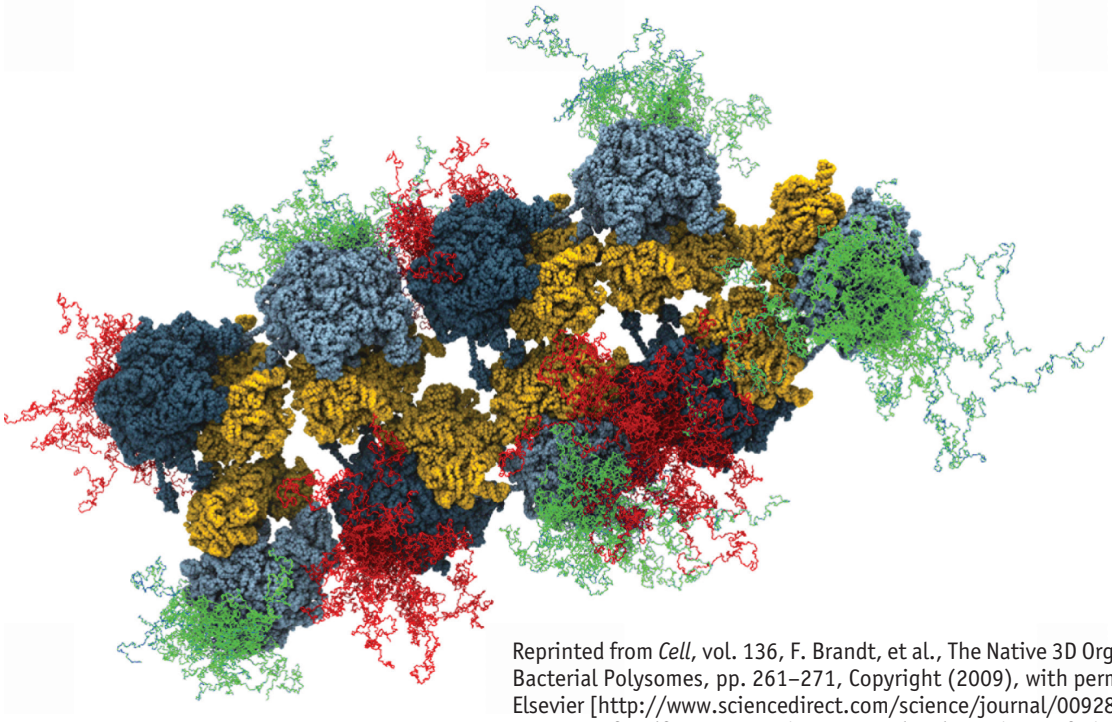
2



Reprinted from *Cell*, vol. 136, F. Brandt, et al., The Native 3D Organization of Bacterial Polysomes, pp. 261–271, Copyright (2009), with permission from Elsevier [http://www.sciencedirect.com/science/journal/00928674]. Photo courtesy of Wolfgang Baumeister, Max-Planck-Institute of Biochemistry.

Genes Code for Proteins

Edited by Esther Siegfried

CHAPTER OUTLINE

2.1 Introduction

The gene is the functional unit of heredity. Each gene is a sequence within the genome that functions by giving rise to a discrete product (which may be a polypeptide or an RNA). The basic behavior of the gene was defined by Mendel more than a century ago. Summarized in his two laws (segregation and independent assortment), the gene was recognized as a "particulate factor" that passes largely unchanged from parent to progeny. A gene may exist in alternative forms. These forms are called **alleles**.

In diploid organisms with two sets of chromosomes, one of each chromosome pair is inherited from each parent. This is also true for genes. One of the two copies of each gene is the paternal allele (inherited from the father), the other is the maternal allele (inherited from the mother). This common pattern of inheritance led to the discovery that chromosomes in fact carry the genes.

Each chromosome consists of a linear array of genes. Each gene resides at a particular location on the chromosome. The location is more formally called a genetic **locus**. The alleles of a gene are the different forms that are found at its locus.

The key to understanding the organization of genes into chromosomes was the discovery of genetic linkage—the tendency for genes on the same chromosome to remain together in the progeny instead of assorting independently as predicted by Mendel's laws. Once the unit of **genetic recombination** (reassortment) was introduced as the measure of linkage, the construction of genetic maps became possible.

