Chapter 2

From genes to genomes

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 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 actually 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 restriction fragments into a map, and in connecting the sequences of the fragments.

Because the use of restriction mapping is central to the molecular analysis of both the genome and the individual gene, we review the principles of the approach briefly before we turn to the structure of the gene itself. In the next section we discuss restriction mapping as such; and the following section discusses its application to construct linkage maps. This puts us in a position to discuss the molecular organization of individual genes, relationships among groups of genes, and the identification of genes in which mutations cause human diseases. In Chapter 3, we consider the overall constitution of the genome and its total number of genes.

Genes can be mapped by restriction cleavage

Once a segment of DNA has been isolated, the first step to obtaining its sequence is to map the nucleic acid at the molecular level. A physical map of any DNA molecule can be obtained by breaking it at defined points whose distance apart can be accurately determined. Specific breaks are made possible by the ability of **restriction enzymes** to recognize rather short sequences of double-stranded DNA as targets for cleavage.

Each restriction enzyme has a particular target in duplex DNA, usually a specific sequence of four to six base pairs. The enzyme cuts the DNA at every point at which its target sequence occurs. Different restriction enzymes have different target sequences, and a large range of these activities (obtained from a wide variety of bacteria) is now available.

A **restriction map** represents a linear sequence of the sites at which particular restriction enzymes find their targets. Distance along such maps is measured directly in base pairs (abbreviated **bp**) for short distances; longer distances are given in **kb,** corresponding to kilobase (10^3) pairs in DNA or to kilobases in RNA. At the level of the chromosome, a map is described in megabase pairs (1 **Mb** = 10^6 bp).

When a DNA molecule is cut with a suitable restriction enzyme, it is cleaved into distinct fragments. These fragments can be separated on the basis of their size by gel electrophoresis. The cleaved DNA is placed on top of a gel made of agarose or polyacrylamide. When an electric current is passed through the gel, each fragment moves down it at a rate that is inversely related to the log of its molecular weight. This movement produces a series of bands. Each band corresponds to a fragment of particular size, decreasing down the gel.

Figure 2.1 DNA can be cleaved by restriction enzymes into fragments that can be separated by gel electrophoresis.

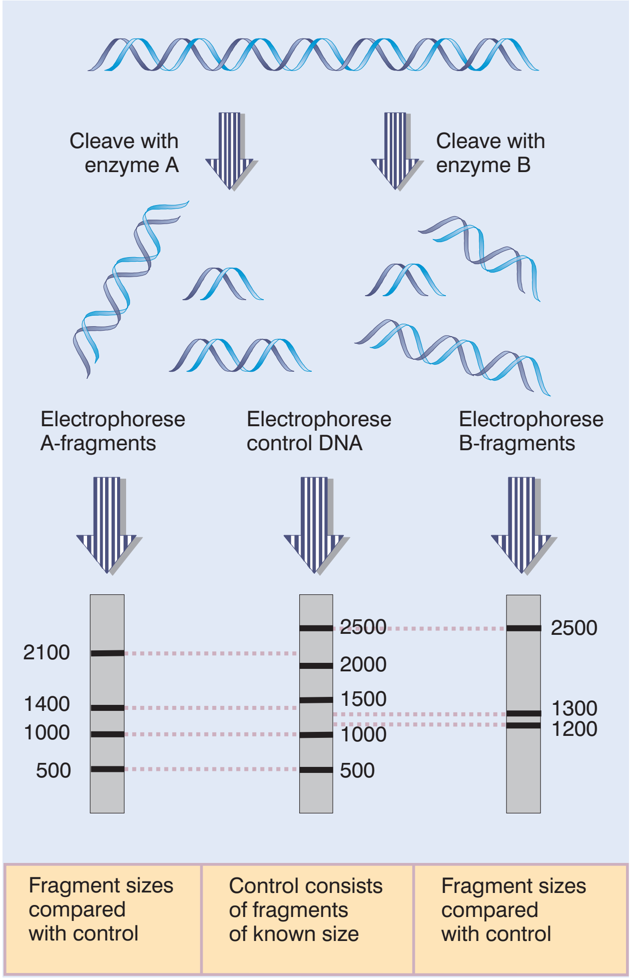

Cleave with enzyme A
Cleave with enzyme B
Electrophorese A-fragments
Electrophorese control DNA
Electrophorese B-fragments
2100
1400
1000
500
2500
2000
1500
1000
500
2500
1300
1200
Fragment sizes compared with control
Control consists of fragments of known size
Fragment sizes compared with control

Figure 2.1 shows an example of this technique. A

DNA molecule of length 5000 bp is incubated separately with two restriction enzymes, A and B. After cleavage the DNA is electrophoresed. The sizes of the individual fragments generated by enzyme A (left) or enzyme B (right) are determined by comparison with the positions of fragments of known size, such as the control shown in the center. This demonstrates that enzyme A has cut the substrate DNA into four fragments (lengths 2100, 1400, 1000, and 500 bp), while enzyme B has generated three fragments (lengths 2500, 1300, and 1200 bp). Can we proceed further from these data to generate a map that places the sites of breakage at defined positions on the DNA?

The patterns of cutting by the two enzymes can be related by several means. **Figure 2.2** illustrates the principle of analysis by **double digestion.** In this technique, the DNA is cleaved simultaneously with two enzymes as well as with either one by itself. The most decisive way to use this technique is to extract each fragment produced in the individual digests with either enzyme A or enzyme B and then to cleave it with the other enzyme. The products of cleavage are analyzed again by electrophoresis.

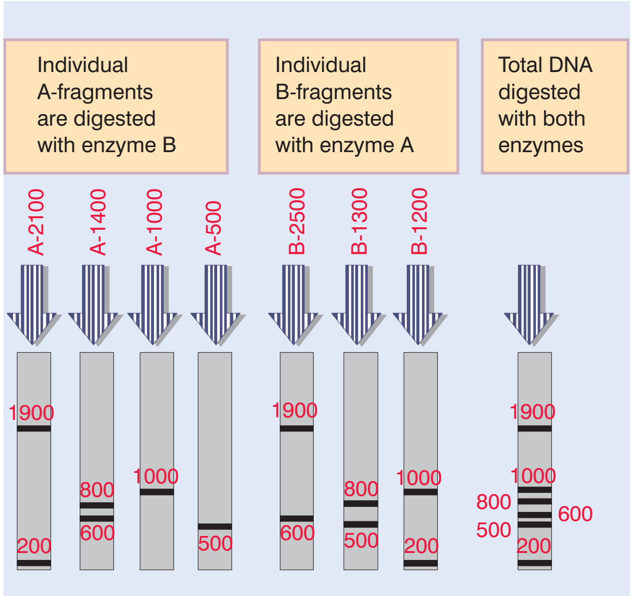
We can use these data to construct a map of the original 5000 bp molecule of DNA, as illustrated by the stages of **Figure 2.3.**

Each gel in Figure 2.2 is labeled according to the fragment that was isolated from the gel in Figure 2.1. A-2100 identifies the fragment of 2100 bp produced by degrading the original DNA molecule with enzyme A. When this fragment is retrieved and subjected to enzyme B, it is cut into fragments of 1900 and 200 bp. So one of the cuts made by enzyme B lies 200 bp from the nearest site cut by enzyme A on one side, and is 1900 bp from the site cut by enzyme A on the other side. This situation is described by the top map in Figure 2.3.

A related pattern of cuts is seen when we examine the susceptibility of fragment B-2500 to enzyme A. It is cut into fragments of 1900 and 600 bp. So the 1900 bp fragment is generated by double cuts, with an A site at one end and a B site at the other end. It can be released from either of the single-cut fragments (A-2100 or B-2500) that contain it. These single-cut fragments must therefore **overlap** in the region of the 1900 bp of the common fragment that can be generated from them. This is described in the second map of Figure 2.3, which extends our map to the right to add a cleavage site for enzyme B.

The key to restriction mapping is the use of overlapping fragments. Because of the overlap of A-2100 and B-2500 in the central region of 1900 bp, we can relate the A site 200 bp to the left of the 1900 bp region with the B site 600 bp to the right. In the same way, we can now extend the map farther on either side. The 200 bp fragment at the left is also produced by cutting B-1200 with enzyme A, so the next B site must lie 1000 bp to the left. The 600 bp fragment at the right is also produced by cutting A-1400 with enzyme B, so the next A site must lie 800 bp to the right. This gives the third map in Figure 2.3.

Figure 2.2 Double digests define the cleavage positions of one enzyme with regard to the other.

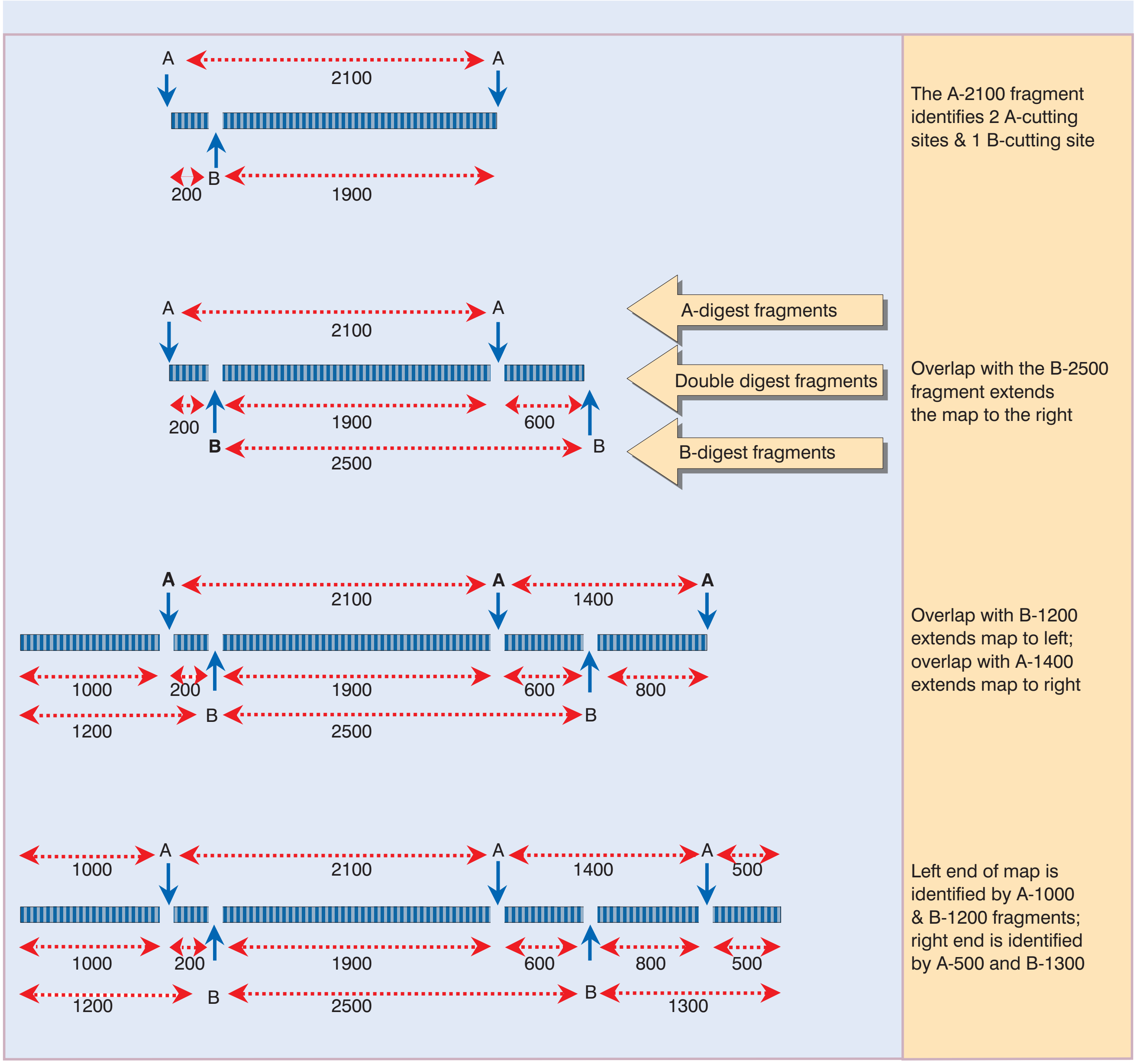

Individual A-fragments are digested with enzyme B
Individual B-fragments are digested with enzyme A
Total DNA digested with both enzymes
A-2100
A-1400
A-1000
A-500
B-2500
B-1300
B-1200
1900
200
800
600
1000
500
1900
600
800
500
1000
200
1900
1000
800
500
600
200

We can now complete the map by identifying the source of the two fragments at each end. At the left end, the 1000 bp fragment arises from B-1200 or in the form of A-1000, which is not cut by enzyme B. So A-1000 lies at the end of the map. Proceeding from the left end of the complete 5000 bp region, it is 1000 bp to the first A site and 1200 bp to the first B site. (This is why a B cut is not shown at the left end of the map above, although formally we treated the end as a B-cutting site in the analysis.)

At the right end of the map, the 800 bp double-cut fragment is generated by cutting B-1300 with enzyme A, so we must add a fragment of 500 bp to the right. This is the terminal fragment, as seen by its presence as A-500 in the single-cut A digest. So our completed map takes the form of the bottom map in Figure 2.3.

The actual construction of a restriction map usually requires recourse to several enzymes, so it becomes necessary to resolve quite a complex pattern of the overlapping fragments generated by the various enzymes. Several other techniques are used in

Figure 2.3 A restriction map can be constructed by relating the A-fragments and B-fragments through the overlaps seen with double digest fragments.



The A-2100 fragment identifies 2 A-cutting sites & 1 B-cutting site

Overlap with the B-2500 fragment extends the map to the right

Overlap with B-1200 extends map to left; overlap with A-1400 extends map to right

Left end of map is identified by A-1000 & B-1200 fragments; right end is identified by A-500 and B-1300

conjunction with comparison of fragments, including **end-labeling,** in which the ends of the DNA molecule are labeled with a radioactive phosphate (certain enzymes can add phosphate moieties specifically to 5′ or to 3′ ends). **Figure 2.4** shows that this allows the fragments containing the ends to be identified directly by their radioactive label. So in the fragment A preparation, A-1000 and A-500 would be placed immediately at opposite ends of the map; similarly, fragments B-1200 and B-1300 would be identified as ends.

