genes in all categories. The only notable concentration of defects is in genes coding for products involved in protein synthesis, where ~50% are essential. Of course, this approach underestimates the number of genes that are essential for the yeast to live in the wild, when it is not so well provided with nutrients.

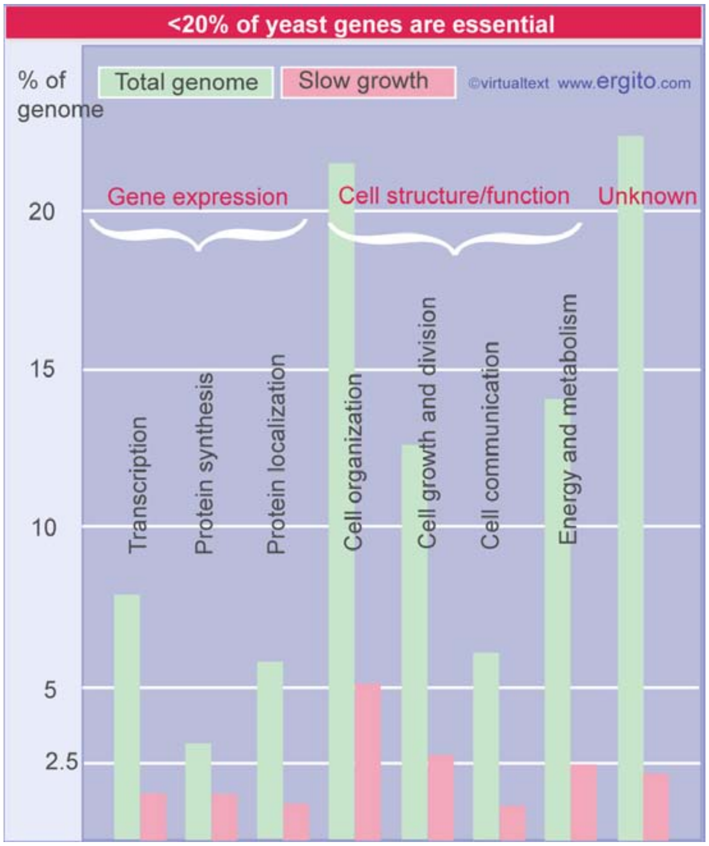


Figure 3.29 Essential yeast genes are found in all classes. Green bars show total proportion of each class of genes, red bars shows those that are essential.

Figure 3.30 summarizes the results of a systematic analysis of the effects of loss of gene function in the worm *C. elegans* (3323). The sequences of individual genes were predicted from the genome sequence, and by targeting an inhibitory RNA against these sequences (see *Molecular Biology 3.11.22 RNA interference is related to gene silencing*), a large set of worms were made in which one (predicted) gene was prevented from functioning in each worm. Detectable effects on the phenotype were only observed for 10% of these knockouts, suggesting that most genes do not play essential roles.

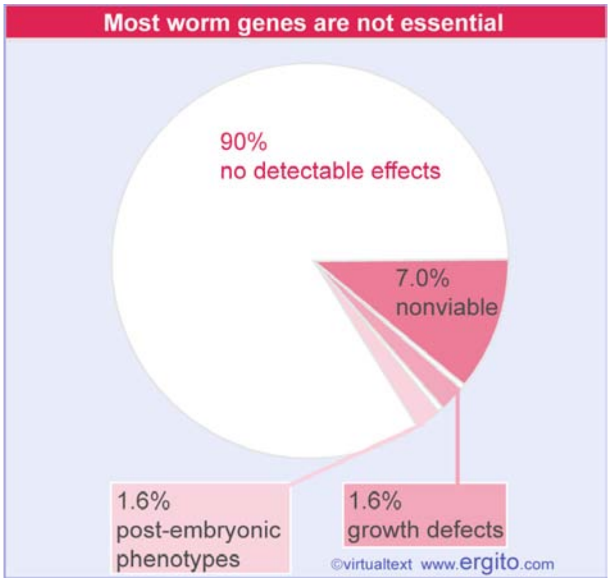


Figure 3.30 A systematic analysis of loss of function for 86% of worm genes shows that only 10% have detectable effects on the phenotype.

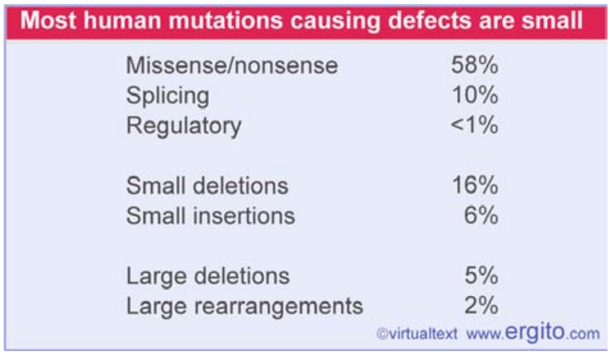
There is a greater proportion of essential genes (21%) among those worm genes that have counterparts in other eukaryotes, suggesting that widely conserved genes tend to play more basic functions. There is also an increased proportion of essential genes among those that are present in only one copy per haploid genome, compared with those where there are multiple copies of related or identical genes. This suggests that many of the multiple genes might be relatively recent duplications that can substitute for one another's functions,

Extensive analyses of essential gene number in a higher eukaryote have been made in *Drosophila* through attempts to correlate visible aspects of chromosome structure with the number of functional genetic units. The notion that this might be possible arose originally from the presence of bands in the polytene chromosomes of *D. melanogaster*. (These chromosomes are found at certain developmental stages and represent an unusually extended physical form, in which a series of bands [more formally called chromomeres] are evident; see *Molecular Biology 5.19.10 Polytene chromosomes form bands*.) From the early concept that the bands might represent a linear order of genes, we have come to the attempt to correlate the organization of genes with the organization of bands. There are ~5000 bands in the *D. melanogaster* haploid set; they vary in size over an order of magnitude, but on average there is ~20 kb of DNA per band.

The basic approach is to saturate a chromosomal region with mutations. Usually the mutations are simply collected as lethals, without analyzing the cause of the lethality. Any mutation that is lethal is taken to identify a locus that is essential for the organism. Sometimes mutations cause visible deleterious effects short of lethality, in which case we also count them as identifying an essential locus. When the mutations are placed into complementation groups, the number can be compared with the number of bands in the region, or individual complementation groups may even be assigned to individual bands. The purpose of these experiments has been to determine whether there is a consistent relationship between bands and genes; for example, does every band contain a single gene?

Totaling the analyses that have been carried out over the past 30 years, the number of lethal complementation groups is ~70% of the number of bands. It is an open question whether there is any functional significance to this relationship. But irrespective of the cause, the equivalence gives us a reasonable estimate for the lethal gene number of ~3600. By any measure, the number of lethal loci in *Drosophila* is significantly less than the total number of genes.

If the proportion of essential human genes is similar to other eukaryotes, we would predict a range of 4000-8000 genes in which mutations would be lethal or produce evidently damaging effects. At the present, 1300 genes have been identified in which mutations cause evident defects. This is a substantial proportion of the expected total, especially in view of the fact that many lethal genes may act so early that we never see their effects. This sort of bias may also explain the results in **Figure 3.31**, which show that the majority of known genetic defects are due to point mutations (where there is more likely to be at least some residual function of the gene).



Most human mutations causing defects are small

Missense/nonsense	58%
Splicing	10%
Regulatory	<1%
Small deletions	16%
Small insertions	6%
Large deletions	5%
Large rearrangements	2%

©virtualtext www.ergito.com

Figure 3.31 Most known genetic defects in human genes are due to point mutations. The majority directly affect the protein sequence. The remainder are due to insertion, deletions, or rearrangements of varying sizes.