The resolution of the recombination map of a multicellular eukaryote is restricted by the small number of progeny that can be obtained from each mating. Recombination occurs so infrequently between nearby points that it is rarely observed between different mutations in the same gene. As a result, classical linkage maps of eukaryotes can place the genes in order, but cannot determine relationships within a gene. By moving to a microbial system in which a very large number of progeny can be obtained from each genetic cross, researchers could demonstrate that recombination occurs within genes. It follows the same rules that

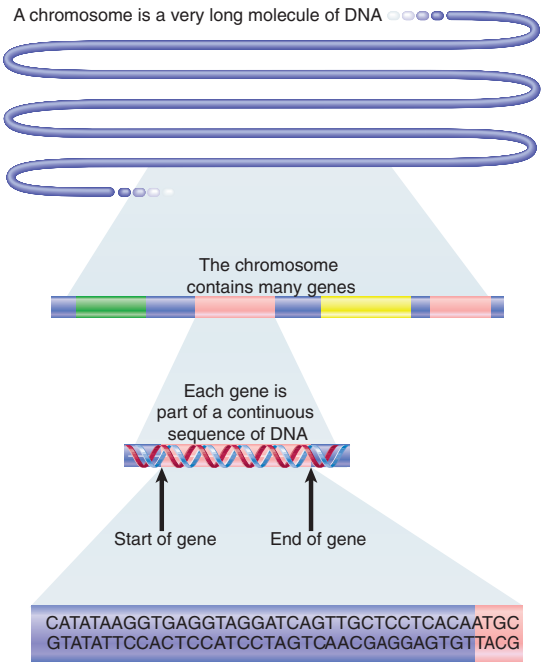


FIGURE 2.1 Each chromosome has a single long molecule of DNA within which are the sequences of individual genes.

were previously deduced for recombination between genes.

Mutations within a gene can be arranged into a linear order, showing that the gene itself has the same linear construction as the array of genes on a chromosome. Thus the genetic map is linear within as well as between loci: it consists of an unbroken sequence within which the genes reside. This conclusion leads naturally into the modern view summarized in **FIGURE 2.1** that the genetic material of a chromosome consists of an uninterrupted length of DNA representing many genes. Having defined the gene as an uninterrupted length of DNA, it should be noted that in eukaryotes many genes are interrupted by sequences in the DNA that are then excised from the mRNA (see Chapter 4, *The Interrupted Gene*).

2.2 A Gene Codes for a Single Polypeptide

Key concepts

- The one gene: one enzyme hypothesis summarizes the basis of modern genetics: that a gene is a stretch of DNA coding for one or more isoforms of a single polypeptide.
- Some genes do not encode polypeptides, but encode structural or regulatory RNAs.
- Most mutations damage gene function and are recessive to the wild-type allele.

The first systematic attempt to associate genes with enzymes, carried out by Beadle and Tatum in the 1940s, showed that each stage in a metabolic pathway is catalyzed by a single enzyme and can be blocked by mutation in a different gene. This led to the **one gene : one enzyme hypothesis**. Each metabolic step is catalyzed by a particular enzyme, whose production is the responsibility of a single gene. A mutation in the gene alters the activity of the protein for which it is responsible.

A modification in the hypothesis is needed to accommodate proteins that consist of more than one subunit. If the subunits are all the same, the protein is a **homomultimer** and is represented by a single gene. If the subunits are different, the protein is a **heteromultimer**. Stated as a more general rule applicable to any heteromultimeric protein, the one gene: one enzyme hypothesis becomes more precisely expressed as the **one gene : one polypeptide hypothesis**. Even this general rule needs to be refined because many genes encode multiple, related polypeptides through alternative splicing of the mRNA (see Chapter 21, *RNA Splicing and Processing*).

Identifying which protein represents a particular gene can be a protracted task. The mutation responsible for Mendel's wrinkled-pea mutant was identified only in 1990 as an alteration that inactivates the gene for a starch-branching enzyme!

It is important to remember that a gene does not directly generate a polypeptide. As shown previously in Figure 1.2, a gene codes for an RNA, which may in turn code for a polypeptide. Many genes code for polypeptides, but some genes code for RNAs that do not give rise to polypeptides. These RNAs may be structural components of the apparatus responsible for synthesizing proteins or, as has become evident in recent years, have roles in regulating gene expression (see Chapter 30, *Regulatory RNA*). The basic principle is that the gene is a sequence of DNA that specifies the sequence of an independent product. The process of gene expression may terminate in a product that is either RNA or polypeptide.

A mutation is a random event with regard to the structure of the gene, so the greatest probability is that it will damage or even abolish gene function. Most mutations that affect gene function are recessive: *they represent an absence of function, because the mutant gene has been prevented from producing its usual product.* **FIGURE 2.2** illustrates the relationship between recessive

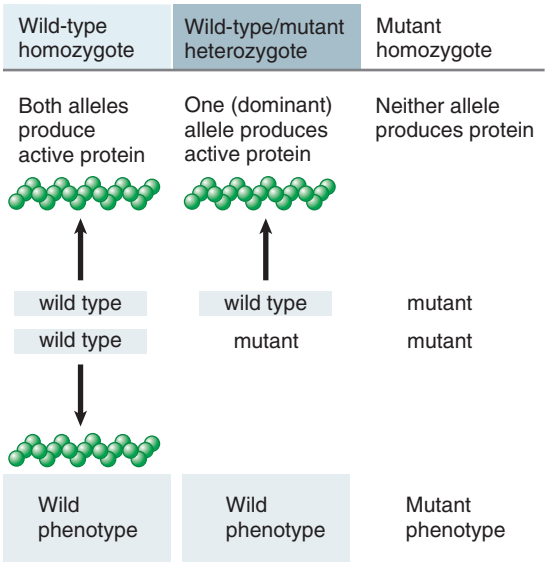


FIGURE 2.2 Genes code for proteins; dominance is explained by the properties of mutant proteins. A recessive allele does not contribute to the phenotype in the wild-type/mutant heterozygote because it produces no protein (or protein that is nonfunctional). If both alleles are the recessive mutant allele, no active protein is produced.