A restriction map of the entire 5000 bp region that was constructed in Figures 2.1–2.3 is recapitulated in its more usual form in **Figure 2.5.** The map shows the positions at which particular restriction enzymes cut DNA; the distances between the sites of cutting are measured in base pairs. *So the DNA is divided into a series of regions of defined lengths that lie between sites recognized by the restriction enzymes.* An important feature is that a restriction map can be obtained for any sequence of DNA, *irrespective of whether mutations have been identified in it,* or, indeed, whether we have any knowledge of its function.

Figure 2.4 When restriction fragments are identified by their possession of a labeled end, each fragment directly shows the distance of a cutting site from the end. Successive fragments increase in length by the distance between adjacent restriction sites.

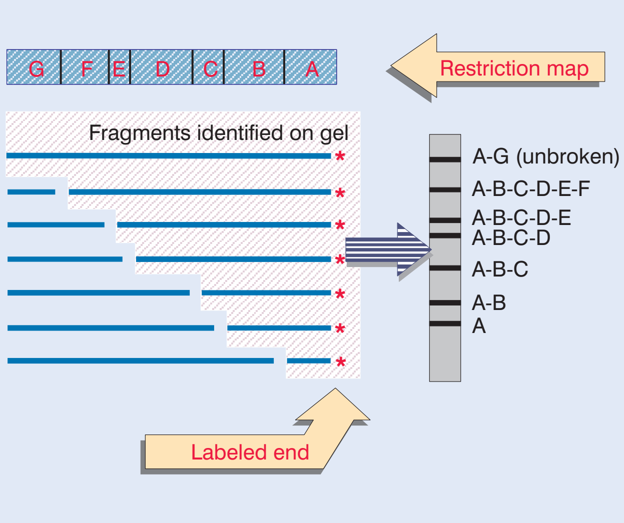
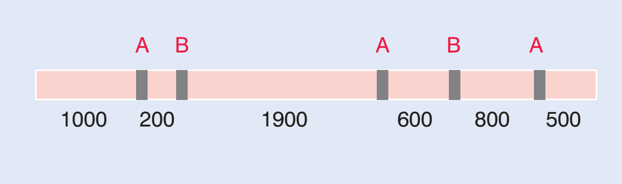


Figure 2.5 A restriction map is a linear sequence of sites separated by defined distances on DNA. The map identifies the sites cleaved by enzymes A and B, as defined by the individual fragments produced by the single and double digests.



How variable are individual genomes?

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.

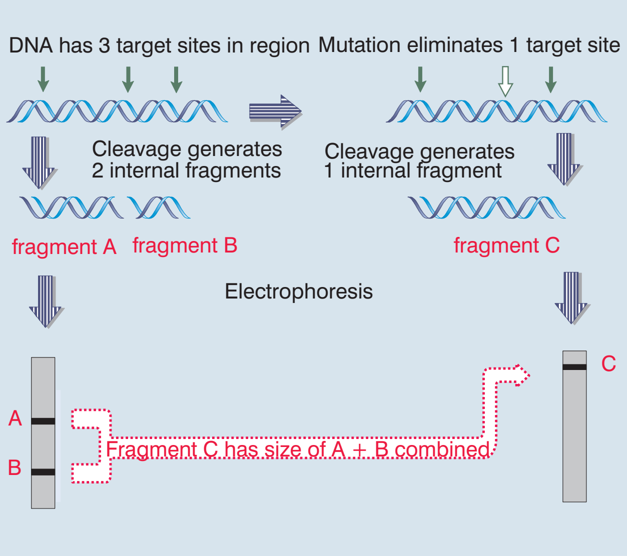
What is the basis for the polymorphism among the mutant alleles? They possess different mutations that alter the protein function, thus producing changes in phenotype. If we compare the restriction maps or the DNA sequences of these alleles, they too will be polymorphic in the sense that each map or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

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 2.6** 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.

Figure 2.6 A point mutation that affects a restriction site is detected by a difference in restriction fragments.


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 has size of A + B combined

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**). 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 2.7** shows a pedigree of a restriction polymorphism followed through three generations. It displays Mendelian segregation at the level of DNA marker fragments.

Recombination frequency can be measured between a restriction marker and a visible phenotypic marker as illustrated in **Figure 2.8.** So a genetic map can include both genotypic and phenotypic markers.

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

Figure 2.7 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.

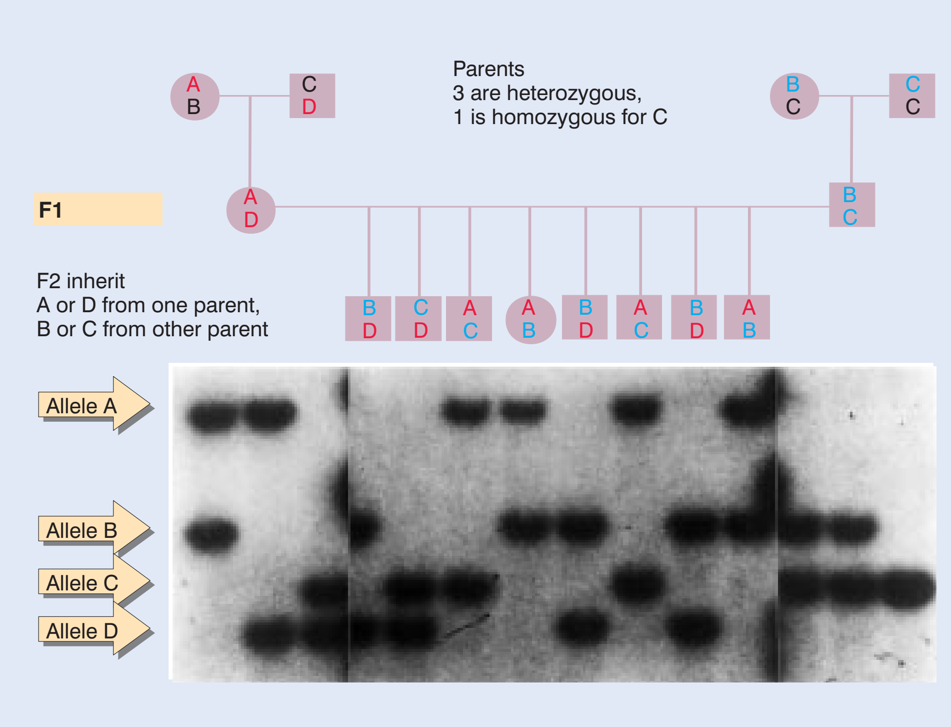
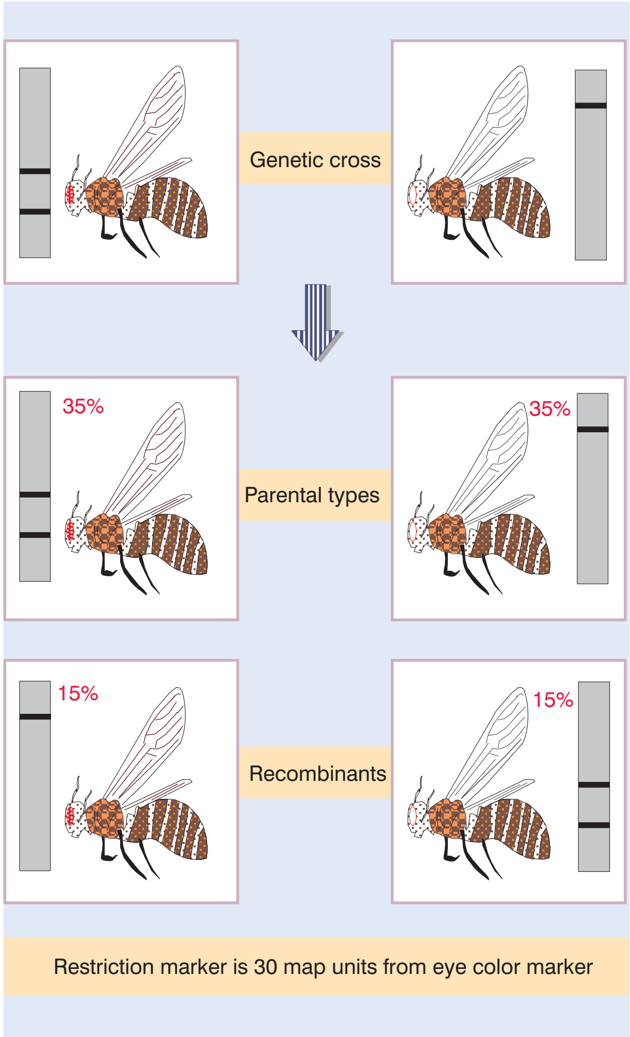


Figure 2.8 A restriction polymorphism can be used as a genetic marker to measure recombination distance from a phenotypic marker (such as eye color). The figure simplifies the situation by showing only the DNA bands corresponding to a haploid genome and omitting the bands corresponding to the allele of the other genome in a diploid.



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 2.9.** 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.

The identification of such a marker has two important consequences:

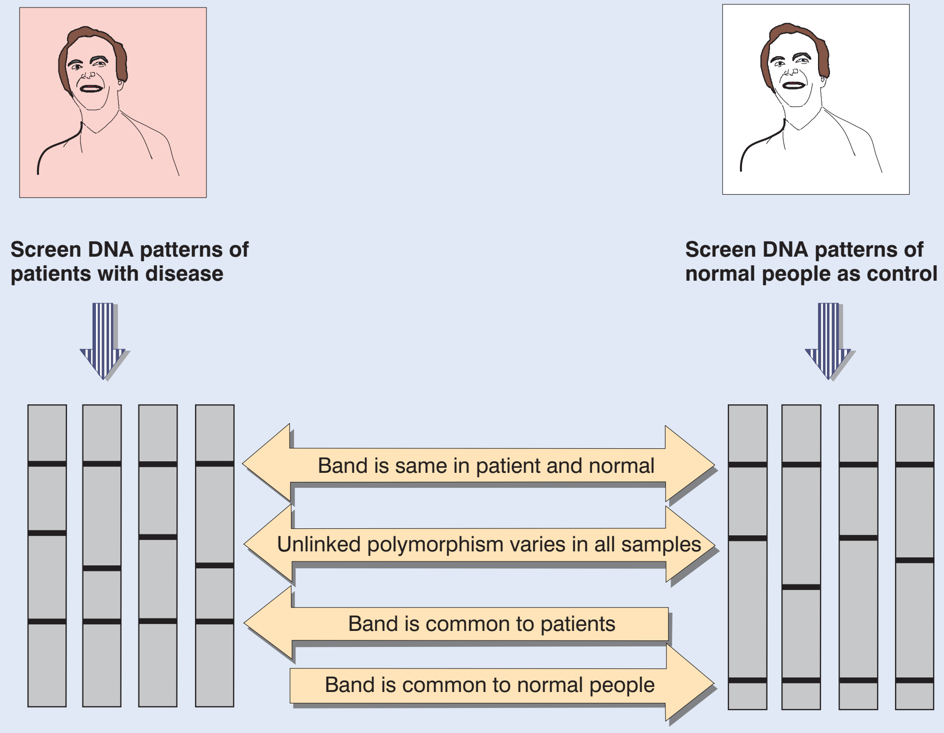
- It may offer a diagnostic procedure for detecting the disease. Some of the human diseases that are genetically well characterized but ill defined in molecular terms cannot be easily diagnosed. If a restriction marker is reliably linked to the phenotype, then its presence can be used to diagnose the disease, either at a prenatal stage or subsequently.
- 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.

RFLPs occur frequently enough in the human genome to be useful for genetic mapping. If allelic sequences are compared between any two individual chromosomes, differences in individual base pairs occur at a frequency of >1 per 1000 bp. Those base changes that affect restriction sites can be detected as RFLPs.

Once an RFLP has been assigned to a linkage group, it can be placed in position on the genetic map, and map distances to its flanking markers determined. An effort to map RFLPs in man and mouse has led to the construction of linkage maps for both genomes. The human map contains >5000 markers separated by an average distance of 1.6 cM (1–2 Mb); the mouse map has >7000 markers with an average spacing of ~0.2 cM (~200 kb). Any unknown site can be tested for linkage to these sites and by this means rapidly placed on to the map.

The frequency of polymorphism means that every individual has a unique constellation of restriction sites. 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 Chapter 24). 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.

Figure 2.9 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.


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

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.** We discuss in more detail in Chapter 4 the use of particularly variable "minisatellite" sequences for mapping in the human genome.