How do we explain the survival of genes whose deletion appears to have no effect? The most likely explanation is that the organism has alternative ways of fulfilling the same function. The simplest possibility is that there is **redundancy**, and that some genes are present in multiple copies. This is certainly true in some cases, in which multiple (related) genes must be knocked out in order to produce an effect. In a slightly more complex scenario, an organism might have two separate pathways capable of providing some activity. Inactivation of either pathway by itself would not be damaging, but the simultaneous occurrence of mutations in genes from both pathways would be deleterious.

Such situations can be tested by combining mutations. In principle, deletions in two genes, neither of which is lethal by itself, are introduced into the same strain. If the double mutant dies, the strain is called a **synthetic lethal**. This technique has been used to great effect with yeast, where the isolation of double mutants can be automated. The procedure is called **synthetic genetic array analysis** (**SGA**). **Figure 3.32** summarizes the results of an analysis in which an SGA screen was made for each of 132 viable deletions, by testing whether it could survive in combination with any one of 4,700 viable deletions. Every one of the test genes had at least one partner with which the combination was lethal, and most of the test genes had many such partners; the median is ~25 partners, and the greatest number is shown by one test gene that had 146 lethal partners(4831). A small proportion (~10%) of the interacting mutant pairs code for proteins that interact physically.

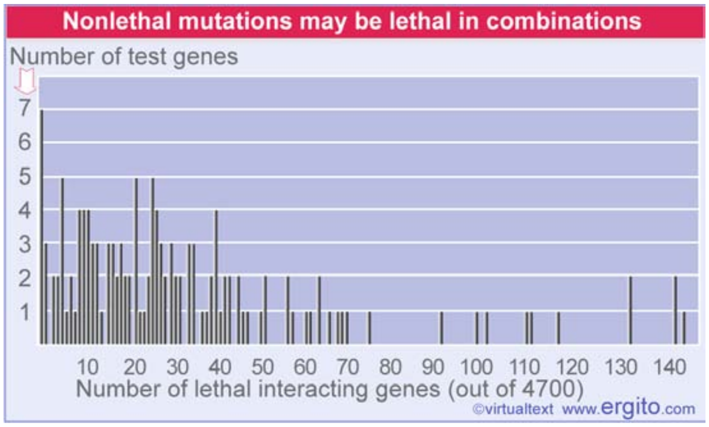


Figure 3.32 Every one of 132 mutant test genes has some combinations that are lethal when it is combined with each of 4700 nonlethal mutations. The chart shows how many lethal interacting genes there are for each test gene.

This result goes some way toward explaining the apparent lack of effect of so many deletions. Natural selection will act against these deletions when they find themselves in lethal pairwise combinations. To some degree, the organism has protected itself against the damaging effects of mutations by building in redundancy. However, it pays a price in the form of accumulating the "genetic load" of mutations that are not deleterious in themselves, but that may cause serious problems when combined with other such mutations in future generations. The theory of natural selection would suggest that the loss of the individual genes in such circumstances produces a sufficient disadvantage to maintain the active gene during the course of evolution.

Last updated on March 9, 2004

References

401. Goebl, M. G. and Petes, T. D. (1986). *Most of the yeast genomic sequences are not essential for cell growth and division*. Cell 46, 983-992.

929. Hutchison, C. A. et al. (1999). *Global transposon mutagenesis and a minimal mycoplasma genome.* Science 286, 2165-2169.

2866. Giaever et al. (2002). *Functional profiling of the S. cerevisiae genome.* Nature 418, 387-391.

3323. Kamath, R. S., Fraser, A. G., Dong, Y., Poulin, G., Durbin, R., Gotta, M., Kanapin, A., Le Bot, N., Moreno, S., Sohrmann, M., Welchman, D. P., Zipperlen, P., and Ahringer, J. (2003). *Systematic functional analysis of the C. elegans genome using RNAi.* Nature 421, 231-237.

4831. Tong, A. H. et al. (2004). *Global mapping of the yeast genetic interaction network.* Science 303, 808-813.

This content is available online at http://www.ergito.com/main.jsp?bcs=MBIO.1.3.15

THE CONTENT OF THE GENOME

1.3.16 Genes are expressed at widely differing levels

Key Terms

The **abundance** of an mRNA is the average number of molecules per cell.

Abundant mRNAs consist of a small number of individual species, each present in a large number of copies per cell.

Scarce mRNA (Complex mRNA) consists of a large number of individual mRNA species, each present in very few copies per cell. This accounts for most of the sequence complexity in RNA.

Key Concepts

- In any given cell, most genes are expressed at a low level.
- Only a small number of genes, whose products are specialized for the cell type, are highly expressed.

The proportion of DNA represented in an mRNA population can be determined by the amount of the DNA that can hybridize with the RNA. Such a saturation analysis typically identifies ~1% of the DNA as providing a template for mRNA. From this we can calculate the number of genes so long as we know the average length of an mRNA. For a lower eukaryote such as yeast, the total number of expressed genes is ~4000. For somatic tissues of higher eukaryotes, the number usually is 10,000-15,000. The value is similar for plants and for vertebrates. (The only consistent exception to this type of value is presented by mammalian brain, where much larger numbers of genes appear to be expressed, although the exact quantitation is not certain.)

Kinetic analysis of the reassociation of an RNA population can be used to determine its sequence complexity (see *Molecular Biology Supplement 32.1 DNA reassociation kinetics*). This type of analysis typically identifies three components in a eukaryotic cell. Just as with a DNA reassociation curve, a single component hybridizes over about two decades of Rot (RNA concentration × time) values, and a reaction extending over a greater range must be resolved by computer curve-fitting into individual components. Again this represents what is really a continuous spectrum of sequences.

An example of an excess mRNA × cDNA reaction that generates three components is given in **Figure 3.33**:

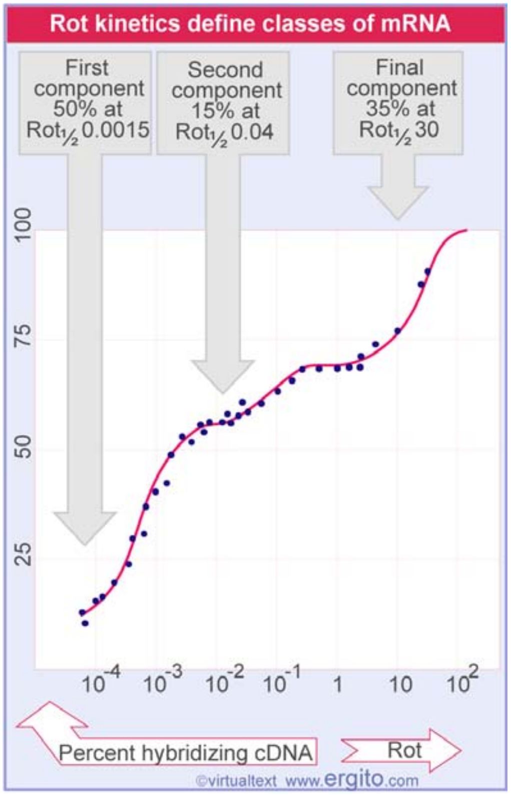


Figure 3.33 Hybridization between excess mRNA and cDNA identifies several components in chick oviduct cells, each characterized by the $Rot_{1/2}$ of reaction.

- The first component has the same characteristics as a control reaction of ovalbumin mRNA with its DNA copy. This suggests that the first component is in fact just ovalbumin mRNA (which indeed occupies about half of the messenger mass in oviduct tissue).
- The next component provides 15% of the reaction, with a total complexity of 15 kb. This corresponds to 7-8 mRNA species of average length 2000 bases.
- The last component provides 35% of the reaction, which corresponds to a complexity of 26 Mb. This corresponds to ~13,000 mRNA species of average length 2000 bases.

From this analysis, we can see that about half of the mass of mRNA in the cell represents a single mRNA, ~15% of the mass is provided by a mere 7-8 mRNAs, and ~35% of the mass is divided into the large number of 13,000 mRNA species. It is therefore obvious that the mRNAs comprising each component must be present in very different amounts.