and wild-type alleles. When a heterozygote contains one wild-type allele and one mutant allele, the wild-type allele is able to direct production of the normal gene product. The wild-type allele is therefore dominant. (This assumes that an adequate *amount* of product is made by the single wild-type allele. When this is not true, the smaller amount made by one allele as compared to two alleles results in the intermediate phenotype of a partially dominant allele in a heterozygote.)

2.3 Mutations in the Same Gene Cannot Complement

Key concepts

- A mutation in a gene affects only the product (protein or RNA) coded by the mutant copy of the gene and does not affect the product coded by any other allele.
- Failure of two mutations to complement (produce wild phenotype) when they are present in *trans* configuration in a heterozygote means that they are part of the same gene.

How do we determine whether two mutations that cause a similar phenotype lie in the same gene? If they map close together, they may be alleles. They could, however, also represent mutations in two *different* genes whose proteins are involved in the same function. The **complementation test** is used to determine whether two mutations lie in the same gene or in different genes. The test consists of making a heterozygote for the two mutations.

If the mutations lie in the same gene, the parental genotypes can be represented as:

$$\frac{m_1}{m_1} \text{ and } \frac{m_2}{m_2}$$

The first parent provides an m_1 mutant allele and the second parent provides an m_2 allele, so that the heterozygote has the constitution:

$$\frac{m_1}{m_2}$$

No wild-type gene is present, so the heterozygote has mutant phenotype and the alleles fail to complement. If the mutations lie in different genes, the parental genotypes can be represented as:

$$\frac{m_1+}{m_1+} \text{ and } \frac{+m_2}{+m_2}$$

Each chromosome has a wild-type copy of one gene (represented by the plus sign) and a mutant copy of the other. Then the heterozygote has the constitution:

$$\frac{m_1+}{+m_2}$$

in which the two parents between them have provided a wild-type copy of each gene. The heterozygote has wild phenotype, and thus the two genes are said to *complement*.

The complementation test is shown in more detail in **FIGURE 2.3**. The basic test consists of the comparison shown in the top part of the figure. If two mutations lie in the same gene, we see a difference in the phenotypes of the *trans* configuration and the *cis* configuration. The *trans* configuration is mutant because each allele has a (different) mutation, whereas the *cis* co nfiguration is wild-type because one allele has two mutations and the other allele has no mutations. The lower part of the figure shows that if the two mutations lie in different genes, we always see a wild phenotype. There is always one wild-type and one mutant allele of each gene, and the configuration is irrelevant. Failure to complement means that two mutations are part of the *same* genetic unit. Mutations that do not complement one another are said to comprise part of the same *complementation group*. Another term used to describe the

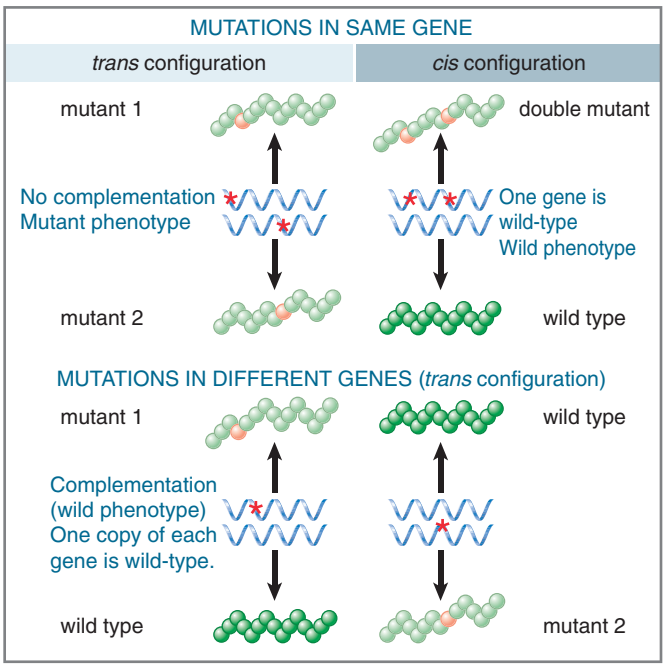


FIGURE 2.3 The cistron is defined by the complementation test. Genes are represented by spirals; red stars identify sites of mutation.