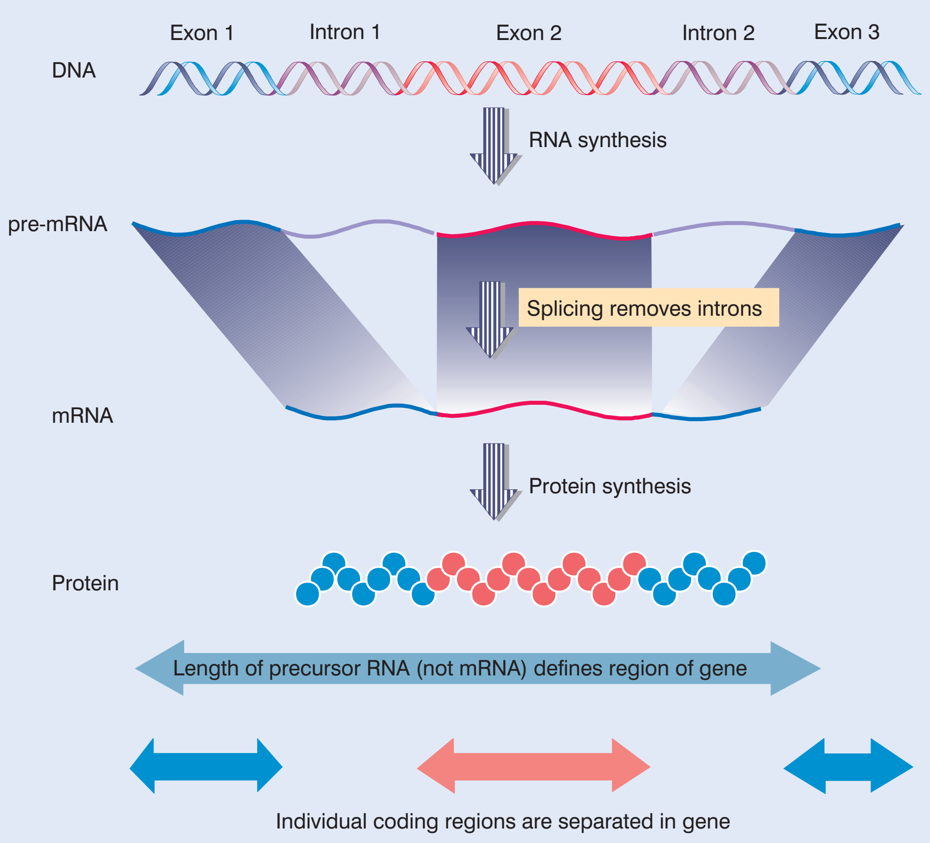
Eukaryotic genes are often interrupted

UNTIL eukaryotic genes were characterized by molecular mapping, we assumed that they would have the same organization as prokaryotic genes. We therefore expected the gene to consist of a length of DNA that is colinear with the protein. But a comparison between the structure of DNA and the corresponding mRNA shows a discrepancy in many cases. The mRNA always includes a nucleotide sequence that corresponds exactly with the protein product according to the rules of the genetic code. *But the gene includes additional sequences that lie within the coding region, interrupting the sequence that represents the protein.*

The sequences of DNA comprising an interrupted gene are divided into the two categories depicted in **Figure 2.10:**

- The **exons** are the sequences represented in the mature RNA. By definition, a gene starts and ends with exons, corresponding to the 5′ and 3′ ends of the RNA.
- The **introns** are the intervening sequences that are

Figure 2.10 Interrupted genes are expressed via a precursor RNA. Introns are removed when the exons are spliced together. The mRNA has only the sequences of the exons.



removed when the primary transcript is processed to give the mature RNA.

The expression of interrupted genes requires an additional step that does not occur for uninterrupted genes. The DNA gives rise to an RNA copy (a **transcript**) that exactly represents the genome sequence. But this RNA is only a precursor; it cannot be used for producing protein. First the introns must be removed from the RNA to give a messenger RNA that consists only of the series of exons. This process is called **RNA splicing.** It involves a precise deletion of an intron from the primary transcript; the ends of the RNA on either side are joined to form a covalently intact molecule. We discuss the mechanisms and regulation of splicing in Chapter 22.

The **structural gene** comprises the region in the genome between points corresponding to the 5′ and 3′ terminal bases of mature mRNA. We know that transcription starts at the 5′ end of the mRNA, but probably it extends beyond the 3′end, which is generated by cleavage of the RNA (see Chapter 22). The definition of the gene can be expanded to include regulatory regions on both sides of the gene that are required for initiating and (sometimes) terminating gene expression.

How does this change our view of the gene? Following splicing, the exons are always joined together in the same order in which they lie in DNA. So the colinearity of gene and protein is maintained between the individual exons and the corresponding parts of the protein chain. The *order* of mutations in the gene remains the same as the order of amino acid replacements in the protein. But the *distances* in the gene do not correspond at all with the distances in the protein. The length of the gene is defined by the length of the initial (precursor) RNA instead of by the length of the messenger RNA.

All the exons are represented on the same molecule of RNA, and their splicing together occurs only as an *intra*molecular reaction. There is usually no joining of exons carried by *different* RNA molecules, so the mechanism excludes any splicing together of sequences representing different alleles. So mutations located in different exons of a gene cannot complement one another; thus they continue to be defined as members of the same complementation group.

What are the effects of mutations in the introns? Since the introns are not part of the messenger RNA, mutations in them cannot directly affect protein structure. However, they can prevent the production of the messenger RNA—for example, by inhibiting the splicing together of exons. A mutation of this sort acts only on the allele that carries it. So it fails to complement any other mutation in that allele, and constitutes part of the same complementation group as the exons.

Eukaryotic genes are not necessarily interrupted. Some correspond directly with the protein product in the same manner as prokaryotic genes. In yeast, most genes are uninterrupted. In higher eukaryotes, most genes are interrupted; and the introns are usually much longer than exons, creating genes that are very much larger than their coding regions.

Organization of interrupted genes may be conserved

WHEN a gene is uninterrupted, the restriction map of its DNA corresponds exactly with the map of its mRNA (obtained by characterizing a cDNA reverse transcript).

When a gene possesses an intron, the map at each end of the gene corresponds with the map at each end of the message sequence. But within the gene, the maps diverge, because additional regions are found in the gene, but are not represented in the message. Each such region corresponds to an intron. The example of **Figure 2.11** compares the restriction maps of a β-globin gene and mRNA. There are two introns. Each intron contains a series of restriction sites that are absent from the cDNA. The pattern of restriction sites in the exons is the same in both the cDNA and the gene.

Ultimately a comparison of the nucleotide sequences of the genomic and cDNA clones precisely defines the introns. Resolution at the sequence level is necessary before we can be sure that all the segments of the gene have been identified. Short introns or exons can be missed in restriction maps if they do not happen to contain an appropriate restriction site. (An intron may pass unnoticed if it lies within a long exon, and an exon that is <50 bp long may fail to hybridize with the cDNA probe, and can therefore pass unnoticed within the introns that flank it.) But a sequence comparison is unambiguous. As indicated in **Figure 2.12,** an intron that lies within a coding region usually interrupts the integrity of the reading frame, but an intact reading frame is found in the cDNA sequence.

No particular rhyme or reason yet has been discerned in the extremely varied structures of eukaryotic

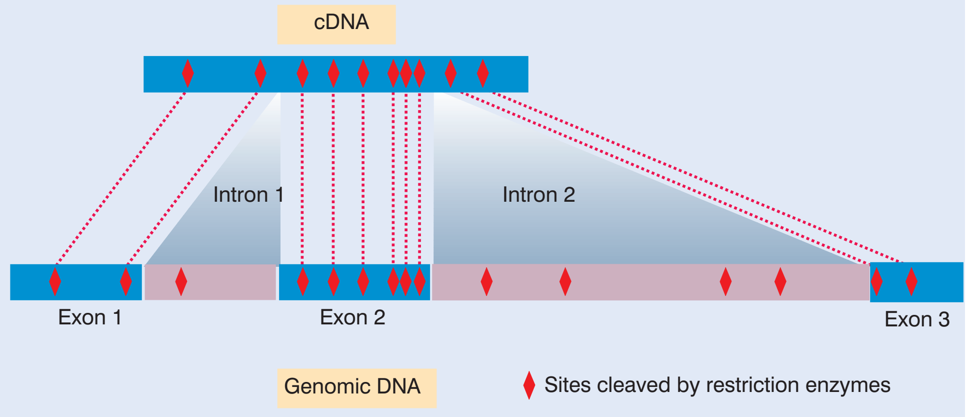


Figure 2.11
Comparison of the restriction maps of cDNA and genomic DNA for mouse -globin shows that the gene has two additional regions not present in the cDNA. The other regions can be aligned exactly between cDNA and gene.

Figure 2.12 An intron is a sequence present in the gene but absent from the mRNA (here shown in terms of the cDNA sequence). The reading frame is indicated by the alternating open and shaded blocks; note that all three possible reading frames are blocked by termination codons in the intron.

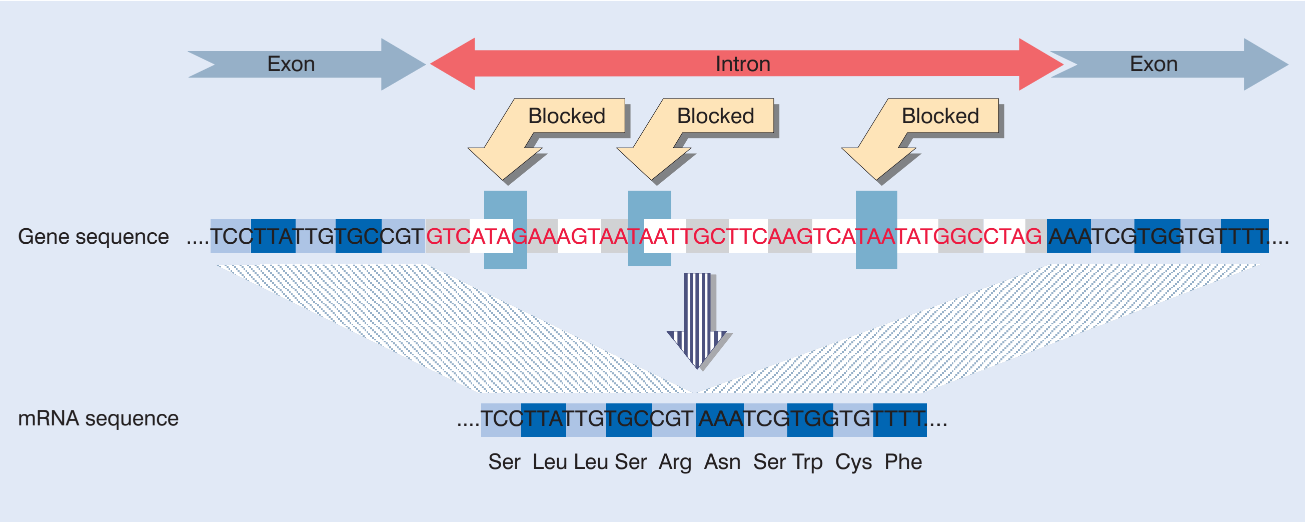
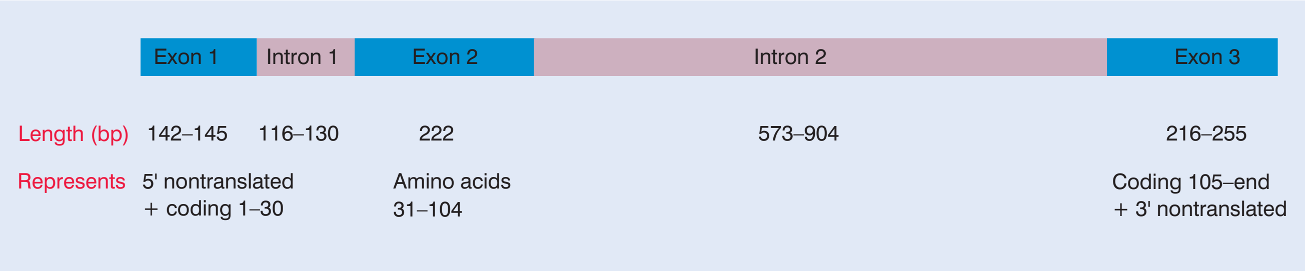

Exon
Intron
Exon
Blocked
Blocked
Blocked
Gene sequence
....TCCTTATTGTGCCGTGTCATAGAAAGTAATAATTGCTTCAAGTCATAATATGGCCTAG AAATCGTGGTGTTTT....
mRNA sequence
....TCCTTATTGTGCCGT AAATCGTGGTGTTTT....
Ser Leu Leu Ser Arg Asn Ser Trp Cys Phe

Figure 2.13 All functional globin genes have an interrupted structure with three exons. The lengths indicated in the figure apply to the mammalian -globin genes.



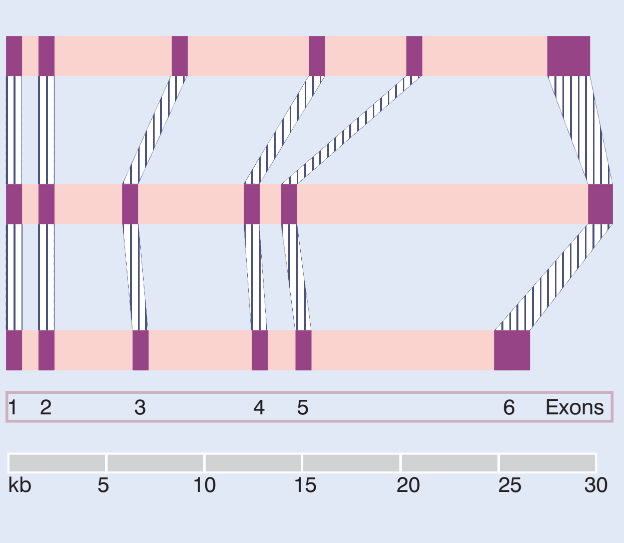
genes. Some genes are uninterrupted, so that the genomic sequence is colinear with that of the mRNA. Most higher eukaryotic genes are interrupted, but the introns vary enormously in both number and size. Introns of nuclear genes generally have termination codons in all reading frames, and have *no coding function*.

All classes of genes may be interrupted: nuclear genes coding for proteins, nucleolar genes coding for rRNA, and genes coding for tRNA. Interruptions are also found in mitochondrial genes in lower eukaryotes, and in chloroplast genes. Interrupted genes do not appear to be excluded from any class of eukaryotes, and have been found in bacteria and bacteriophages, although they are extremely rare in prokaryotic genomes.

Some interrupted genes possess only one or a few introns. The globin genes provide an extensively studied example (see Chapter 4). The two general types of globin gene, α and β, share a common type of structure. The consistency of the organization of mammalian globin genes is evident from the structure of the "generic" globin gene summarized in **Figure 2.13.**

Interruptions occur at homologous positions (relative to the coding sequence) in all known active globin genes, including those of mammals, birds, and frogs. The first intron is always fairly short, and the second is usually longer, but the actual lengths can vary. Most of the variation in overall lengths between different globin genes results from the variation in the second intron. In the mouse, the second intron in the α-globin gene is only 150 bp long, so the overall length of the gene is 850 bp, compared with the 1382 bp of the major β-globin gene. So the variation in length of the genes is

Figure 2.14 Mammalian genes for DHFR have the same relative organization of rather short exons and very long introns, but vary extensively in the lengths of corresponding introns.



much greater than the range of lengths of the mRNAs (α-globin mRNA = 585 bases, β-globin mRNA = 620 bases).

The example of DHFR, a somewhat larger gene, is shown in **Figure 2.14.** The mammalian DHFR (dihydrofolate reductase) gene is organized into six exons that correspond to the 2000 base mRNA. But they extend over a much greater length of DNA because the introns are exceedingly long. In three mammals the exons remain essentially the same, and the relative positions of the introns are unaltered, but the lengths of individual introns vary extensively, resulting in a variation in the length of the gene from 25 to 31 kb.