The average number of molecules of each mRNA per cell is called its **abundance**. It can be calculated quite simply if the total mass of RNA in the cell is known. In the example shown in **Figure 3.33**, the total mRNA can be accounted for as 100,000 copies of the first component (ovalbumin mRNA), 4000 copies of each of the 7-8 mRNAs in the second component, but only ~5 copies of each of the 13,000 mRNAs that constitute the last component.

We can divide the mRNA population into two general classes, according to their abundance:

- The oviduct is an extreme case, with so much of the mRNA represented in only one species, but most cells do contain a small number of RNAs present in many copies each. This **abundant mRNA** component typically consists of <100 different mRNAs present in 1000-10,000 copies per cell. It often corresponds to a major part of the mass, approaching 50% of the total mRNA.
- About half of the mass of the mRNA consists of a large number of sequences, of the order of 10,000, each represented by only a small number of copies in the mRNA – say, <10. This is the **scarce mRNA** or **complex mRNA** class. It is this class that drives a saturation reaction (396).

References

396. Hastie, N. B. and Bishop, J. O. (1976). *The expression of three abundance classes of mRNA in mouse tissues*. Cell 9, 761-774.

This content is available online at http://www.ergito.com/main.jsp?bcs=MBIO.1.3.16

THE CONTENT OF THE GENOME

1.3.17 How many genes are expressed?

Key Terms

The **transcriptome** is the complete set of RNAs present in a cell, tissue, or organism. Its complexity is due mostly to mRNAs, but it also includes noncoding RNAs.

Housekeeping genes (Constitutive gene) are those (theoretically) expressed in all cells because they provide basic functions needed for sustenance of all cell types.

Luxury genes are those coding for specialized functions synthesized (usually) in large amounts in particular cell types.

Key Concepts

- mRNAs expressed at low levels overlap extensively when different cell types are compared.
- The abundantly expressed mRNAs are usually specific for the cell type.
- ~10,000 expressed genes may be common to most cell types of a higher eukaryote.

Many somatic tissues of higher eukaryotes have an expressed gene number in the range of 10,000-20,000. How much overlap is there between the genes expressed in different tissues? For example, the expressed gene number of chick liver is ~11,000-17,000, compared with the value for oviduct of ~13,000-15,000. How many of these two sets of genes are identical? How many are specific for each tissue? These questions are usually addressed by analyzing the **transcriptome** – the set of sequences represented in RNA.

We see immediately that there are likely to be substantial differences among the genes expressed in the abundant class. Ovalbumin, for example, is synthesized only in the oviduct, not at all in the liver. This means that 50% of the mass of mRNA in the oviduct is specific to that tissue.

But the abundant mRNAs represent only a small proportion of the number of expressed genes. In terms of the total number of genes of the organism, and of the number of changes in transcription that must be made between different cell types, we need to know the extent of overlap between the genes represented in the scarce mRNA classes of different cell phenotypes.

Comparisons between different tissues show that, for example, ~75% of the sequences expressed in liver and oviduct are the same. In other words, ~12,000 genes are expressed in both liver and oviduct, ~5000 additional genes are expressed only in liver, and ~3000 additional genes are expressed only in oviduct.

The scarce mRNAs overlap extensively. Between mouse liver and kidney, ~90% of the scarce mRNAs are identical, leaving a difference between the tissues of only 1000-2000 in terms of the number of expressed genes. The general result obtained in several comparisons of this sort is that only ~10% of the mRNA sequences of a cell are unique to it. The majority of sequences are common to many, perhaps even all, cell types.

This suggests that the common set of expressed gene functions, numbering perhaps ~10,000 in a mammal, comprise functions that are needed in all cell types. Sometimes this type of function is referred to as a **housekeeping gene** or **constitutive gene**. It contrasts with the activities represented by specialized functions (such as ovalbumin or globin) needed only for particular cell phenotypes. These are sometimes called **luxury genes**.

Last updated on 12-13-2001

This content is available online at http://www.ergito.com/main.jsp?bcs=MBIO.1.3.17

THE CONTENT OF THE GENOME

1.3.18 Expressed gene number can be measured *en masse*

Key Concepts

- "Chip" technology allows a snapshot to be taken of the expression of the entire genome in a yeast cell.
- ~75% (~4500 genes) of the yeast genome is expressed under normal growth conditions.
- Chip technology allows detailed comparisons of related animal cells to determine (for example) the differences in expression between a normal cell and a cancer cell.

Recent technology allows more systematic and accurate estimates of the number of expressed genes. One approach (SAGE, serial analysis of gene expression) allows a unique sequence tag to be used to identify each mRNA. The technology then allows the abundance of each tag to be measured. This approach identifies 4,665 expressed genes in *S. cerevisiae* growing under normal conditions, with abundances varying from 0.3 to >200 transcripts/cell. This means that ~75% of the total gene number (~6000) is expressed under these conditions (397). **Figure 3.34** summarizes the number of different mRNAs that is found at each different abundance levels.

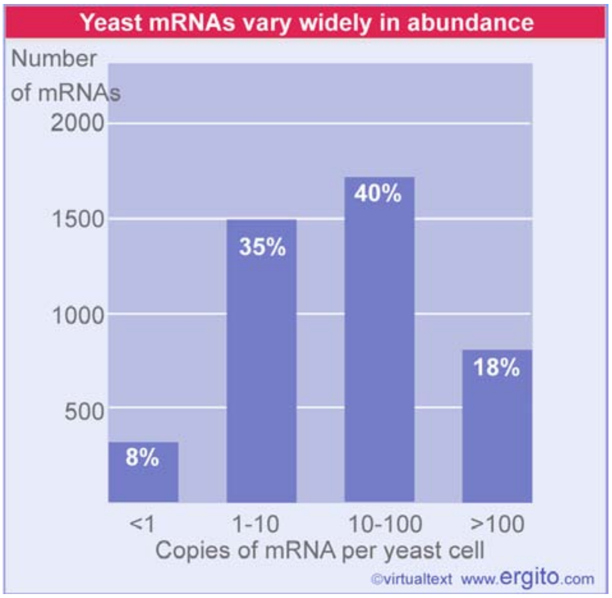


Figure 3.34 The abundancies of yeast mRNAs vary from <1 per cell (meaning that not every cell has a copy of the mRNA) to >100 per cell (coding for the more abundant proteins).

The most powerful new technology uses chips that contain high-density oligonucleotide arrays (HDAs). Their construction is made possibly by knowledge of the sequence of the entire genome. In the case of *S. cerevisiae*, each of 6181 ORFs is represented on the HDA by 20 25-mer oligonucleotides that perfectly match the sequence of the message and 20 mismatch oligonucleotides that differ at one base position. The expression level of any gene is calculated by subtracting the average signal of a mismatch from its perfect match partner. The entire yeast genome can be represented on 4 chips. This technology is sensitive enough to detect transcripts of 5460 genes (~90% of the genome), and shows that many genes are expressed at low levels, with abundances of 0.1-2 transcripts/cell. An abundance of <1 transcript/cell means that not all cells have a copy of the transcript at any given moment.

The technology allows not only measurement of levels of gene expression, but also detection of differences in expression in mutant cells compared with wild-type, cells growing under different growth conditions, and so on (1201; for review see 1200). The results of comparing two states are expressed in the form of a grid, in which each square represents a particular gene, and the relative change in expression is indicated by color. The upper part of **Figure 3.35** shows the effect of a mutation in RNA polymerase II, the enzyme that produces mRNA, which as might be expected causes the expression of most genes to be heavily reduced. By contrast, the lower part shows that a mutation in an ancillary component of the transcription apparatus (*SRB10*) has much more restricted effects, causing increases in expression of some genes (398).