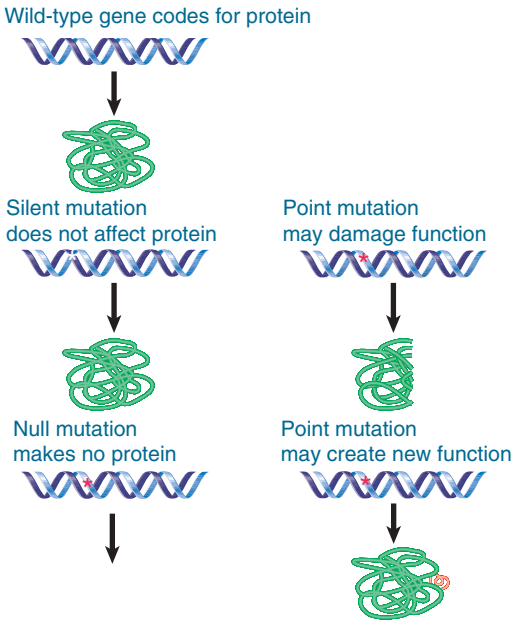


FIGURE 2.4 Mutations that do not affect protein sequence or function are silent. Mutations that abolish all protein activity are null. Point mutations that cause loss-of-function are recessive; those that cause gain-of-function are dominant.

unit defined by the complementation test is the **cistron**. This is the same as the gene. Basically these three terms all describe a stretch of DNA that functions as a unit to give rise to an RNA or protein product. The properties of the gene with regard to complementation are explained by the fact that this product is a single molecule that behaves as a functional unit.

2.4 Mutations May Cause Loss-of-Function or Gain-of-Function

Key concepts

- Recessive mutations are due to loss-of-function by the protein product.
- Dominant mutations result from a gain-of-function.
- Testing whether a gene is essential requires a null mutation (one that completely eliminates its function).
- Silent mutations have no effect, either because the base change does not change the sequence or amount of protein, or because the change in protein sequence has no effect.

The various possible effects of mutation in a gene are summarized in FIGURE 2.4.

When a gene has been identified, insight into its function in principle can be gained by generating a mutant organism that entirely lacks the gene. A mutation that completely eliminates gene function—usually because the gene has been deleted—is called a **null mutation**. If a gene is essential, a null mutation is lethal when homozygous or hemizygous.

To determine what effect a gene has upon the phenotype, it is essential to characterize a null mutant. Generally, if a null mutant fails to affect a phenotype, we may safely conclude that the gene function is not necessary. Some genes have overlapping functions, though, and removal of one gene is not sufficient to significantly affect the phenotype. Null mutations, or other mutations that impede gene function (but do not necessarily abolish it entirely), are called **loss-of-function mutations**. A loss-of-function mutation is recessive (as in the example of Figure 2.2). Loss-of-function mutations that affect protein activity but retain sufficient activity so that the phenotype is not altered are referred to as *leaky mutations*. Sometimes a mutation has the opposite effect and causes a protein to acquire a new function or expression pattern; such a change is called a **gain-of-function mutation**. A gain-of-function mutation is dominant.

Not all mutations in protein-coding genes lead to a detectable change in the phenotype. Mutations without apparent effect are called **silent mutations**. They comprise two types: One type involves base changes in DNA that do not cause any change in the amino acid present

in the corresponding protein. The second type changes the amino acid, but the replacement in the protein does not affect its activity; these are called **neutral substitutions**.

2.5 A Locus May Have Many Different Mutant Alleles

Key concept

- The existence of multiple alleles allows heterozygotes that represent any pairwise combination of alleles to exist.

If a recessive mutation is produced by every change in a gene that prevents the production of an active protein, there should be a large number of such mutations in any one gene. Many amino acid replacements may change the structure of the protein sufficiently to impede its function.

Different variants of the same gene are called *multiple alleles*, and their existence makes possible a heterozygote with two mutant alleles. The relationship between these multiple alleles takes various forms.

In the simplest case, a wild-type allele codes for a product that is functional. Mutant allele(s) code for products that are nonfunctional. There are often cases, though, in which a series of mutant alleles affect the same phenotype to differing extents. For example, wild-type function of the *white* locus of *Drosophila melanogaster* is required for development of the normal red color of the eye. The locus is named for the effect of extreme (null) mutations, which cause the fly to have white eyes in mutant homozygotes.

To denote wild-type and mutant alleles, the wild-type genotype is indicated by a plus superscript after the name of the locus (w^+ is the wild-type allele for [red] eye color in *D. melanogaster*). Sometimes + is used by itself to describe the wild-type allele, and only the mutant alleles are indicated by the name of the locus.

An entirely defective form of the gene (or absence of phenotype) may be indicated by a minus superscript. To distinguish among a variety of mutant alleles with different effects, other superscripts may be introduced, such as w^i or w^a.