The globin and DHFR genes present examples of a general phenomenon: *genes that are related by evolution have related organizations, with conservation of the positions of (at least some) of the introns. Variations in the lengths of the genes are primarily determined by the lengths of the introns.*

Exon sequences are conserved but introns vary

Is a structural gene unique in its genome? The answer can be ambiguous. The entire length of the gene is unique as such, but its exons often are related to those of other genes. As a general rule, when two genes are related, the relationship between their exons is closer than the relationship between the introns. In an extreme case, the exons of two genes may code for the same protein sequence, but the introns may be different. This implies that the two genes originated by a duplication of some common ancestral gene. Then differences accumulated between the copies, but they were restricted in the exons by the need to code for protein functions.

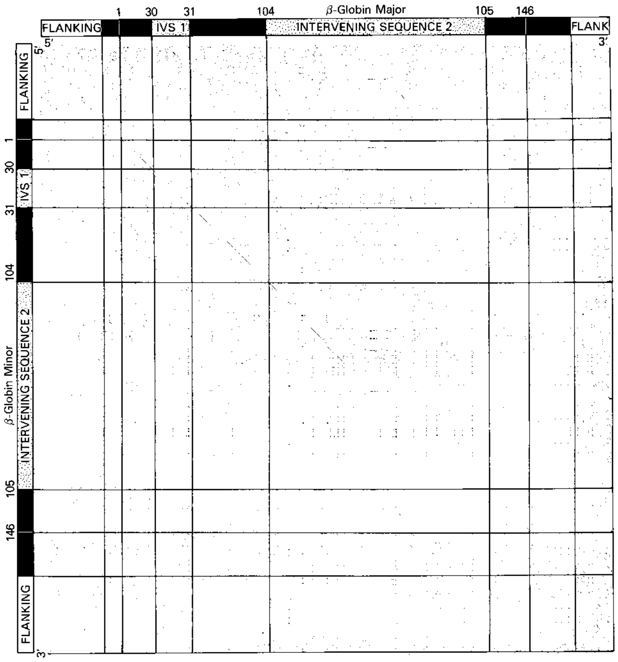
As we see later when we consider the evolution of the gene, exons can be considered as basic building blocks that are assembled in various combinations. A gene may have some exons that are related to exons of another gene, but the other exons may be unrelated. Usually the introns are not related at all in such cases. Such genes may arise by duplication and translocation of individual exons.

The relationship between two genes can be plotted in the form of the dot matrix comparison of **Figure 2.15.** A dot is placed to indicate each position at which the same sequence is found in each gene. The dots form a line at an angle of 45° if two sequences are identical. The line is broken by regions that lack similarity, and it is displaced laterally or vertically by deletions or insertions in one sequence relative to the other.

When the two β-globin genes of the mouse are compared, such a line extends through the three exons and through the small intron. The line peters out in the flanking regions and in the large intron. This is a typical pattern, in which coding sequences are well related, the relationship can extend beyond the boundaries of the exons, but it is lost in longer introns and the regions on either side of the gene.

The overall degree of divergence between two exons is related to the differences between the proteins. It is caused mostly by base substitutions. In the translated regions, the exons are under the constraint of needing to code for amino acid sequences, so they are limited in their potential to change sequence. Many of the changes do not affect codon meanings, because they

Figure 2.15 The sequences of the mouse ^{maj} and ^{min} globin genes are closely related in coding regions, but differ in the flanking regions and large intron. Data kindly provided by Philip Leder.



change one codon into another that represents the same amino acid. Changes occur more freely in nontranslated regions (corresponding to the 5′ leader and 3′ trailer of the mRNA).

In corresponding introns, the pattern of divergence involves both changes in size (due to deletions and insertions) and base substitutions. Introns evolve much more rapidly than exons. When a gene is compared in different species, sometimes the exons are homologous, while the introns have diverged so much that corresponding sequences cannot be recognized.

Mutations occur at the same rate in both exons and introns, but are removed more effectively from the exons by adverse selection. However, in the absence of the constraints imposed by a coding function, an intron is able quite freely to accumulate point substitutions and other changes. These changes imply that the intron does not have a sequence-specific function. Whether its presence is at all necessary for gene function is not clear.

Genes can be isolated by the conservation of exons

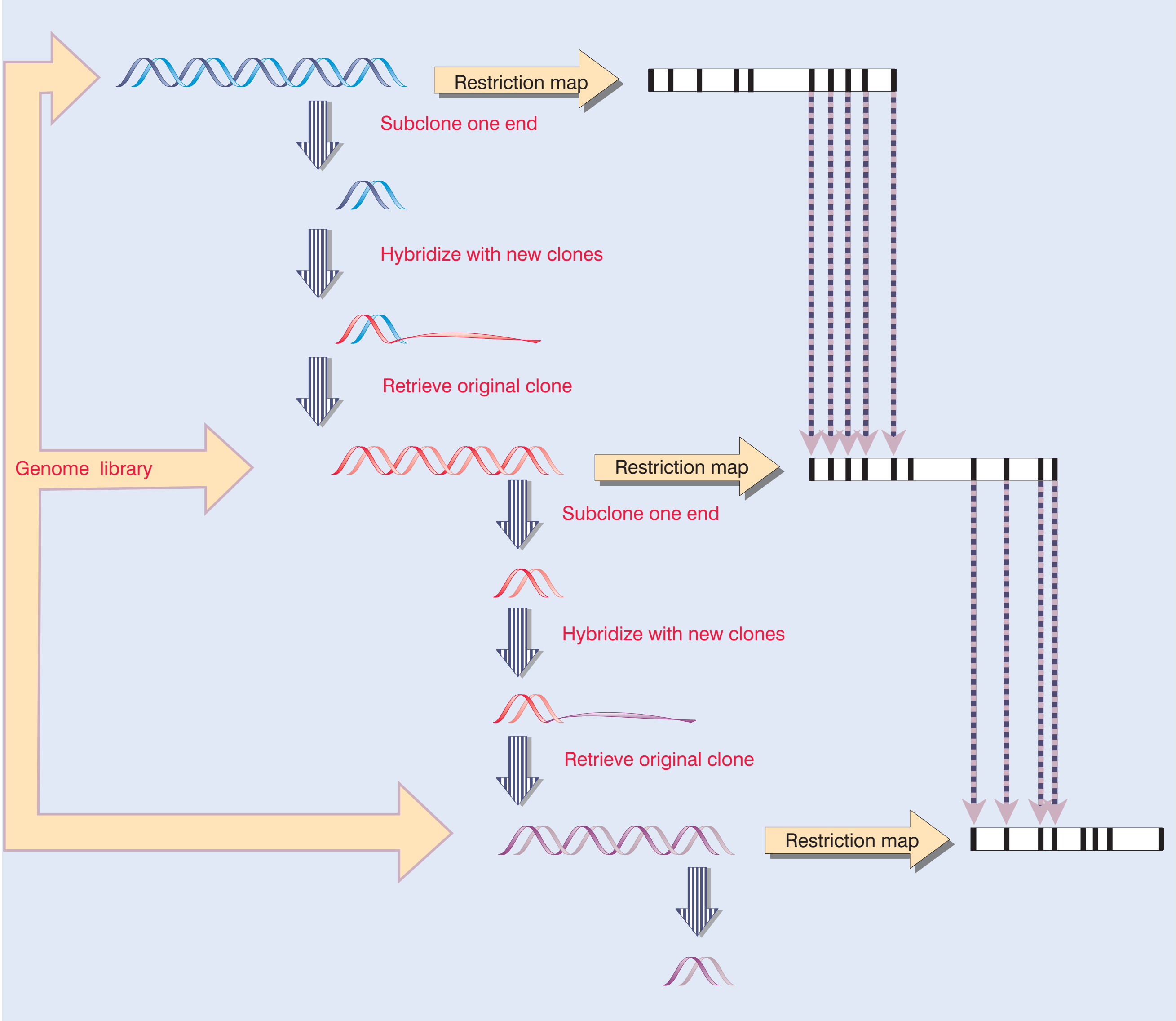
Some major approaches to identifying genes are based on the contrast between the conservation of exons and the variation of introns. In a region containing a gene whose function has been conserved among a range of species, the sequence representing the protein should have two distinctive properties: it must of course have an open reading frame; and it is likely to have a related sequence in other species. These features can be used to isolate genes.

Suppose we know by genetic data that a particular genetic trait is located in a given chromosomal region. If we lack knowledge about the nature of the gene product, how are we to identify the gene in a region that may be (for example) >1 Mb?

We start with a clone that lies in the general vicinity of this region and then we "walk" through the region by identifying overlapping clones from a library. As shown **in Figure 2.16,** a subfragment from one end of the first clone is used to isolate clones that extend farther along the chromosome. These clones in turn are used to isolate the next set. In each cycle, a new clone is selected because its restriction map coincides at one end with the end of the previous clone, but at the other end has new material. It is possible to walk for hundreds of kb, typically at a rate of >100 kb per month. Chromosome walking allows large contiguous regions of the chromosome to be represented in a library of clones.

Of course, it becomes much easier to identify a particular gene once the sequence of a chromosome has been obtained. This can be done either by sequencing a contiguous series of clones that were obtained by walking, or by relating the clones by other means (such as direct comparisons of sequences). With a sequence in

Figure 2.16 Chromosome walking is accomplished by successive hybridizations between overlapping genomic clones.



hand, a gene can be identified by comparison with either the RNA or protein product, or by the identification of a mutation in the sequence.

In the meantime, until full sequence information is available, a heroic approach that has proved successful with some genes of medical importance is to screen relatively short fragments from the region for the two properties expected of a conserved gene. First we seek to identify fragments that cross-hybridize with the genomes of other species. Then we examine these fragments for open reading frames.

The first criterion is applied by performing a **zoo blot.** We use short fragments from our chromosome walk as (radioactive) probes to test for related DNA from a variety of species by Southern blotting. If we find hybridizing fragments in several species related to that of the probe—the probe is usually human—the probe becomes a candidate for an exon of the gene.

The candidates are sequenced and, if they contain open reading frames, are used to isolate surrounding genomic regions. If these appear to be part of an exon, we may then use them to identify the entire gene, to isolate the corresponding cDNA or mRNA, and ultimately to identify the protein.

This approach is valuable for genes whose existence is implied by genetics, but whose nature is unknown.

One example is the gene *zfy* located on the human Y chromosome. **Figure 2.17** shows a zoo blot using a probe from this region. It hybridizes specifically with sex chromosomes of mammals and also with other species. It contains an open reading frame, which identifies a conserved gene.

This approach is especially important when the target gene is spread out because it has many large introns. This proved to be the case with Duchenne muscular dystrophy (DMD), a degenerative disorder of muscle, which is X-linked and affects 1 in 3500 of human male births. The steps in identifying the gene are summarized in **Figure 2.18.**

Linkage analysis localized the DMD locus to chromosomal band Xp21. Patients with the disease often have chromosomal rearrangements involving this band. By comparing the ability of X-linked DNA probes to hybridize with DNA from patients and with normal DNA, cloned fragments were obtained that correspond to the region that was rearranged or deleted in patients' DNA.

A chromosomal walk was used to construct a restriction map of the region on either side of the probe, covering a region of >100 kb. Analysis of the DNA from a series of patients identified large deletions in this region, extending in either direction. The most telling deletion is one contained entirely within the region,

Figure 2.17 A zoo blot with a probe from the human Y chromosomal gene *zfy* identifies cross-hybridizing fragments on the sex chromosomes of other mammals and birds. There is one reacting fragment on the Y chromosome and another on the X chromosome. Data kindly provided by David Page.

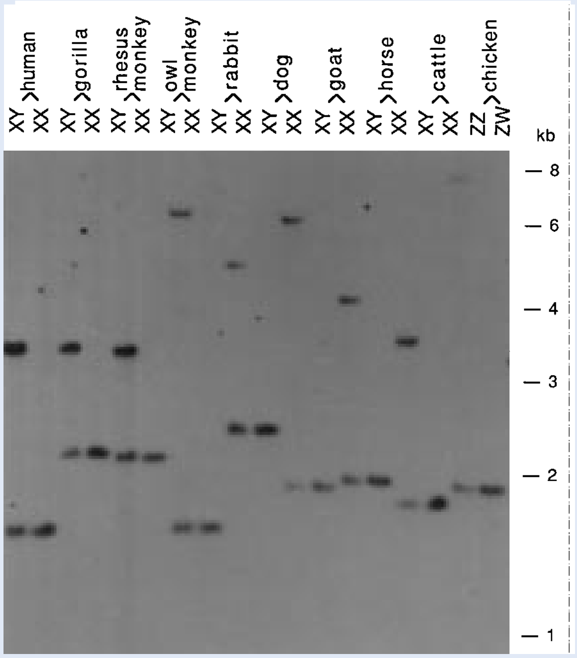
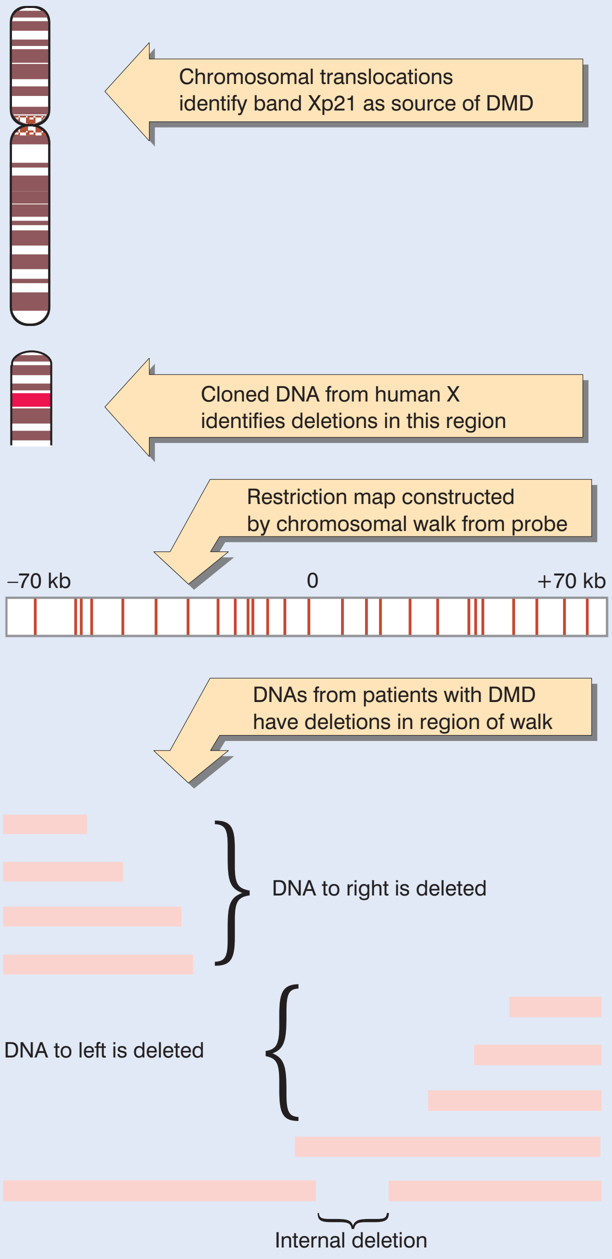


Figure 2.18 The gene involved in Duchenne muscular dystrophy has been tracked down by chromosome mapping and walking to a region in which deletions can be identified with the occurrence of the disease.