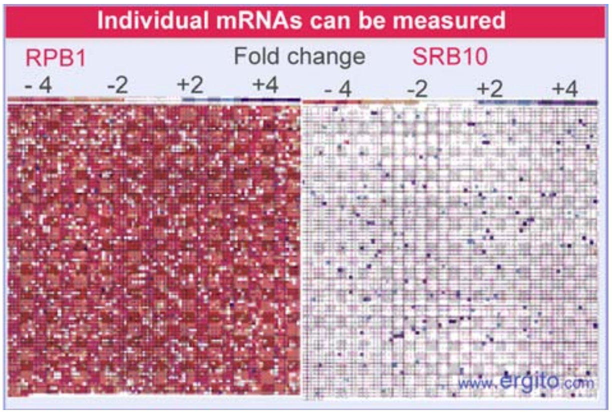


Figure 3.35 HDA analysis allows change in expression of each gene to be measured. Each square represents one gene (top left is first gene on chromosome I, bottom right is last gene on chromosome XVI). Change in expression relative to wild type is indicated by red (reduction), white (no change) or blue (increase). Photograph kindly provided by Rick Young.

The extension of this technology to animal cells will allow the general descriptions based on RNA hybridization analysis to be replaced by exact descriptions of the genes that are expressed, and the abundances of their products, in any given cell type (399).

Last updated on January 16, 2004

Reviews

399. Mikos, G. L. G. and Rubin, G. M. (1996). *The role of the genome project in determining gene function: insights from model organisms*. Cell 86, 521-529.

1200. Young, R. A. (2000). *Biomedical discovery with DNA arrays.* Cell 102, 9-15.

References

397. Velculescu, V. E. et al. (1997). *Characterization of the yeast transcriptosome*. Cell 88, 243-251.

398. Holstege, F. C. P. et al. (1998). *Dissecting the regulatory circuitry of a eukaryotic genome*. Cell 95, 717-728.

1201. Hughes, T. R., Marton, M. J., Jones, A. R., Roberts, C. J., Stoughton, R., Armour, C. D., Bennett, H. A., Coffey, E., Dai, H., He, Y. D., Kidd, M. J., King, A. M., Meyer, M. R., Slade, D., Lum, P. Y., Stepaniants, S. B., Shoemaker, D. D. et al., (2000). *Functional discovery via a compendium of expression profiles.* Cell 102, 109-126.

This content is available online at http://www.ergito.com/main.jsp?bcs=MBIO.1.3.18

THE CONTENT OF THE GENOME

1.3.19 Organelles have DNA

Key Terms

Maternal inheritance describes the preferential survival in the progeny of genetic markers provided by one parent.

Extranuclear genes reside outside the nucleus in organelles such as mitochondria and chloroplasts.

Cytoplasmic inheritance is a property of genes located in mitochondria or chloroplasts.

Key Concepts

- Mitochondria and chloroplasts have genomes that show nonMendelian inheritance. Typically they are maternally inherited.
- Organelle genomes may undergo somatic segregation in plants.
- Comparisons of mitochondrial DNA suggest that humans are descended from a single female who lived 200,000 years ago in Africa.

The first evidence for the presence of genes outside the nucleus was provided by nonMendelian inheritance in plants (observed in the early years of this century, just after the rediscovery of Mendelian inheritance). NonMendelian inheritance is sometimes associated with the phenomenon of somatic segregation. They have a similar cause:

- NonMendelian inheritance is defined by the failure of the progeny of a mating to display Mendelian segregation for parental characters. It reflects lack of association between the segregating character and the meiotic spindle.
- Somatic segregation describes a phenomenon in which parental characters segregate in somatic cells, and therefore display heterogeneity in the organism. This is a notable feature of plant development. It reflects lack of association between the segregating character and the mitotic spindle.

*NonMendelian inheritance and somatic segregation are therefore taken to indicate the presence of genes that reside outside the nucleus and do not utilize segregation on the meiotic and mitotic spindles to distribute replicas to gametes or to daughter cells, respectively.***Figure 3.36** shows that this happens when the mitochondria inherited from the male and female parents have different alleles, and by chance a daughter cell receives an unbalanced distribution of mitochondria that represents only one parent (see *Molecular Biology 4.13.24 How do mitochondria replicate and segregate?*).

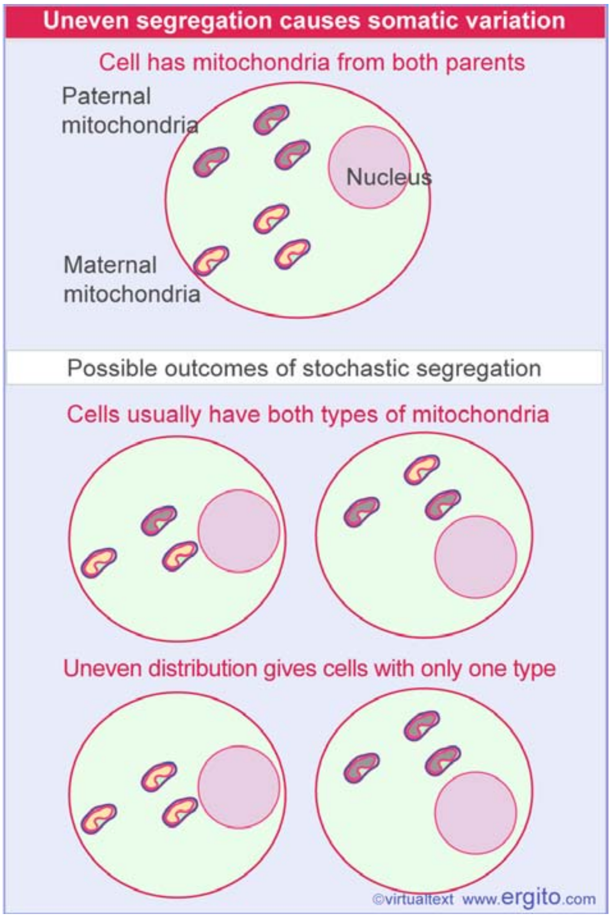

Uneven segregation causes somatic variation
Cell has mitochondria from both parents
Paternal mitochondria
Nucleus
Maternal mitochondria
Possible outcomes of stochastic segregation
Cells usually have both types of mitochondria
Uneven distribution gives cells with only one type
©virtualtext www.ergito.com

Figure 3.36 When paternal and maternal mitochondrial alleles differ, a cell has two sets of mitochondrial DNAs. Mitosis usually generates daughter cells with both sets. Somatic variation may result if unequal segregation generates daughter cells with only one set.

The extreme form of nonMendelian inheritance is uniparental inheritance, when the genotype of only one parent is inherited and that of the other parent is permanently lost. In less extreme examples, the progeny of one parental genotype exceed those of the other genotype. Usually it is the mother whose genotype is preferentially (or solely) inherited. This effect is sometimes described as **maternal inheritance**. The important point is that the genotype contributed by the parent of one particular sex predominates, as seen in abnormal segregation ratios when a cross is made between mutant and wild type. This contrasts with the behavior of Mendelian genetics when reciprocal crosses show the contributions of both parents to be equally inherited.

The bias in parental genotypes is established at or soon after the formation of a zygote. There are various possible causes. The contribution of maternal or paternal information to the organelles of the zygote may be unequal; in the most extreme case, only one parent contributes. In other cases, the contributions are equal, but the information provided by one parent does not survive. Combinations of both effects are possible. Whatever the cause, the unequal representation of the information from the two parents contrasts with nuclear genetic information, which derives equally from each parent.

NonMendelian inheritance results from the presence in mitochondria and chloroplasts of DNA genomes that are inherited independently of nuclear genes. In effect, the organelle genome comprises a length of DNA that has been physically sequestered in a defined part of the cell, and is subject to its own form of expression and regulation. An organelle genome can code for some or all of the RNAs, but codes for only some of the proteins needed to perpetuate the organelle. The other proteins are coded in the nucleus, expressed via the cytoplasmic protein synthetic apparatus, and imported into the organelle.