The w^+ allele is dominant over any other allele in heterozygotes. There are many different mutant alleles. **FIGURE 2.5** shows a (small) sample. Although some alleles produce no visible pigment, and therefore the eyes are white, many alleles produce some color. Each of these

Allele	Phenotype of homozygote
w^+	red eye (wild type)
w^{bl}	blood
w^{ch}	cherry
w^{bf}	buff
w^h	honey
w^a	apricot
w^e	eosin
w^I	ivory
w^z	zeste (lemon-yellow)
w^{sp}	mottled, color varies
w^1	white (no color)

FIGURE 2.5 The *w* locus has an extensive series of alleles whose phenotypes extend from wild-type (red) color to complete lack of pigment.

mutant alleles must therefore represent a different mutation of the gene, which does not eliminate its function entirely, but leaves a residual activity that produces a characteristic phenotype. These alleles are named for the color of the eye in a homozygote. (Most *w* alleles affect the quantity of pigment in the eye. The examples in the figure are arranged in [roughly] declining amount of color, but others, such as w^{sp}, affect the pattern in which it is deposited.)

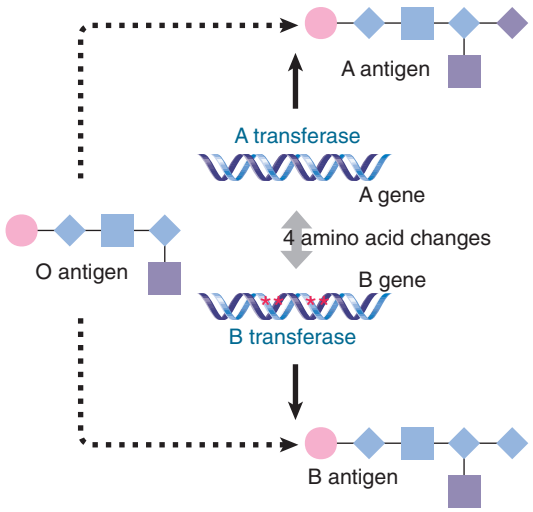
When multiple alleles exist, an organism may be a heterozygote that carries two different mutant alleles. The phenotype of such a heterozygote depends on the nature of the residual activity of each allele. The relationship between two mutant alleles is in principle no different from that between wild-type and mutant alleles: one allele may be dominant, there may be partial dominance, or there may be codominance.

2.6 A Locus May Have More Than One Wild-type Allele

Key concept

- A locus may have a polymorphic distribution of alleles with no individual allele that can be considered to be the sole wild-type.

There is not necessarily a unique wild-type allele at any particular locus. Control of the human blood group system provides an example. Lack of function is represented by the null type *O* group. The functional alleles *A* and *B*, however, provide activities that are codominant with one another and dominant over *O* group. The basis for this relationship is illustrated in **FIGURE 2.6**.



Phenotype	Genotype	Transferase Activity
O	OO	None
A	AO or AA	N-Ac-gal transferase
B	BO or BB	Gal transferase
AB	AB	GalN-Ac-Gal-transferase

FIGURE 2.6 The ABO blood group locus codes for a galactosyltransferase whose specificity determines the blood group.

The O (or H) antigen is generated in all individuals and consists of a particular carbohydrate group that is added to proteins. The *ABO* locus codes for a galactosyltransferase enzyme that adds a further sugar group to the O antigen. The specificity of this enzyme determines the blood group. The *A* allele produces an enzyme that uses the cofactor UDP-N-acetylgalactose, creating the A antigen. The *B* allele produces an enzyme that uses the cofactor UDP-galactose, creating the B antigen. The A and B versions of the transferase protein differ in four amino acids that presumably affect its recognition of the type of cofactor. The *O* allele has a transferase mutation (a small deletion) that eliminates activity, so no modification of the O antigen occurs.

This explains why *A* and *B* alleles are dominant in the *AO* and *BO* heterozygotes: the corresponding transferase activity creates the A or B antigen. The *A* and *B* alleles are codominant in *AB* heterozygotes, because both transferase activities are expressed. The *OO* homozygote is a null that has neither activity and therefore lacks both antigens.

Neither *A* nor *B* can be regarded as uniquely wild type, because they represent alternative activities rather than loss or gain of function. A situation such as this, in which there are multiple functional alleles in a population, is described as a **polymorphism** (see *Section 5.3, Individual Genomes Show Extensive Variation*).

2.7 Recombination Occurs by Physical Exchange of DNA

Key concepts

- Recombination is the result of crossing-over that occurs at chiasmata and involves two of the four chromatids.
- Recombination occurs by a breakage and reunion that proceeds via an intermediate of hybrid DNA that depends on the complementarity of the two strands of DNA.
- The frequency of recombination between two genes is proportional to their physical distance; recombination between genes that are very closely linked is rare.
- For genes that are very far apart on a single chromosome, the frequency of recombination is not proportional to their physical distance because recombination happens so frequently.

Genetic recombination describes the generation of new combinations of alleles that occurs at each generation in diploid organisms. The two copies of each chromosome may have different alleles at some loci. By exchanging corresponding parts between the chromosomes, recombinant chromosomes that are different from the parental chromosomes can be generated.