Chromosomal translocations identify band Xp21 as source of DMD
Cloned DNA from human X identifies deletions in this region
Restriction map constructed by chromosomal walk from probe
–70 kb
0
+70 kb
DNAs from patients with DMD have deletions in region of walk
DNA to right is deleted
DNA to left is deleted
Internal deletion

since this delineates a segment that must be important in gene function and indicates that the gene, or at least part of it, lies in this region.

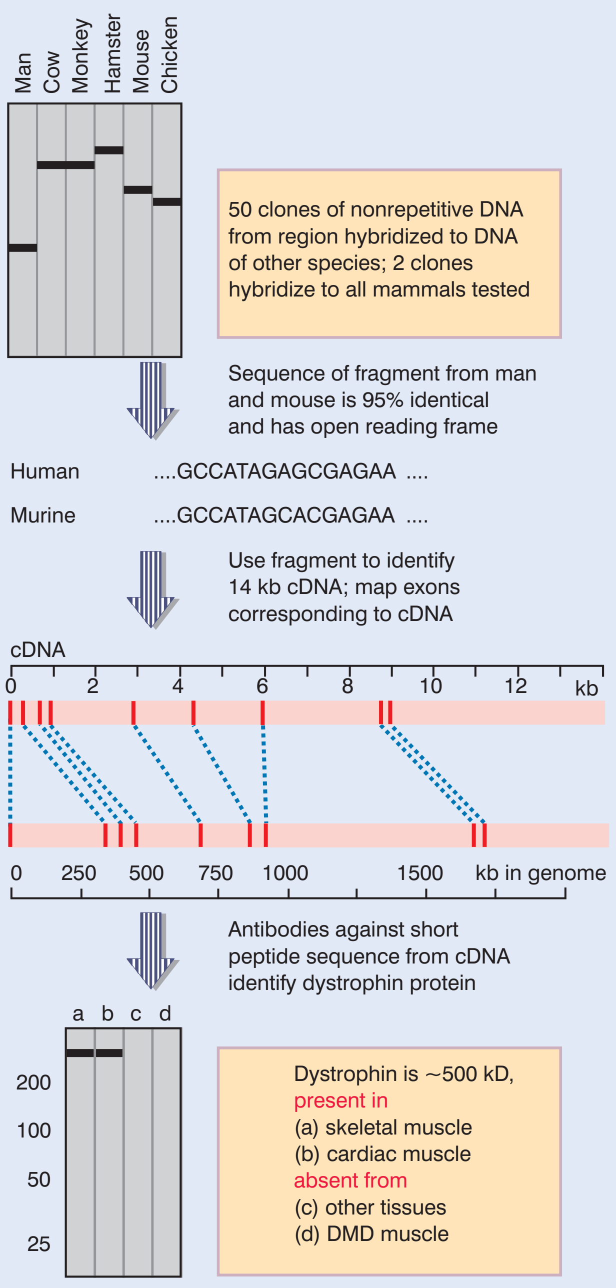
Having now come into the region of the gene, we need to identify its exons and introns. A zoo blot identified fragments that cross-hybridize with the mouse X chromosome and with other mammalian DNAs. As summarized in **Figure 2.19,** these were scrutinized for open reading frames and the sequences typical of exon-intron junctions. Fragments that met these criteria were used as probes to identify homologous sequences in a cDNA library prepared from muscle mRNA.

The cDNA corresponding to the gene identifies an unusually large mRNA, ~14 kb. Hybridization back to the genome shows that the mRNA is represented in >60 exons, which are spread over ~2000 kb of DNA. This makes DMD the longest gene identified; in fact, it is 10× longer than any other known gene.

The gene codes for a protein of ~500 kD, called dystrophin, which is a component of muscle, present in rather low amounts. All patients with the disease have deletions at this locus, and lack (or have defective) dystrophin.

Another technique that allows genomic fragments to be scanned rapidly for the presence of exons is called **exon trapping. Figure 2.20** shows that it starts with a vector that contains a strong promoter, and has a single intron between two exons. When this vector is transfected into cells, its transcription generates large amounts of an RNA containing the sequences of the two exons. A restriction cloning site lies within the intron, and is used to insert genomic fragments from a region of interest. If a fragment does not contain an exon, there is no change in the splicing pattern, and the RNA contains only the same sequences as the parental vector. But if the genomic fragment contains an exon flanked by two partial intron sequences, the splicing sites on either side of this exon are recognized, and the sequence of the exon is inserted into the RNA between the two exons of the vector. This can be detected readily by reverse transcribing the cytoplasmic RNA into cDNA, and using PCR to amplify the sequences between the two exons of the vector. So the appearance in the amplified population of sequences from the genomic fragment indicates that an exon has been trapped. Because introns are usually large and exons are small in animal cells, there is a high probability that a random piece of genomic DNA will contain the required structure of an exon surrounded by partial introns.

Figure 2.19 The Duchene muscular dystrophy gene has been characterized by zoo blotting, cDNA hybridization, genomic hybridization, and identification of the protein.



50 clones of nonrepetitive DNA from region hybridized to DNA of other species; 2 clones hybridize to all mammals tested

Sequence of fragment from man and mouse is 95% identical and has open reading frame

Human ....GCCATAGAGCGAGAA ....

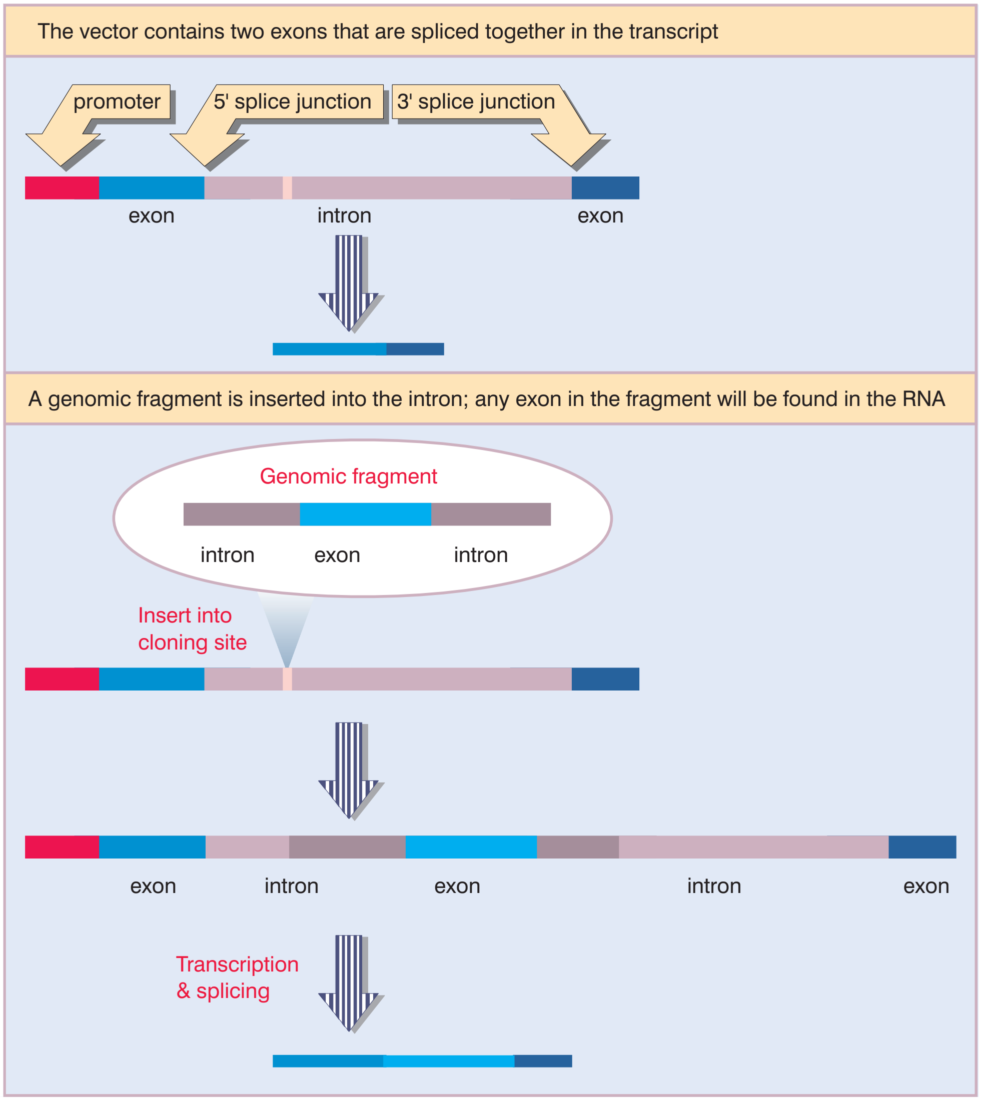
Murine ....GCCATAGCACGAGAA ....

Use fragment to identify 14 kb cDNA; map exons corresponding to cDNA

Antibodies against short peptide sequence from cDNA identify dystrophin protein

Dystrophin is ~500 kD,
present in
(a) skeletal muscle
(b) cardiac muscle
absent from
(c) other tissues
(d) DMD muscle

Figure 2.20 A special splicing vector is used for exon trapping. If an exon is present in the genomic fragment, its sequence will be recovered in the cytoplasmic RNA, but if the genomic fragment consists solely of an intron, it will be spliced out and lost.


The vector contains two exons that are spliced together in the transcript
promoter
5' splice junction
3' splice junction
exon
intron
exon
A genomic fragment is inserted into the intron; any exon in the fragment will be found in the RNA
Genomic fragment
intron
exon
intron
Insert into cloning site
exon
intron
exon
intron
exon
Transcription & splicing

Genes show a wide distribution of sizes

The existence of interrupted genes makes it evident that the gene can be much larger than the unit that codes for protein. As genome size increases, the tendency is for introns to become rather large, while exons remain quite small.

Figure 2.21 shows that the exons coding for stretches of protein tend to be fairly small relative to the size of the gene. Most code for less than 100 amino acids (often less than 50 in vertebrates), and the general distribution fits well with the idea that genes have evolved by the slow addition of units that code for small, individual domains of proteins (see later). There is no very significant difference in the sizes of exons in different types of organism, except perhaps for an apparent absence of larger exons in the vertebrates. (The peak of exons coding for >300 amino acids in fungi and *Drosophila* mostly represents the presence of uninterrupted genes, that is, genes that consist of one exon.) There are some much larger exons coding for untranslated 5′ and 3′ regions (not included in the figure).

Figure 2.21 Exons coding for proteins are usually short.

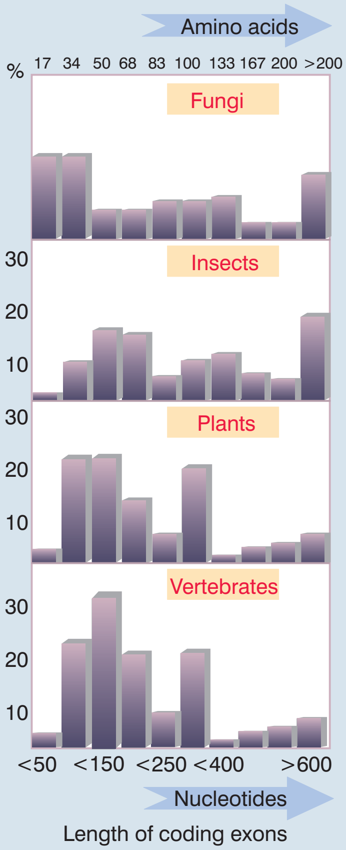


Figure 2.22 Introns in vertebrate genes range from very short to very long.

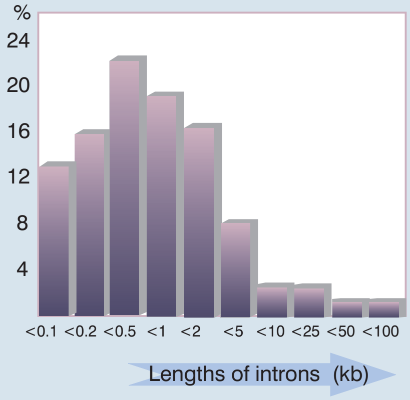


Figure 2.22 shows that introns are longer than exons. Their size distribution extends from approximately the same size as the exons (<200 bp) to lengths measured in 10s of kbs, and extending up to 50–60 kb in extreme cases.

Figure 2.23 shows the overall organization of genes in yeasts, insects, and mammals. In *S. cerevisiae,* the great majority of genes (>96%) are not interrupted, and those that have exons usually remain reasonably compact. There are virtually no *S. cerevisiae* genes with more than four exons.