Genes not residing within the nucleus are generally described as **extranuclear genes**; they are transcribed and translated in the *same* organelle compartment (mitochondrion or chloroplast) in which they reside. By contrast, *nuclear* genes are expressed by means of *cytoplasmic* protein synthesis. (The term **cytoplasmic inheritance** is sometimes used to describe the behavior of genes in organelles. However, we shall not use this description, since it is important to be able to distinguish between events in the general cytosol and those in specific organelles.)

Higher animals show maternal inheritance, which can be explained if the mitochondria are contributed entirely by the ovum and not at all by the sperm. **Figure 3.37** shows that the sperm contributes only a copy of the nuclear DNA. So the mitochondrial genes are derived exclusively from the mother; and in males they are discarded each generation.

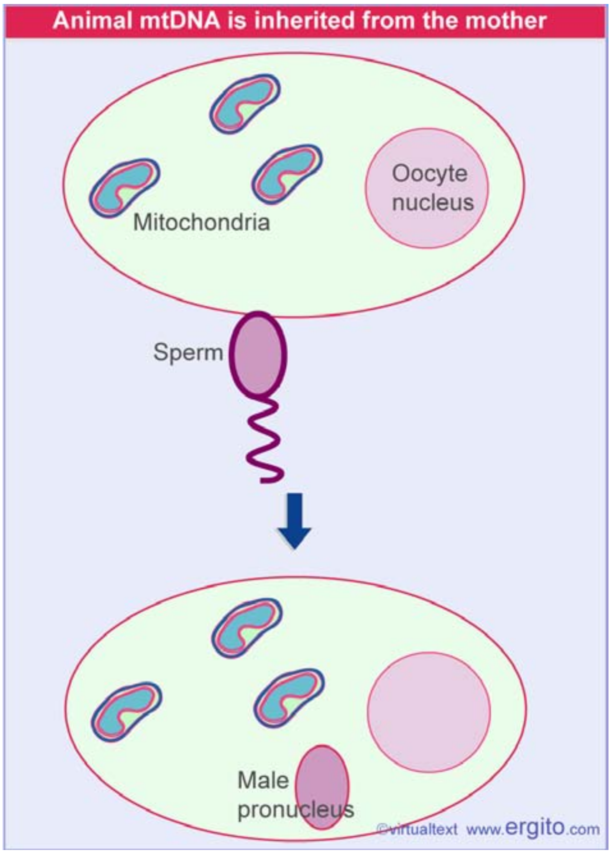

Animal mtDNA is inherited from the mother
Mitochondria
Oocyte nucleus
Sperm
Male pronucleus
©virtualtext www.ergito.com

Figure 3.37 DNA from the sperm enters the oocyte to form the male pronucleus in the fertilized egg, but all the mitochondria are provided by the oocyte.

Conditions in the organelle are different from those in the nucleus, and organelle DNA therefore evolves at its own distinct rate. If inheritance is uniparental, there can be no recombination between parental genomes; and usually recombination does not occur in those cases where organelle genomes are inherited from both parents. Since organelle DNA has a different replication system from that of the nucleus, the error rate during replication may be different. Mitochondrial DNA accumulates mutations more rapidly than nuclear DNA in mammals, but in plants the accumulation in the mitochondrion is slower than in the nucleus (the chloroplast is intermediate).

One consequence of maternal inheritance is that the sequence of mitochondrial DNA is more sensitive than nuclear DNA to reductions in the size of the breeding population. Comparisons of mitochondrial DNA sequences in a range of human populations allow an evolutionary tree to be constructed. The divergence among human mitochondrial DNAs spans 0.57%. A tree can be constructed in which the mitochondrial variants diverged from a common (African) ancestor. The rate at which mammalian mitochondrial DNA accumulates mutations is 2-4% per million years, >10× faster than the rate for globin. Such a rate would generate the observed divergence over an evolutionary period of 140,000-280,000 years. This implies that the human race is descended from a single female, who lived in Africa ~200,000 years ago (414).

Last updated on 2-6-2002

References

414. Cann, R. L., Stoneking, M., and Wilson, A. C. (1987). *Mitochondrial DNA and human evolution*. Nature 325, 31-36.

This content is available online at http://www.ergito.com/main.jsp?bcs=MBIO.1.3.19

THE CONTENT OF THE GENOME

1.3.20 Organelle genomes are circular DNAs that code for organelle proteins

Key Terms

Mitochondrial DNA (mtDNA) is an independent DNA genome, usually circular, that is located in the mitochondrion.

Chloroplast DNA (ctDNA) is an independent genome (usually circular) found in a plant chloroplast.

Key Concepts

- Organelle genomes are usually (but not always) circular molecules of DNA.
- Organelle genomes code for some but not all of the proteins found in the organelle.

Most organelle genomes take the form of a single circular molecule of DNA of unique sequence (denoted **mtDNA** in the mitochondrion and **ctDNA** in the chloroplast). There are a few exceptions where mitochondrial DNA is a linear molecule, generally in lower eukaryotes.

Usually there are several copies of the genome in the individual organelle. Since there are multiple organelles per cell, there are many organelle genomes per cell. Although the organelle genome itself is unique, it constitutes a repetitive sequence relative to any nonrepetitive nuclear sequence.

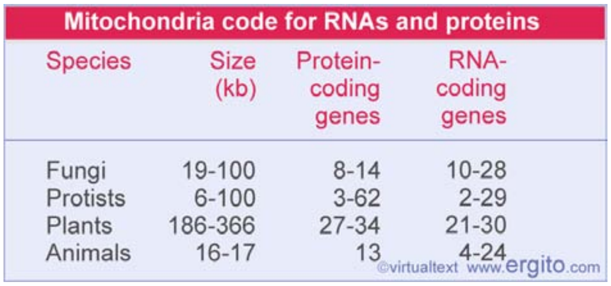
Chloroplast genomes are relatively large, usually ~140 kb in higher plants, and <200 kb in lower eukaryotes. This is comparable to the size of a large bacteriophage, for example, T4 at ~165 kb. There are multiple copies of the genome per organelle, typically 20-40 in a higher plant, and multiple copies of the organelle per cell, typically 20-40.

Mitochondrial genomes vary in total size by more than an order of magnitude. Animal cells have small mitochondrial genomes, ~16.5 kb in mammals. There are several hundred mitochondria per cell. Each mitochondrion has multiple copies of the DNA. The total amount of mitochondrial DNA relative to nuclear DNA is small, <1%.

In yeast, the mitochondrial genome is much larger. In *S. cerevisiae*, the exact size varies among different strains, but is ~80 kb. There are ~22 mitochondria per cell, which corresponds to ~4 genomes per organelle. In growing cells, the proportion of mitochondrial DNA can be as high as 18%.

Plants show an extremely wide range of variation in mitochondrial DNA size, with a minimum of ~100 kb. The size of the genome makes it difficult to isolate intact, but restriction mapping in several plants suggests that the mitochondrial genome is usually a single sequence, organized as a circle. Within this circle, there are short homologous sequences. Recombination between these elements generates smaller, subgenomic circular molecules that coexist with the complete, "master" genome, explaining the apparent complexity of plant mitochondrial DNAs.

With mitochondrial genomes sequenced from many organisms, we can now see some general patterns in the representation of functions in mitochondrial DNA (for review see981). **Figure 3.38** summarizes the distribution of genes in mitochondrial genomes. The total number of protein-coding genes is rather small, but does not correlate with the size of the genome. Mammalian mitochondria use their 16 kb genomes to code for 13 proteins, whereas yeast mitochondria use their 60-80 kb genomes to code for as few as 8 proteins. Plants, with much larger mitochondrial genomes, code for more proteins. Introns are found in most mitochondrial genomes, although not in the very small mammalian genomes.