Recombination results from a physical exchange of chromosomal material. This is visible in the form of the *crossing-over* that occurs during meiosis (the specialized division that produces haploid germ cells). Meiosis starts with a cell that has duplicated its chromosomes, so that it has four copies of each chromosome. Early in meiosis, all four copies are closely associated (synapsed) in a structure called a *bivalent*. Each individual chromosomal unit is called a *chromatid* at this stage. Pairwise exchanges of material occur between two nonidentical (nonsister) chromatids.

The visible result of a crossing-over event is called a **chiasma** and is illustrated diagrammatically in **FIGURE 2.7**. A chiasma represents a site at which two of the chromatids in a bivalent have been broken at corresponding points. The broken ends have been rejoined crosswise, generating new chromatids. Each new chromatid consists of material derived from one chromatid on one side of the junction point, with material from the other chromatid on the opposite side.

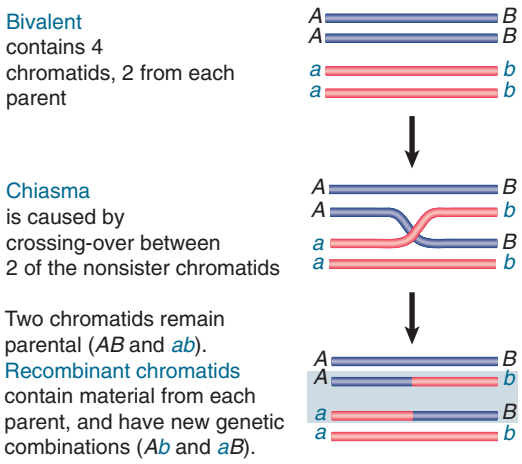


FIGURE 2.7 Chiasma formation is responsible for generating recombinants.

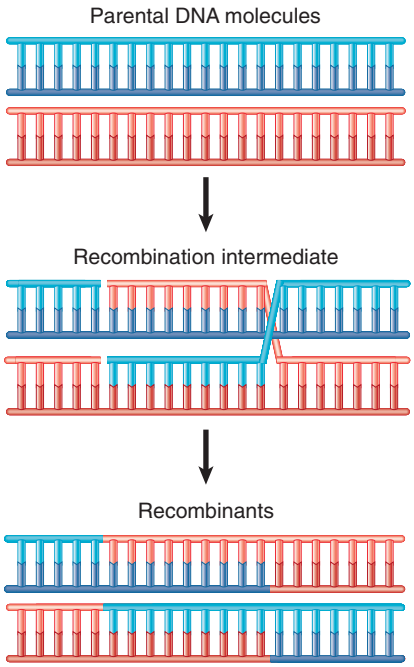


FIGURE 2.8 Recombination involves pairing between complementary strands of the two parental duplex DNAs.

The two recombinant chromatids have reciprocal structures. The event is described as a *breakage and reunion*. Its nature explains why a single recombination event can produce only 50% recombinants: each individual recombination event involves only two of the four associated chromatids.

The complementarity of the two strands of DNA is essential for the recombination process. Each of the chromatids shown in Figure 2.7 consists of a very long duplex of DNA. For them to be broken and reconnected without any loss of material requires a mechanism to recognize exactly corresponding positions through complementary base pairing.

Recombination involves a process in which the single strands in the region of the crossover exchange their partners. **FIGURE 2.8** shows that this creates a stretch of *hybrid DNA*, in which the single strand of one duplex is paired with its complement from the other duplex. Each duplex DNA corresponds to one of the chromatids involved in recombination in Figure 2.7. The mechanism, of course, involves other stages (strands must be broken and resealed), which we discuss in more detail in Chapter 15 (*Homologous and Site-Specific Recombination*), but the crucial feature that makes precise recombination possible is the complementarity of DNA strands. The figure shows only some stages of the reaction, but we see that a stretch of hybrid DNA forms in the recombination intermediate when a single strand crosses over from one duplex to the other. Each recombinant consists of one parental duplex DNA at the left, which is connected by a stretch of hybrid DNA to the other parental duplex at the right.

The formation of hybrid DNA requires the sequences of the two recombining duplexes to be close enough to allow pairing between the complementary strands. If there are no differences between the two parental genomes in this region, formation of hybrid DNA will be perfect. The reaction can be tolerated, however, even when there are small differences. In this case, the hybrid DNA has points of mismatch, at which a base in one strand faces a base in the other strand that is not complementary to it. The correction of such mismatches is another feature of genetic recombination (see Chapter 16, *Repair Systems Handle Damage to DNA*).

Over chromosomal distances, recombination events occur more or less at random with a characteristic frequency. The probability that a crossover will occur within any specific region of the chromosome is more or less proportional to the length of the region, up to a saturation point. For example, a large human chromosome usually has three or four crossover events per meiosis, whereas a small chromosome has only one on average.