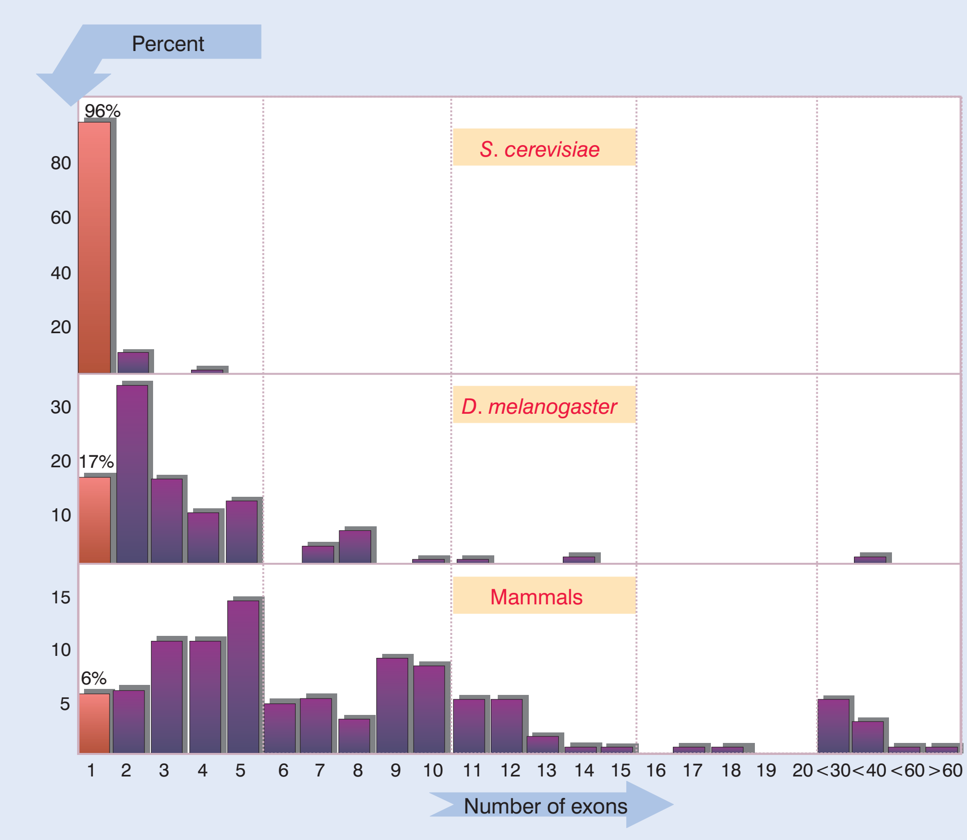
In insects and mammals, the situation is reversed. Only a few genes have uninterrupted coding sequences (6% in mammals). Insect genes tend to have a fairly small number of exons, typically fewer than 10. Mammalian genes are split into more pieces, and some have several 10s of exons; ~50% of mammalian genes have >10 introns.

If we now examine the consequences of this type of organization for the overall size of the gene, we see in **Figure 2.24** that there is a striking difference between yeast and the higher eukaryotes. The average yeast gene is 1.4 kb long, and very few are longer than 5 kb. By contrast, relatively few genes in flies or mammals are shorter than 2 kb, and many have lengths between 5 kb and 100 kb.

The switch from largely uninterrupted to largely interrupted genes occurs in the lower eukaryotes. In fungi (excepting the yeasts), the majority of genes are interrupted, but they have a relatively small number of exons (<6) and are fairly short (<5 kb). The switch to long genes occurs within the higher eukaryotes, and genes become significantly larger in the insects. Perhaps genes become large at the same point where the relationship between genome complexity and organism complexity is lost (see Figure 3.1).

Very long genes are the result of very long introns, not the result of coding for longer products. There is no correlation between gene size and mRNA size in higher eukaryotes; nor is there a good correlation between gene size and the number of exons. The size of a gene therefore depends primarily on the lengths of its individual introns. In mammals, insects, and birds, the "average" gene is approximately 5× the length of its mRNA.

Figure 2.23 Most genes are uninterrupted in yeast, but most genes are interrupted in flies and mammals. (Uninterrupted genes have only 1 exon, and are totalled in the leftmost column.)



Some DNA sequences code for more than one protein

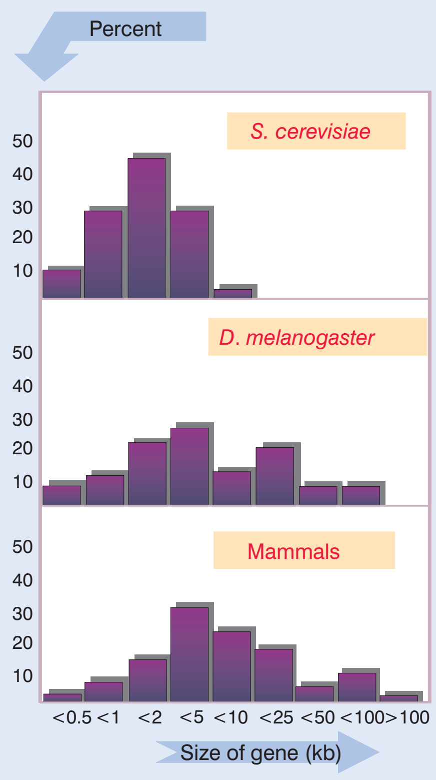
Most genes consist of a sequence of DNA that is devoted solely to the purpose of coding for one protein (although the gene may include noncoding regions at either end and introns within the coding region). However, there are some cases in which a single sequence of DNA codes for more than one protein.

Overlapping genes occur in the relatively simple situation in which one gene is part of the other. The first half (or second half) of a gene is used independently to specify a protein that represents the first (or second) half of the protein specified by the full gene. This relationship is illustrated in **Figure 2.25.** The end result is much the same as though a partial cleavage took place in the protein product to generate part-length as well as full-length forms.

Two genes overlap in a more subtle manner when the same sequence of DNA is shared between two *nonhomologous* proteins. This situation arises when the same sequence of DNA is translated in more than one reading frame. In cellular genes, a DNA sequence usually is read in only one of the three potential reading frames, but in some viral and mitochondrial genes, there is an overlap between two adjacent genes that are read in different reading frames. This situation is illustrated in **Figure 2.26.** The distance of overlap is usually relatively short, so that most of the sequence representing the protein retains a unique coding function.

In some genes, *alternative* patterns of gene expression create switches in the pathway for connecting the exons. A single gene may generate a variety of mRNA products that differ in their content of exons. The difference may be that certain exons are optional—they may be included or spliced out. Or there may be exons that are treated as mutually exclusive—one or the other is included, but not both. The alternative forms produce proteins in which one part is common while the

Figure 2.24 Yeast genes are small, but genes in flies and mammals have a dispersed distribution extending to very large sizes.



other part is different. **Figure 2.27** illustrates an example in which alternative splicing leads to the inclusion of an exon in some mRNAs, while it is left out of others. (Other types of combinations that are produced by alternative splicing are discussed in Chapter 22.)

A single type of transcript is made from the gene in Figure 2.27, but it can be spliced in either of two ways. In the first pathway, two introns are spliced out, and the three exons are joined together. In the second pathway, the second exon is not recognized. As a result, a single large intron is spliced out. This intron consists of intron 1 + exon 2 + intron 2. In effect, exon 2 has been treated in this pathway as part of the single intron. The pathways produce two proteins that are the same at their ends, but one of which has an additional sequence in the middle. So the region of DNA codes for more than one protein.

Sometimes two pathways operate simultaneously, a certain proportion of the RNA being spliced in each way; sometimes the pathways are alternatives that are expressed under different conditions, one in one cell type and one in another cell type.

In some cases, the alternative means of expression do not affect the sequence of the protein; for example, changes that affect the 5′ nontranslated leader or the 3′ nontranslated trailer may have regulatory consequences, but the same protein is made. In other cases, one exon is substituted for another, as indicated in **Figure 2.28.**

Figure 2.25 Two proteins can be generated from a single gene by starting (or terminating) expression at different points.

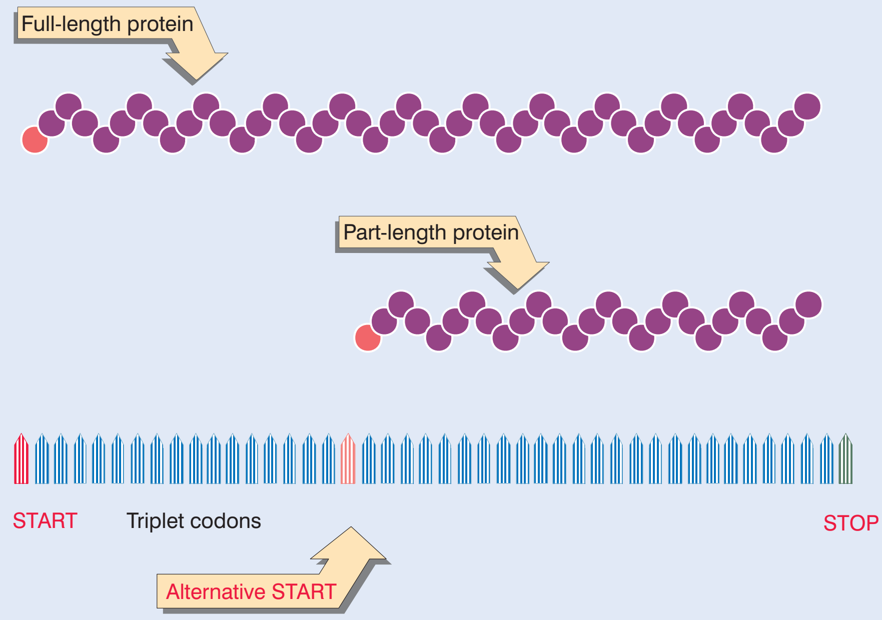


Figure 2.26 Two genes may share the same sequence by reading the DNA in different frames.

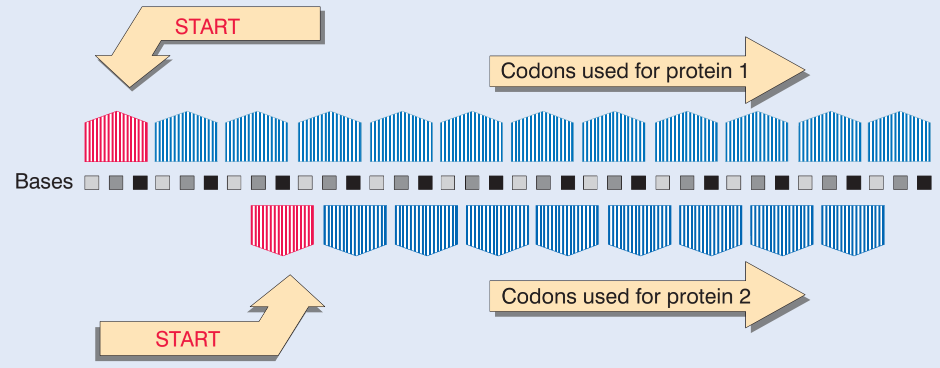


Figure 2.27 Alternative splicing uses the same pre-mRNA to generate mRNAs that have different combinations of exons.

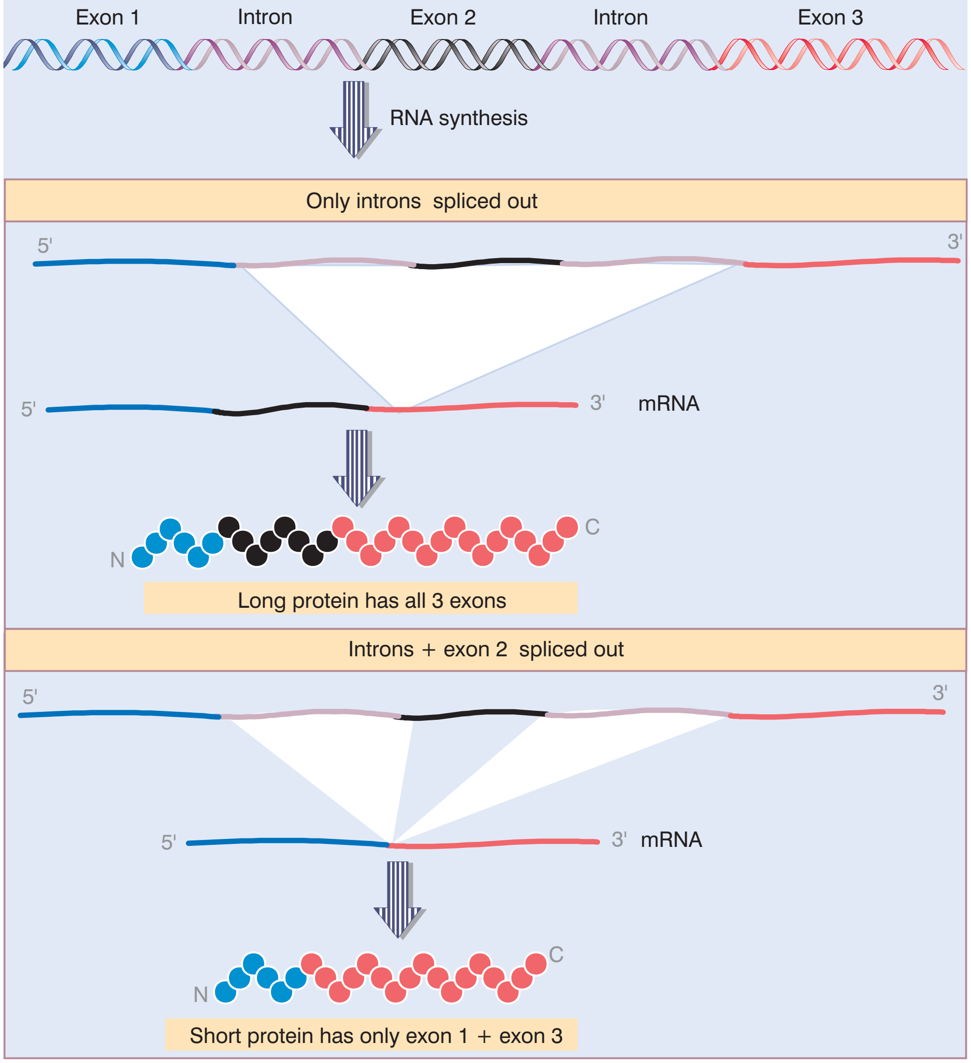
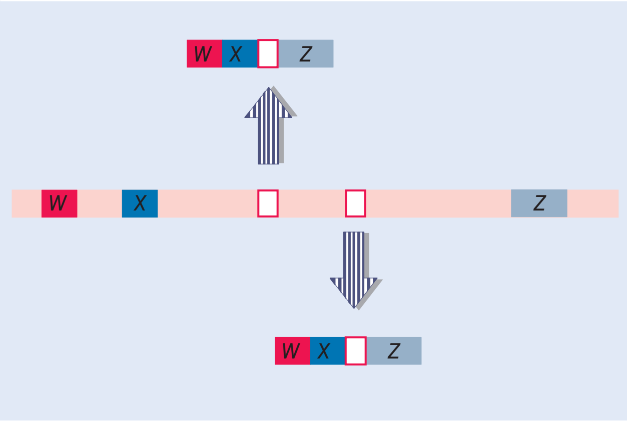


Figure 2.28 Alternative splicing generates the and variants of troponin T.