Mitochondria code for RNAs and proteins

Species	Size (kb)	Protein-coding genes	RNA-coding genes
Fungi	19-100	8-14	10-28
Protists	6-100	3-62	2-29
Plants	186-366	27-34	21-30
Animals	16-17	13	4-24

©virtualtext www.ergito.com

Figure 3.38 Mitochondrial genomes have genes coding for (mostly complex I-IV) proteins, rRNAs, and tRNAs.

The two major rRNAs are always coded by the mitochondrial genome. The number of tRNAs coded by the mitochondrial genome varies from none to the full complement (25-26 in mitochondria). This accounts for the variation in **Figure 3.38**.

The major part of the protein-coding activity is devoted to the components of the multisubunit assemblies of respiration complexes I-IV. Many ribosomal proteins are coded in protist and plant mitochondrial genomes, but there are few or none in fungi and animal genomes. There are genes coding for proteins involved in import in many protist mitochondrial genomes.

Last updated on 5-15-2000

Reviews

981. Lang, B. F., Gray, M. W., and Burger, G. (1999). *Mitochondrial genome evolution and the origin of eukaryotes*. Annu. Rev. Genet. 33, 351-397.

This content is available online at http://www.ergito.com/main.jsp?bcs=MBIO.1.3.20

THE CONTENT OF THE GENOME

1.3.21 Mitochondrial DNA organization is variable

Key Concepts

- Animal cell mitochondrial DNA is extremely compact and typically codes for 13 proteins, 2 rRNAs, and 22 tRNAs.
- Yeast mitochondrial DNA is 5× longer than animal cell mtDNA because of the presence of long introns.

Animal mitochondrial DNA is extremely compact. There are extensive differences in the detailed gene organization found in different animal phyla, but the general principle is maintained of a small genome coding for a restricted number of functions. In mammalian mitochondrial genomes, the organization is extremely compact. There are no introns, some genes actually overlap, and almost every single base pair can be assigned to a gene. With the exception of the D loop, a region concerned with the initiation of DNA replication, no more than 87 of the 16,569 bp of the human mitochondrial genome can be regarded as lying in intercistronic regions.

The complete nucleotide sequences of mitochondrial genomes in animal cells show extensive homology in organization (1398). The map of the human mitochondrial genome is summarized in **Figure 3.39**. There are 13 protein-coding regions. All of the proteins are components of the apparatus concerned with respiration. These include cytochrome *b*, 3 subunits of cytochrome oxidase, one of the subunits of ATPase, and 7 subunits (or associated proteins) of NADH dehydrogenase (411; for review see 18; 19; 21).

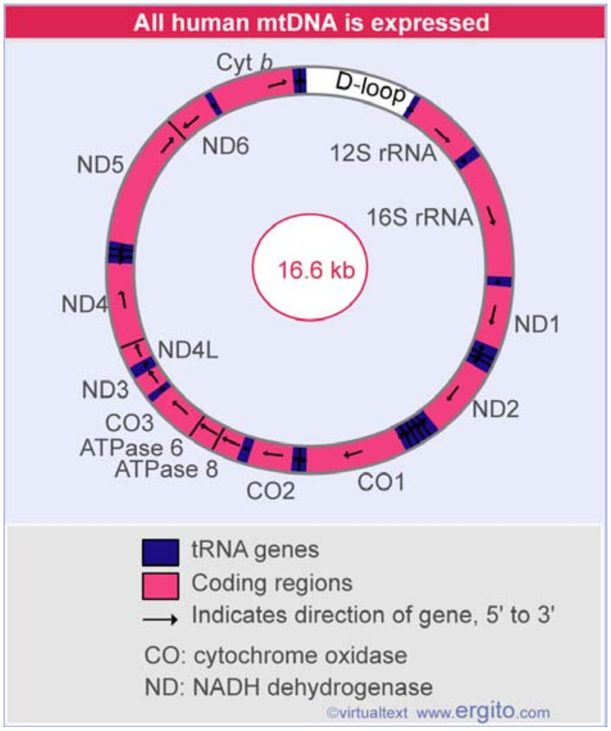


Figure 3.39 Human mitochondrial DNA has 22 tRNA genes, 2 rRNA genes, and 13 protein-coding regions. 14 of the 15 protein-coding or rRNA-coding regions are transcribed in the same direction. 14 of the tRNA genes are expressed in the clockwise direction and 8 are read counter clockwise.

The five-fold discrepancy in size between the *S. cerevisiae* (84 kb) and mammalian (16 kb) mitochondrial genomes alone alerts us to the fact that there must be a great difference in their genetic organization in spite of their common function. The number of endogenously synthesized products concerned with mitochondrial enzymatic functions appears to be similar. Does the additional genetic material in yeast mitochondria represent other proteins, perhaps concerned with regulation, or is it unexpressed?

The map shown in **Figure 3.40** accounts for the major RNA and protein products of the yeast mitochondrion. The most notable feature is the dispersion of loci on the map.

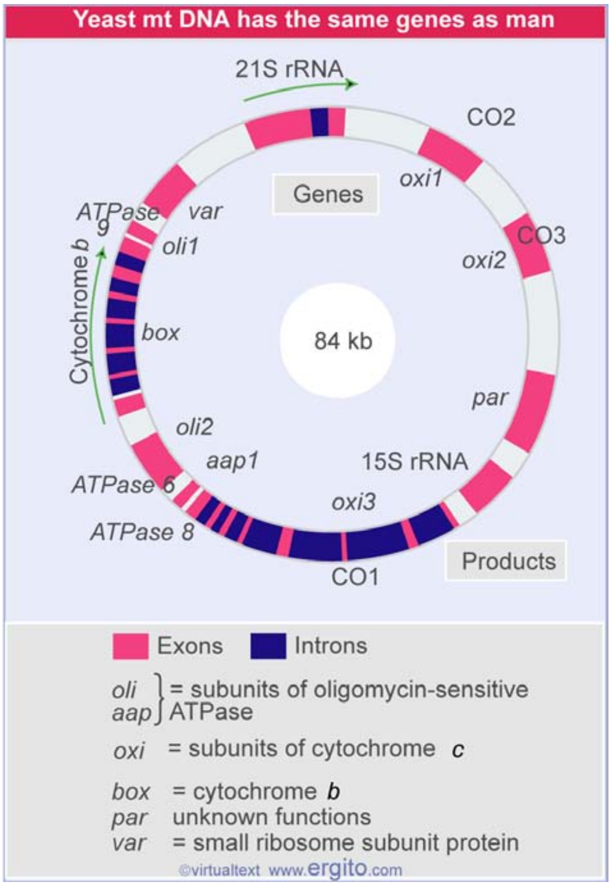

Yeast mt DNA has the same genes as man
21S rRNA
CO2
oxi1
Genes
ATPase 9
var
oli1
CO3
oxi2
Cytochrome b
box
84 kb
par
oli2
aap1
15S rRNA
ATPase 6
oxi3
ATPase 8
Products
CO1
Exons
Introns
oli
aap
= subunits of oligomycin-sensitive ATPase
oxi = subunits of cytochrome c
box = cytochrome b
par unknown functions
var = small ribosome subunit protein
©virtualtext www.ergito.com

Figure 3.40 The mitochondrial genome of *S. cerevisiae* contains both interrupted and uninterrupted protein-coding genes, rRNA genes, and tRNA genes (positions not indicated). Arrows indicate direction of transcription.

The two most prominent loci are the interrupted genes *box* (coding for cytochrome *b*) and *oxi3* (coding for subunit 1 of cytochrome oxidase). Together these two genes are almost as long as the entire mitochondrial genome in mammals! Many of the long introns in these genes have open reading frames in register with the preceding exon (see *Molecular Biology 5.26.5 Some group I introns code for endonucleases that sponsor mobility*). This adds several proteins, all synthesized in low amounts, to the complement of the yeast mitochondrion.