FIGURE 2.9 compares three situations: two genes on different chromosomes, two genes that are far apart on the same chromosome, and two genes that are close together on the same chromosome. Genes on different chromosomes segregate independently according to Mendel's laws, resulting in the production

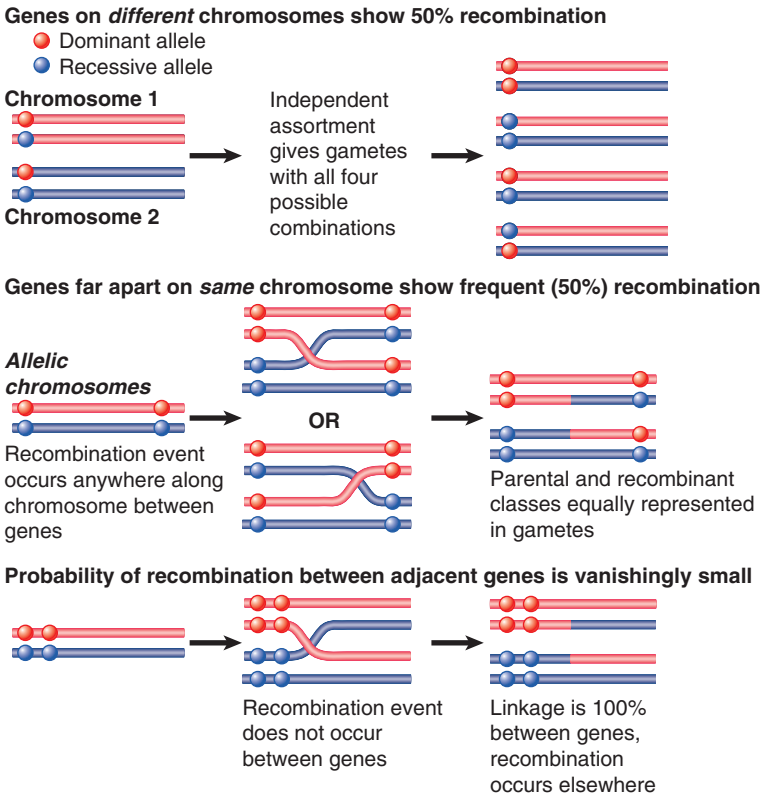


FIGURE 2.9 Genes on different chromosomes segregate independently so that all possible combinations of alleles are produced in equal proportions. Recombination occurs so frequently between genes that are far apart on the same chromosome that they effectively segregate independently. Recombination is reduced, however, when genes are closer together, and for adjacent genes may hardly ever occur.

of 50% parental types and 50% recombinant types during meiosis. When genes are sufficiently far apart on the same chromosome, the probability of one or more recombination events in the region between them becomes so high that they behave in the same way as genes on different chromosomes and show 50% recombination.

When genes are close together, though, the probability of a recombination event between them is reduced, and occurs only in some proportion of meioses. For example, if it occurs in one quarter of the meioses, the overall rate of recombination is 12.5% (because a single recombination event produces 50% recombination, and this occurs in 25% of meioses). When genes are very close together, as shown in the bottom panel of Figure 2.9, recombination between them may never be observed in phenotypes of higher eukaryotes.

This leads us to the view that a chromosome contains an array of many genes. Each protein-coding gene is an independent unit of expression, and is represented in one or more polypeptide chains. The properties of a gene can be changed by mutation. The allelic combinations present on a chromosome can be changed by recombination. We can now ask, "what is the relationship between the sequence of a gene and the sequence of the polypeptide chain it represents?"

2.8 The Genetic Code Is Triplet

Key concepts

- The genetic code is read in triplet nucleotides called codons.
- The triplets are nonoverlapping and are read from a fixed starting point.
- Mutations that insert or delete individual bases cause a shift in the triplet sets after the site of mutation.
- Combinations of mutations that together insert or delete three bases (or multiples of three) insert or delete amino acids, but do not change the reading of the triplets beyond the last site of mutation.

Each gene represents a particular polypeptide chain. The concept that each protein consists of a particular series of amino acids dates from Sanger's characterization of insulin in

the 1950s. The discovery that a gene consists of DNA presents us with the issue of how a sequence of nucleotides in DNA represents a sequence of amino acids in protein.

The sequence of nucleotides in DNA is important not because of its structure *per se*, but because it *codes* for the sequence of amino acids that constitutes the corresponding polypeptide. The relationship between a sequence of DNA and the sequence of the corresponding protein is called the **genetic code**.

The structure and/or enzymatic activity of each protein follows from its primary sequence of amino acids and its overall conformation, which is determined by interactions between the amino acids. By determining the sequence of amino acids in each protein, the gene is able to carry all the information needed to specify an active polypeptide chain. In this way, a single type of structure—the gene—is able to represent itself in innumerable polypeptide forms.

Together the various protein products of a cell undertake the catalytic and structural activities that are responsible for establishing its phenotype. Of course, in addition to sequences that code for proteins, DNA also contains certain sequences whose function is to be recognized by regulator molecules, usually proteins. Here the function of the DNA is determined by its sequence directly, not via any intermediary code. Both types of region—genes expressed as proteins and sequences recognized as such—constitute genetic information.