In this example, the proteins produced by the two mRNAs contain sequences that overlap extensively, but that are different within the alternatively spliced region. The 3′ half of the troponin T gene of rat muscle contains five exons, but only four are used to construct an individual mRNA. Three exons, *WXZ,* are the same in both expression patterns. However, in one pattern the α exon is spliced between *X* and *Z;* in the other pattern, the β exon is used. The α and β forms of troponin T therefore differ in the sequence of the amino acids present between sequences W and Z, depending on which of the alternative exons, α or β, is used. Either one of the α and β exons can be used to form an individual mRNA, but both cannot be used in the same mRNA.

So alternative (or differential) splicing can generate proteins with overlapping sequences from a single stretch of DNA. It is curious that the higher eukaryotic genome is extremely spacious in having large genes that are often quite dispersed, but at the same time it may make multiple products from an individual locus. It is not possible to say how many genes have alternative modes of expression, but on an anecdotal basis, the number seems to be a few percent.

How did interrupted genes evolve?

WHAT was the original form of genes that today are interrupted?

- The "introns early" model supposes that introns have always been an integral part of the gene. Genes originated as interrupted structures, and those without introns have lost them in the course of evolution.
- The "introns late" model supposes that the ancestral protein-coding units consisted of uninterrupted sequences of DNA. Introns were subsequently inserted into them.

A test of the models is to ask whether the difference between eukaryotic and prokaryotic genes can be accounted for by the acquisition of introns in the eukaryotes or by the loss of introns from the prokaryotes.

The introns early model suggests that the mosaic structure of genes is a remnant of an ancient approach to the reconstruction of genes to make novel proteins. Suppose that an early cell had a number of separate protein-coding sequences. One aspect of its evolution is likely to have been the reorganization and juxtaposition of different polypeptide units to build up new proteins.

If the protein-coding unit must be a continuous series of codons, every such reconstruction would require a precise recombination of DNA to place the two protein-coding units in register, end to end in the same reading frame. Furthermore, if this combination is not successful, the cell has been damaged, because it has lost the original protein-coding units.

But if an approximate recombination of DNA could place the two protein-coding units within the same transcription unit, splicing patterns could be tried out at the level of RNA to combine the two proteins into a single polypeptide chain. And if these combinations are not successful, the original protein-coding units remain available for further trials. Such an approach essentially allows the cell to try out controlled deletions in RNA without suffering the damaging instability that could occur from applying this procedure to DNA.

If current proteins evolved by combining ancestral proteins that were originally separate, the accretion of units is likely to have occurred sequentially over some period of time, with one exon added at a time. Can the different functions from which these genes were pieced together be seen in their present structure? In other words, can we equate particular functions of current proteins with individual exons?

In some cases, there is a clear relationship between the structures of the gene and protein. The example *par excellence* is provided by the immunoglobulin proteins, which are coded by genes in which every exon corresponds exactly with a known functional domain of the protein. **Figure 2.29** compares the structure of an immunoglobulin with its gene.

An immunoglobulin is a tetramer of two light chains and two heavy chains, which aggregate to generate a protein with several distinct domains. Light chains and heavy chains differ in structure, and there are several types of heavy chain. Each type of chain is expressed from a gene that has a series of exons corresponding with the structural domains of the protein.

In many instances, some of the exons of a gene can be identified with particular functions. In secretory proteins, the first exon, coding for the N-terminal region of the polypeptide, often specifies the signal sequence involved in membrane secretion. An example is insulin.

Sometimes the evolution of a gene involves the duplication of exons, creating an internally repetitious sequence in the protein. In chicken collagen, a 54 bp exon appears to have been multiplied many times, generating a series of exons that are either 54 bp or multiples of 54 bp in length.

Sequences held in common between genes that are related only in part may represent exons that have migrated or been recruited between genes. **Figure 2.30** summarizes the relationship between the receptor for human LDL (plasma low density lipoprotein) and other proteins.

In the center of the LDL receptor gene is a series of exons related to the exons of the gene for the precursor for EGF (epidermal growth factor). In the N-terminal part of the protein, a series of exons codes for a sequence related to the blood protein complement factor C9. So the LDL receptor gene was created by assembling *modules* for its various functions. These modules are also used in other proteins.

Figure 2.29 Immunoglobulin light chains and heavy chains are coded by genes whose structures (in their expressed forms) correspond with the distinct domains in the protein. Each protein domain corresponds to an exon; introns are numbered 1–5.

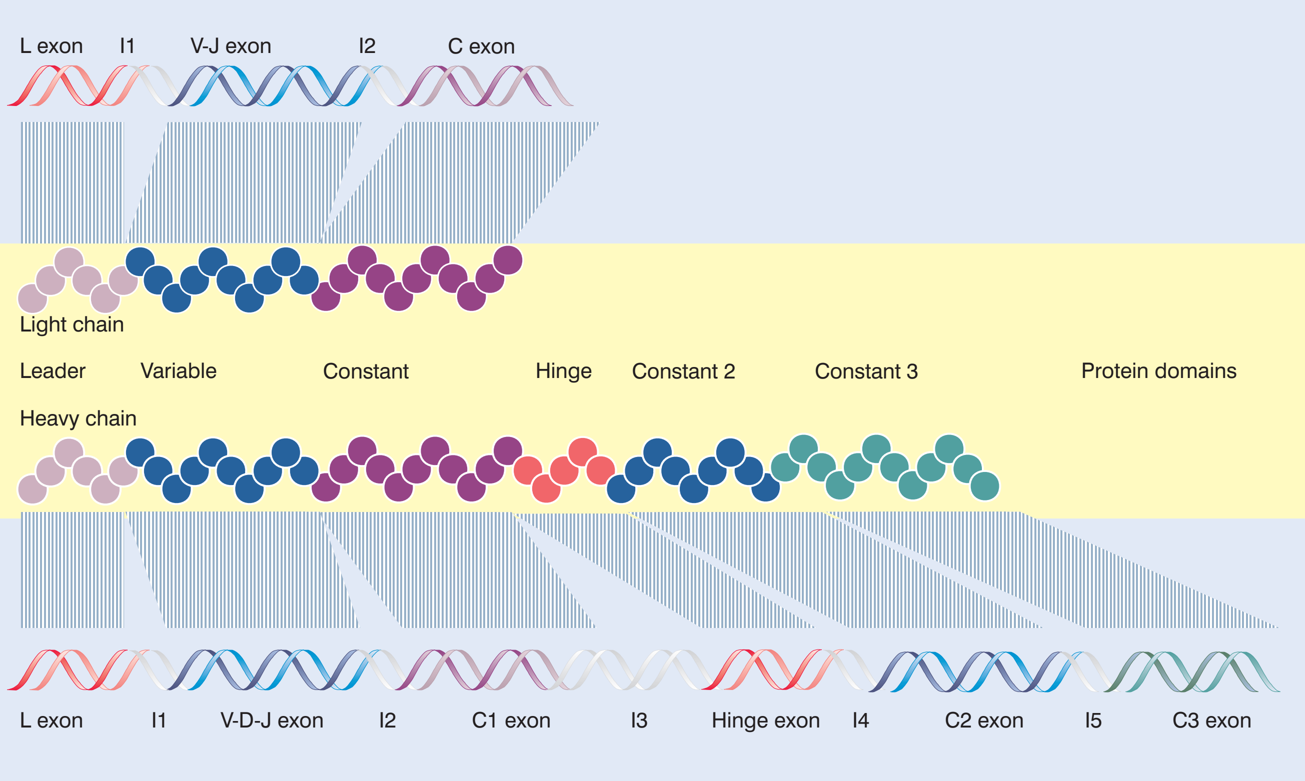
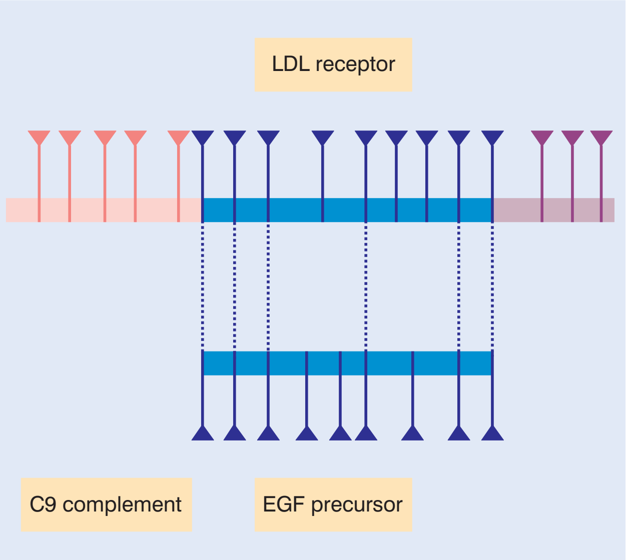


Figure 2.30 The LDL receptor gene consists of 18 exons, some of which are related to EGF precursor and some to the C9 blood complement gene. Triangles mark the positions of introns. Only some of the introns in the region related to EGF precursor are identical in position to those in the EGF gene.



The relationship between exons and protein domains is somewhat erratic in known genes. In some cases there is a clear 1:1 relationship; in others no pattern is to be discerned. One possibility is that removal of introns has fused the adjacent exons. This means that the intron must have been precisely removed, without changing the integrity of the coding region. An alternative is that some introns arose by insertion into a coherent domain; here the difficulty is that we must suppose that the intron carried with it the ability to be spliced out.

Exons tend to be fairly small (see Figure 2.20), around the size of the smallest polypeptide that can assume a stable folded structure, ~20–40 residues. Perhaps proteins were originally assembled from rather small modules. Each module need not necessarily correspond to a current function; several modules could have combined to generate a function. The number of exons in a gene tends to increase with the length of its protein, which is consistent with the view that proteins acquire multiple functions by successively adding appropriate modules.

This idea might explain another feature of protein structure: it seems that the sites represented at exon-intron boundaries often are located at the surface of a protein. As modules are added to a protein, the connections, at least of the most recently added modules, could tend to lie at the surface.

A fascinating case of evolutionary conservation is presented by the globins, all of whose genes have three exons (see Figure 2.13). The two introns are located at constant positions relative to the coding sequence. The central exon appears to represent the heme-binding domain of the globin chain. The α- and β-globin genes have similar structures.

Another perspective on this structure is provided by the existence of two other types of protein that are related to globin. Myoglobin is a monomeric oxygen-binding protein of animals, whose amino acid sequence suggests a common (though ancient) origin with the globin subunits. Leghemoglobins are oxygen-binding proteins present in the legume class of plants; like myoglobin, they are monomeric. They too share a common origin with the other heme-binding proteins. Together, the globins, myoglobin, and leghemoglobin constitute the globin "superfamily," a set of gene families all descended from some (distant) common ancestor.

Myoglobin is represented by a single gene in the human genome, whose structure is essentially the same as that of the globin genes. The three-exon structure therefore predates the evolution of separate myoglobin and globin functions.

Leghemoglobin genes contain three introns, the first and last of which occur at points in the coding sequence that are homologous to the locations of the two introns in the globin genes. This remarkable similarity suggests an exceedingly ancient origin for the heme-binding proteins in the form of a split gene, as illustrated in **Figure 2.31.**

The central intron of leghemoglobin separates two exons that together code for the sequence corresponding to the single central exon in globin. Could the central exon of the globin gene have been derived by a fusion of two central exons in the ancestral gene? Or is the single central exon the ancestral form; in this case, an intron must have been inserted into it at the start of plant evolution?

Cases in which homologous genes differ in structure may provide information about their evolution. An example is insulin. Mammals and birds have only one gene for insulin, except for the rodents, which have two genes. **Figure 2.32** illustrates the structures of these genes.

The principle we use in comparing the organization of related genes in different species is that *a common feature identifies a structure that predated the evolutionary separation of the two species.* In chicken, the single

Figure 2.31 The exon structure of globin genes corresponds with protein function, but leghemoglobin has an extra intron in the central domain.