The remaining genes are uninterrupted. They correspond to the other two subunits of cytochrome oxidase coded by the mitochondrion, to the subunit(s) of the ATPase, and (in the case of *var1*) to a mitochondrial ribosomal protein. The total number of yeast mitochondrial genes is unlikely to exceed ~25.

Reviews

18. Clayton, D. A. (1984). *Transcription of the mammalian mitochondrial genome.* Annu. Rev. Biochem. 53, 573-594.

19. Attardi, G. (1985). *Animal mitochondrial DNA: an extreme example of economy*. Int. Rev. Cytol. 93, 93-146.

21. Gray, M. W. (1989). *Origin and evolution of mitochondrial DNA.* Annu. Rev. Cell Biol. 5, 25-50.

1398. Boore, J. L. (1999). *Animal mitochondrial genomes.* Nucleic Acids Res. 27, 1767-1780.

References

411. Anderson, S. et al. (1981). *Sequence and organization of the human mitochondrial genome*. Nature 290, 457-465.

This content is available online at http://www.ergito.com/main.jsp?bcs=MBIO.1.3.21

THE CONTENT OF THE GENOME

1.3.22 Mitochondria evolved by endosymbiosis

How did a situation evolve in which an organelle contains genetic information for some of its functions, while others are coded in the nucleus? **Figure 3.41** shows the endosymbiosis model for mitochondrial evolution, in which primitive cells captured bacteria that provided the functions that evolved into mitochondria and chloroplasts. At this point, the proto-organelle must have contained all of the genes needed to specify its functions.

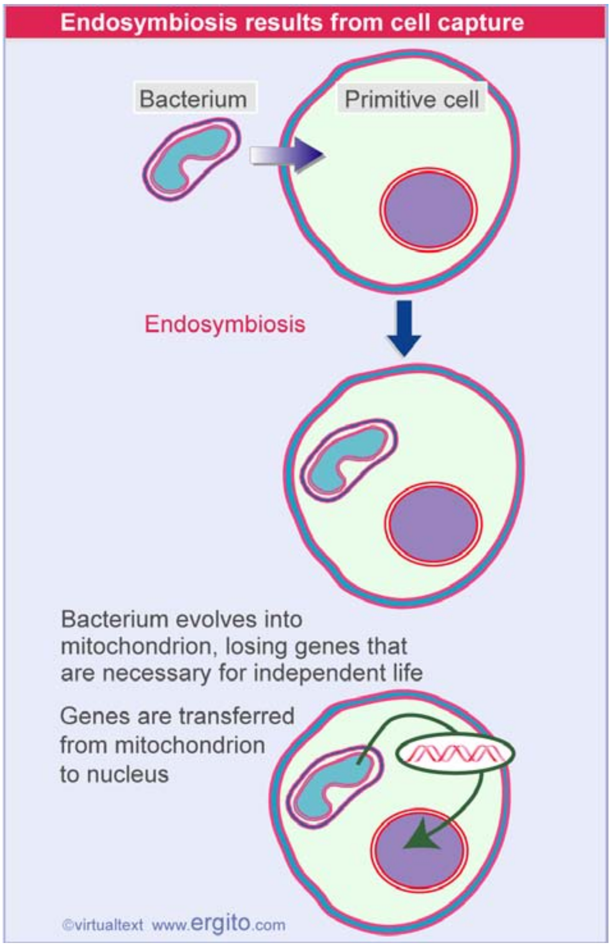

Endosymbiosis results from cell capture
Bacterium
Primitive cell
Endosymbiosis
Bacterium evolves into mitochondrion, losing genes that are necessary for independent life
Genes are transferred from mitochondrion to nucleus
©virtualtext www.ergito.com

Figure 3.41 Mitochondria originated by a endosymbiotic event when a bacterium was captured by a eukaryotic cell.

Sequence homologies suggest that mitochondria and chloroplasts evolved separately, from lineages that are common with eubacteria, with mitochondria sharing an origin with α-purple bacteria, and chloroplasts sharing an origin with cyanobacteria. The closest known relative of mitochondria among the bacteria is *Rickettsia* (the causative agent of typhus), which is an obligate intracellular parasite that is probably descended from free-living bacteria. This reinforces the idea that mitochondria originated in an endosymbiotic event involving an ancestor that is also common to *Rickettsia* (for review see 981).

Two changes must have occurred as the bacterium became integrated into the recipient cell and evolved into the mitochondrion (or chloroplast). The organelles have far fewer genes than an independent bacterium, and have lost many of the gene functions that are necessary for independent life (such as metabolic pathways). And since the majority of genes coding for organelle functions are in fact now located in the nucleus, these genes must have been transferred there from the organelle.

Transfer of DNA between organelle and nucleus has occurred over evolutionary time periods, and still continues. The rate of transfer can be measured directly by introducing into an organelle a gene that can function only in the nucleus, for example, because it contains a nuclear intron, or because the protein must function in the cytosol. In terms of providing the material for evolution, the transfer rates from organelle to nucleus are roughly equivalent to the rate of single gene mutation. DNA introduced into mitochondria is transferred to the nucleus at a rate of 2×10^{-5} per generation. Experiments to measure transfer in the reverse direction, from nucleus to mitochondrion, suggest that it is much lower, $<10^{-10}$ (3691). When a nuclear-specific antibiotic resistance gene is introduced into chloroplasts, its transfer to the nucleus and successful expression can be followed by screening seedlings for resistance to the antibiotic. This shows that transfer occurs at a rate of 1 in 16,000 seedlings, or 6×10^{-5} (3690).

Transfer of a gene from an organelle to the nucleus requires physical movement of the DNA, of course, but successful expression also requires changes in the coding sequence. Organelle proteins that are coded by nuclear genes have special sequences that allow them to be imported into the organelle after they have been synthesized in the cytoplasm (see *Molecular Biology 2.8.17 Post-translational membrane insertion depends on leader sequences*). These sequences are not required by proteins that are synthesized within the organelle. Perhaps the process of effective gene transfer occurred at a period when compartments were less rigidly defined, so that it was easier both for the DNA to be relocated, and for the proteins to be incorporated into the organelle irrespective of the site of synthesis.

Phylogenetic maps show that gene transfers have occurred independently in many different lineages. It appears that transfers of mitochondrial genes to the nucleus occurred only early in animal cell evolution, but it is possible that the process is still continuing in plant cells (1399). The number of transfers can be large; there are >800 nuclear genes in *Arabidopsis* whose sequences are related to genes in the chloroplasts of other plants (1403). These genes are candidates for evolution from genes that originated in the chloroplast

Last updated on 3-12-2003

Reviews

981. Lang, B. F., Gray, M. W., and Burger, G. (1999). *Mitochondrial genome evolution and the origin of eukaryotes*. Annu. Rev. Genet. 33, 351-397.

References

1399. Adams, K. L., Daley, D. O., Qiu, Y. L., Whelan, J., and Palmer, J. D. (2000). *Repeated, recent and diverse transfers of a mitochondrial gene to the nucleus in flowering plants*. Nature 408, 354-357.

1403. The Arabidopsis Genome Initiative. (2000). *Analysis of the genome sequence of the flowering plant Arabidopsis thaliana.* Nature 408, 796-815.

3690. Huang, C. Y., Ayliffe, M. A., and Timmis, J. N. (2003). *Direct measurement of the transfer rate of chloroplast DNA into the nucleus.* Nature 422, 72-76.

3691. Thorsness, P. E. and Fox, T. D. (1990). *Escape of DNA from mitochondria to the nucleus in S. cerevisiae.* Nature 346, 376-379.