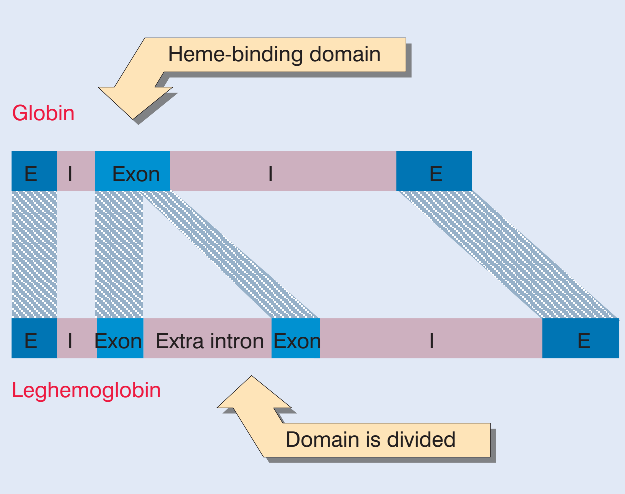
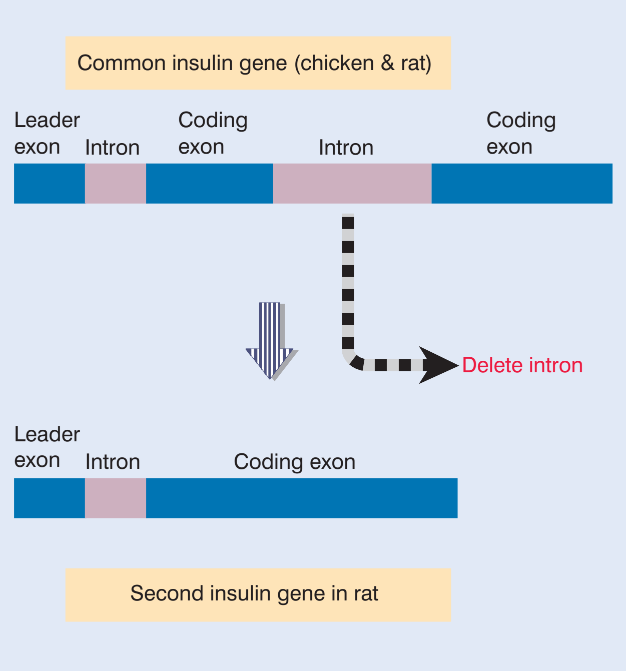


Figure 2.32 The rat insulin gene with one intron evolved by losing an intron from an ancestor with two interruptions.



insulin gene has two introns; one of the two rat genes has the same structure. The common structure implies that the ancestral insulin gene had two introns. However, the second rat gene has only one intron. It must have evolved by a gene duplication in rodents that was followed by the precise removal of one intron from one of the copies.

The organization of some genes shows extensive discrepancies between species. In these cases, there must have been extensive removal or insertion of introns during evolution.

A well characterized case is represented by the actin genes. The typical actin gene has a nontranslated leader of <100 bases, a coding region of ~1200 bases, and a trailer of ~200 bases. Most actin genes are interrupted; the positions of the introns can be aligned with regard to the coding sequence (except for a single intron sometimes found in the leader).

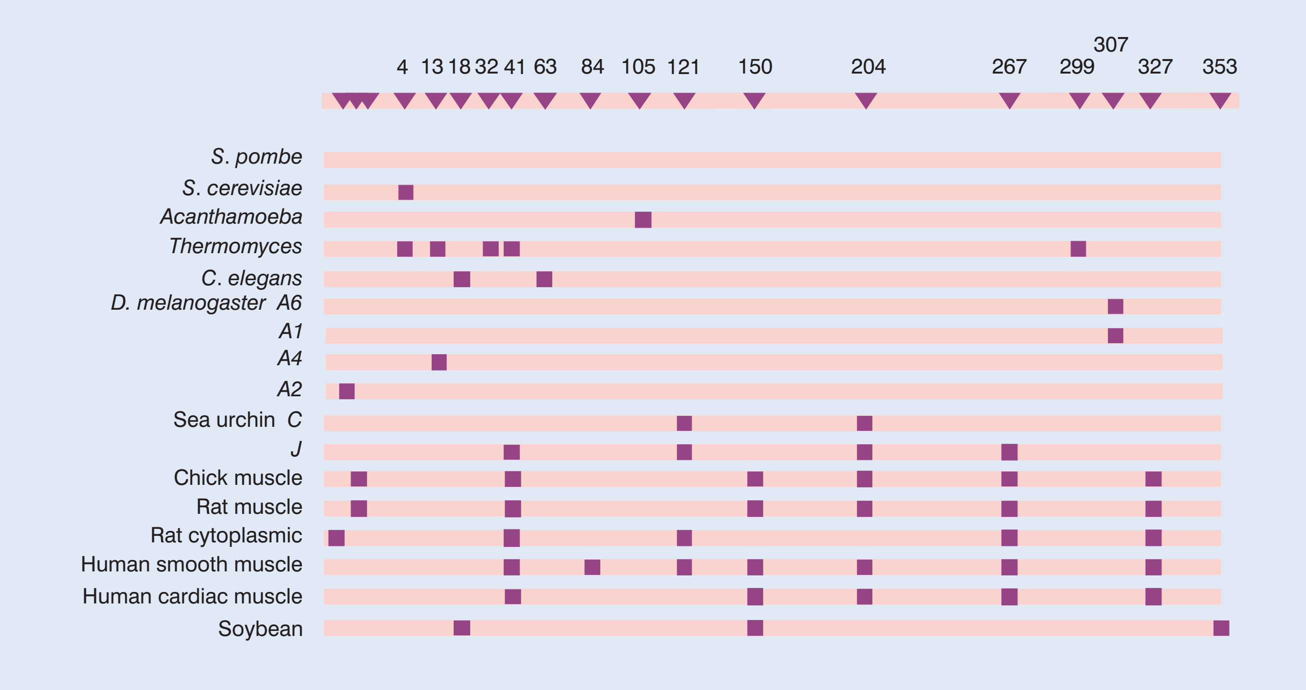
Figure 2.33 shows that almost every actin gene is different in its pattern of interruptions. Taking all the genes together, introns occur at 12 different sites. However, no individual gene has more than six introns; some genes have only one intron, and one is uninterrupted altogether. How did this situation arise? If we suppose that the primordial actin gene was interrupted, and all current actin genes are related to it by loss of introns, different introns have been lost in each evolutionary branch. Probably some introns have been lost entirely, so the primordial gene could well have had 20 or more. The alternative is to suppose that a process of intron insertion continued independently in the different lines of evolution. The relationships between the intron locations found in different species may be used ultimately to construct a tree for the evolution of the gene.

The equation of at least some exons with protein domains, and the appearance of related exons in different proteins, leaves no doubt that the duplication and juxtaposition of exons has played an important role in evolution. It is possible that the number of ancestral exons, from which all proteins have been derived by duplication, variation, and recombination, could be relatively small (a few thousands or tens of thousands). By taking exons as the building blocks of evolution, this view implicitly accepts the introns early model for the origin of genes coding for proteins.

The highly interrupted structure of eukaryotic genes suggests a picture of the eukaryotic genome as a sea of introns (mostly but not exclusively unique in sequence), in which islands of exons (sometimes very short) are strung out in individual archipelagoes that constitute genes.

Alternative forms of genes for rRNA and tRNA are sometimes found, with and without introns. In the case of the tRNAs, where all the molecules conform to the same general structure, it seems unlikely that evolution brought together the two regions of the gene. After all, the different regions are involved in the base

Figure 2.33 Actin genes vary widely in their organization. The sites of introns are indicated in purple; the number identifies the codon interrupted by the intron.



pairing that gives significance to the structure. So here it must be that the introns were inserted into continuous genes.

Organelle genomes provide some striking connections between the prokaryotic and eukaryotic worlds. Because of many general similarities between mitochondria or chloroplasts and bacteria, it seems likely that the organelles originated by an **endosymbiosis** in which an early bacterial prototype was inserted into eukaryotic cytoplasm. Yet in contrast to the resemblances with bacteria—for example, as seen in protein or RNA synthesis—some organelle genes possess introns, and therefore resemble eukaryotic nuclear genes.

Introns are found in several chloroplast genes, including some that have homologies with genes of *E. coli*. This suggests that the endosymbiotic event occurred before introns were lost from the prokaryotic line. If a suitable gene can be found, it may therefore be possible to trace gene lineage back to the period when endosymbiosis occurred.

The mitochondrial genome presents a particularly striking case. The genes of yeast and mammalian mitochondria code for virtually identical mitochondrial proteins, in spite of a considerable difference in gene organization. Vertebrate mitochondrial genomes are very small, with an extremely compact organization of continuous genes, whereas yeast mitochondrial genomes are larger and have some complex interrupted genes. Which is the ancestral form? The yeast mitochondrial introns often have the property of mobility—they are self-contained sequences that can splice out of the RNA and insert DNA copies elsewhere—which suggests that they may have arisen by insertions into the genome (see Chapter 16).

The scope of the paradigm

The concept of the gene has evolved significantly in the past few years. The question of what's in a name is especially appropriate for the gene. We can no longer say that a gene is a sequence of DNA that continuously and uniquely codes for a particular protein. In situations in which a stretch of DNA is responsible for production of one particular protein, current usage regards the entire sequence of DNA, from the first point represented in the messenger RNA to the last point corresponding to its end, as comprising the "gene," exons, introns, and all.

When the sequences representing proteins overlap or have alternative forms of expression, we may reverse the usual description of the gene. Instead of saying "one gene-one polypeptide," we may describe the relationship as "one polypeptide-one gene." So we regard the sequence actually responsible for production of the polypeptide (including introns as well as exons) as constituting the gene, while recognizing that from the perspective of another protein, part of this same sequence also belongs to *its* gene. This allows the use of descriptions such as "overlapping" or "alternative" genes.

We can now see how far we have come from the original one gene : one enzyme hypothesis. Up to that time, the driving question was the nature of the gene. Once it was discovered that genes represent proteins, the paradigm became fixed in the form of the concept that every genetic unit functions through the synthesis of a particular protein.

This view remains the central paradigm of molecular biology: a sequence of DNA functions either by directly coding for a particular protein or by being necessary for the use of an adjacent segment that actually codes for the protein. How far does this paradigm take us beyond explaining the basic relationship between genes and proteins?

The development of multicellular organisms rests on the use of different genes to generate the different cell phenotypes of each tissue. The expression of genes is determined by a regulatory network that probably takes the form of a cascade. Expression of the first set of genes at the start of embryonic development leads to expression of the genes involved in the next stage of development, which in turn leads to a further stage, and so on until all the tissues of the adult are functioning. The molecular nature of this regulatory network is largely unknown, but we assume that it consists of genes that code for products (probably protein, perhaps sometimes RNA) that act on other genes.

While such a series of interactions is almost certainly the means by which the developmental program is executed, we can ask whether it is entirely sufficient. One specific question concerns the nature and role of **positional information.** We know that all parts of a fertilized egg are not equal; one of the features responsible for development of different tissue parts from different regions of the egg is location of information (presumably specific macromolecules) within the cell.

We do not know how these particular regions are formed. But we may speculate that the existence of positional information in the egg leads to the differential expression of genes in the cells subsequently formed in these regions, which leads to the development of the adult organism, which leads to the development of an egg with the appropriate positional information…

This possibility prompts us to ask whether some information needed for development of the organism is contained in a form that we cannot directly attribute to a sequence of DNA (although the expression of particular sequences may be needed to perpetuate the positional information). Put in a more general way, we might ask: if we could read out the entire sequence of DNA comprising the genome of some organism and interpret it in terms of proteins and regulatory regions, could we then construct an organism (or even a single living cell) by controlled expression of the proper genes?

Summary

Genes and genomes can be mapped by the use of overlapping restriction fragments. Ultimately this can be extended into a sequence. Restriction sites can be used as genetic markers. The existence of polymorphisms (RFLPs) allows linkage maps to be constructed using restriction fragments.

All types of eukaryotic genomes contain interrupted genes. The proportion of interrupted genes is low in yeasts and increases in the lower eukaryotes; few genes are uninterrupted in higher eukaryotes.

Introns are found in all classes of eukaryotic genes. The structure of the interrupted gene is the same in all tissues, exons are joined together in RNA in the same order as their organization in DNA, and the introns usually have no coding function. Introns are removed from RNA by splicing. Some genes are expressed by alternative splicing patterns, in which a particular sequence is removed as an intron in some situations, but retained as an exon in others.

Positions of introns are conserved when the organization of homologous genes is compared between species. Intron sequences vary, and may even be unrelated, although exon sequences remain well related. The conservation of exons can be used to isolate related genes in different species.

The size of a gene is determined primarily by the lengths of its introns. Introns become larger early in the higher eukaryotes, when gene sizes therefore increase significantly. The range of gene sizes in mammals is generally 1–100 kb, but it is possible to have even larger genes; the longest known case is dystrophin at 2000 kb.

Some genes share only some of their exons with other genes, suggesting that they have been assembled by addition of exons representing individual modules of the protein. Such modules may have been incorporated into a variety of different proteins. The idea that genes have been assembled by accretion of exons implies that introns were present in genes of primitive organisms. Some of the relationships between homologous genes can be explained by loss of introns from the primordial genes, with different introns being lost in different lines of descent.

Further reading

Reviews

Nathans, D. and Smith, H. O. (1975). Restriction endonucleases in the analysis and restructuring of DNA molecules. *Ann. Rev. Biochem.* **46,** 273–293.

Wilson, A. C. *et al.* (1977). Biochemical evolution. *Ann. Rev. Biochem.* **46,** 573–639.

Breathnach, R. and Chambon, P. (1981). Organization and expression of eukaryotic split genes coding for proteins. *Ann. Rev. Biochem.* **50,** 349–383.

Blake, C. C. (1985). Exons and the evolution of proteins. *Int. Rev. Cytol.* **95,** 149–185.

Wu, R. (1978). DNA sequence analysis. *Ann. Rev. Biochem.* **47,** 607–734.

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

Diener, T. O. (1986). Viroid processing: a model involving the central conserved region and hairpin. *Proc. Nat. Acad. Sci. USA* **83,** 58–62.

Gusella, J. F. (1986). DNA polymorphism and human disease. *Ann. Rev. Biochem.* **55,** 831–854.

Interrupted genes

Wenskink, P. *et al.* (1974). A system for mapping DNA sequences in the chromosomes of *D. melanogaster*. *Cell* **3,** 315–325.

Berget, S. M., Moore, C., and Sharp, P. (1977). Spliced segments at the 5′ terminus of adenovirus 2 late mRNA. *Proc. Nat. Acad. Sci. USA* **74,** 3171–3175.

Chow, L. T., Gelinas, R. E., Broker, T. R., and Roberts, R. J. (1977). An amazing sequence arrangement at the 5′ ends of adenovirus 2 mRNA. *Cell* **12,** 1–8.

Glover, D. M. and Hogness, D. S. (1977). A novel arrangement of the 18S and 28S sequences in a repeating unit of *D. melanogaster* rDNA. *Cell* **10,** 167–176.

Jeffreys, A. J. and Flavell, R. A. (1977). The rabbit β-globin gene contains a large insert in the coding sequence. *Cell* **12,** 1097–1108.

RFLP maps

Danna, K. J., Sack, G. H., and Nathans, D. (1973). Studies of SV40 DNA. VII. A cleavage map of the SV40 genome. *J. Mol. Biol.* **78,** 363–376.

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

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

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