This content is available online at http://www.ergito.com/main.jsp?bcs=MBIO.1.3.22

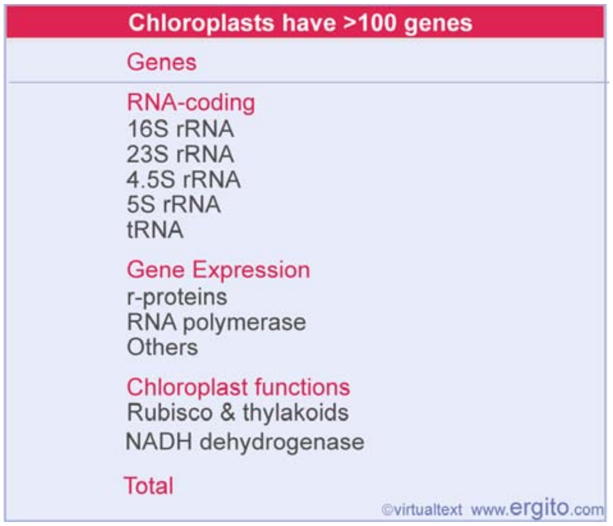
THE CONTENT OF THE GENOME

1.3.23 The chloroplast genome codes for many proteins and RNAs

Key Concepts

- Chloroplast genomes vary in size, but are large enough to code for 50-100 proteins as well as the rRNAs and tRNAs.

What genes are carried by chloroplasts? Chloroplast DNAs vary in length from 120-190 kb. The sequenced chloroplast genomes (>10 in total) have 87-183 genes (413; for review see 20; 3055). **Figure 3.42** summarizes the functions coded by the chloroplast genome in land plants. There is more variation in the chloroplast genomes of algae.



Chloroplasts have >100 genes
Genes
RNA-coding
16S rRNA
23S rRNA
4.5S rRNA
5S rRNA
tRNA
Gene Expression
r-proteins
RNA polymerase
Others
Chloroplast functions
Rubisco & thylakoids
NADH dehydrogenase
Total

©virtualtext www.ergito.com

Figure 3.42 The chloroplast genome in land plants codes for 4 rRNAs, 30 tRNAs, and ~60 proteins.

The situation is generally similar to that of mitochondria, except that more genes are involved. The chloroplast genome codes for all the rRNA and tRNA species needed for protein synthesis. The ribosome includes two small rRNAs in addition to the major species. The tRNA set may include all of the necessary genes. The chloroplast genome codes for ~50 proteins, including RNA polymerase and ribosomal proteins. Again the rule is that organelle genes are transcribed and translated by the apparatus of the organelle.

About half of the chloroplast genes code for proteins involved in protein synthesis. The endosymbiotic origin of the chloroplast is emphasized by the relationships between these genes and their counterparts in bacteria. The organization of the rRNA genes in particular is closely related to that of a cyanobacterium, which pins down more precisely the last common ancestor between chloroplasts and bacteria.

Introns in chloroplasts fall into two general classes. Those in tRNA genes are usually (although not inevitably) located in the anticodon loop, like the introns found in yeast nuclear tRNA genes (see *Molecular Biology 5.24.14 Yeast tRNA splicing involves cutting and rejoining*). Those in protein-coding genes resemble the introns of mitochondrial genes (see *Molecular Biology 5.26 Catalytic RNA*). This places the endosymbiotic event at a time in evolution before the separation of prokaryotes with uninterrupted genes.

The role of the chloroplast is to undertake photosynthesis. Many of its genes code for proteins of complexes located in the thylakoid membranes. The constitution of these complexes shows a different balance from that of mitochondrial complexes. Although some complexes are like mitochondrial complexes in having some subunits coded by the organelle genome and some by the nuclear genome, other chloroplast complexes are coded entirely by one genome.

Last updated on 10-21-2002

Reviews

20. Palmer, J. D. (1985). *Comparative organization of chloroplast genomes.* Annu. Rev. Genet. 19, 325-354.

413. Shimada, H et al. (1991). *Fine structural features of the chloroplast genome: comparison of the sequenced chloroplast genomes*. Nucleic Acids Res. 11, 983-995.

3055. Sugiura, M., Hirose, T., and Sugita, M. (1998). *Evolution and mechanism of translation in chloroplasts.* Annu. Rev. Genet. 32, 437-459.

This content is available online at http://www.ergito.com/main.jsp?bcs=MBIO.1.3.23

THE CONTENT OF THE GENOME

1.3.24 Summary

Genomes that have been sequenced include many bacteria and archaea, yeasts, and a worm, fly, mouse, and Man. The minimum number of genes required to make a living cell (an obligatory intracellular parasite) is ~470. The minimum number required to make a free-living cell is ~1700. A typical gram-negative bacterium has ~1500 genes. Strains of *E. coli* vary from 4300 to 5400 genes. The average bacterial gene is ~1000 bp long and is separated from the next gene by a space of ~100 bp. The yeasts *S. pombe* and *S. cerevisiae* have 5000 and 6000 genes, respectively.

Although the fly *D. melanogaster* is a more complex organism and has a larger genome than the worm *C. elegans*, the fly has fewer genes (13,600) than the worm (14,100). The plant *Arabidopsis* has 25,000 genes, and the lack of a clear relationship between genome size and gene number is shown by the fact that the rice genome is 4× larger, but contains only a 50% increase in gene number, to ~40,000. Mouse has ~30,000 genes. Man has <40,000 genes, which is much less than had been expected. The complexity of development of an organism may depend on the nature of the interactions between genes as well as their total number.

About 8000 genes are common to prokaryotes and eukaryotes and are likely to be involved in basic functions. A further 12,000 genes are found in multicellular organisms. Another 8000 genes are added to make an animal, and a further 8000 (largely involved with the immune and nervous systems) are found in vertebrates. In each organism that has been sequenced, only ~50% of the genes have defined functions. Analysis of lethal genes suggests that only a minority of genes are essential in each organism.

The sequences comprising a eukaryotic genome can be classified in three groups: nonrepetitive sequences are unique; moderately repetitive sequences are dispersed and repeated a small number of times in the form of related but not identical copies; and highly repetitive sequences are short and usually repeated as tandem arrays. The proportions of the types of sequence are characteristic for each genome, although larger genomes tend to have a smaller proportion of nonrepetitive DNA. Almost 50% of the human genome consists of repetitive sequences, the vast majority corresponding to transposon sequences. Most structural genes are located in nonrepetitive DNA. The complexity of nonrepetitive DNA is a better reflection of the complexity of the organism than the total genome complexity; nonrepetitive DNA reaches a maximum complexity of ~2×10^9 bp.

Genes are expressed at widely varying levels. There may be 10^5 copies of mRNA for an abundant gene whose protein is the principal product of the cell, 10^3 copies of each mRNA for <10 moderately abundant messages, and <10 copies of each mRNA for >10,000 scarcely expressed genes. Overlaps between the mRNA populations of cells of different phenotypes are extensive; the majority of mRNAs are present in most cells.

NonMendelian inheritance is explained by the presence of DNA in organelles in the cytoplasm. Mitochondria and chloroplasts both represent membrane-bounded systems in which some proteins are synthesized within the organelle, while others are imported. The organelle genome is usually a circular DNA that codes for all of the RNAs and for some of the proteins that are required.

Mitochondrial genomes vary greatly in size from the 16 kb minimalist mammalian genome to the 570 kb genome of higher plants. It is assumed that the larger genomes code for additional functions. Chloroplast genomes range from 120-200 kb. Those that have been sequenced have a similar organization and coding functions. In both mitochondria and chloroplasts, many of the major proteins contain some subunits synthesized in the organelle and some subunits imported from the cytosol.

Mammalian mtDNAs are transcribed into a single transcript from the major coding strand, and individual products are generated by RNA processing. Rearrangements occur in mitochondrial DNA rather frequently in yeast; and recombination between mitochondrial or between chloroplast genomes has been found. Transfers of DNA have occurred from chloroplasts or mitochondria to nuclear genomes.

This content is available online at http://www.ergito.com/main.jsp?bcs=MBIO.1.3.24