The genetic code is deciphered by a complex apparatus that interprets the nucleic acid sequence. This apparatus is essential if the information carried in DNA is to have meaning. In any given region, only one of the two strands of DNA codes for protein, so we write the genetic code as a sequence of bases (rather than base pairs).

The genetic code is read in groups of three nucleotides, each group representing one amino acid. Each trinucleotide sequence is called a **codon**. A gene includes a series of codons that is read sequentially from a starting point at one end to a termination point at the other end. Written in the conventional 5′ to 3′ direction, the nucleotide sequence of the DNA strand that codes for protein corresponds to the amino acid sequence of the protein written in the direction from N-terminus to C-terminus.

The genetic code is read in *nonoverlapping triplets from a fixed starting point*:

- The use of a *fixed starting point* means that assembly of a protein must start at one end and work to the other, so that different parts of the coding sequence cannot be read independently.

The nature of the code predicts that two types of mutations, base substitution and base insertion/deletion, will have different effects. If a particular sequence is read sequentially, such as:

UUU AAA GGG CCC (codons)
aa1 aa2 aa3 aa4 (amino acids)

then a base substitution, or point mutation, will affect only one amino acid. For example, the substitution of an A by some other base (X) causes aa2 to be replaced by aa5:

UUU AAX GGG CCC
aa1 aa5 aa3 aa4

because only the second codon has been changed.

A mutation that inserts or deletes a single base, though, will change the triplet sets for the entire subsequent sequence. A change of this sort is called a **frameshift**. An insertion might take the form:

UUU AAX AGG GCC C
aa1 aa5 aa6 aa7

The new sequence of triplets is completely different from the old one, and as a result the entire amino acid sequence of the protein is altered beyond the site of mutation. Thus the function of the protein is likely to be lost completely.

Frameshift mutations are induced by the **acridines**. The acridines are compounds that bind to DNA and distort the structure of the double helix, causing additional bases to be incorporated or omitted during replication. Each mutagenic event sponsored by an acridine results in the addition or removal of a single base pair.

If an acridine mutant is produced by, say, addition of a nucleotide, it should revert to wild-type by deletion of the nucleotide. Reversion also can be caused by deletion of a different base, though, at a site close to the first. Combinations of such mutations provided revealing evidence about the nature of the genetic code.

FIGURE 2.10 illustrates the properties of frameshift mutations. An insertion or deletion changes the entire protein sequence following the site of mutation. The combination of an insertion *and* a deletion, though, causes the code to be read incorrectly only between the

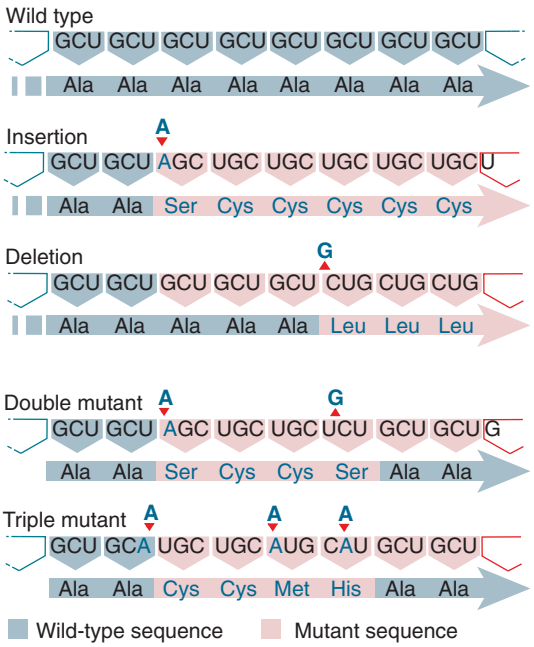


FIGURE 2.10 Frameshift mutations show that the genetic code is read in triplets from a fixed starting point.

two sites of mutation; correct reading resumes after the second site.

In 1961, genetic analysis of acridine mutations in the *rII* region of the phage T4 showed that all the mutations could be classified into one of two sets, described as (+) and (–). Either type of mutation by itself causes a frameshift: the (+) type by virtue of a base addition, and the (–) type by virtue of a base deletion. Double mutant combinations of the types (+ +) and (– –) continue to show mutant behavior. Combinations of the types (+ –) or (– +), however, suppress one another, giving rise to a description in which one mutation is described as a *frameshift suppressor* of the other. (In the context of this work, "suppressor" is used in an unusual sense because the second mutation is in the same gene as the first.)

These results show that the genetic code must be read as a sequence that is fixed by the starting point. Thus additions or deletions compensate for each other, whereas double additions or double deletions remain mutant. This does not, however, reveal how many nucleotides make up each codon.

When triple mutants are constructed, only (+ + +) and (– – –) combinations show the wild phenotype, whereas other combinations remain mutant. If we take three additions or three deletions to correspond respectively to the addition or omission overall of a single amino acid, this implies that the code is read in triplets. An incorrect amino acid sequence is found between the two outside sites of mutation and the sequence on either side remains wild-type, as indicated in Figure 2.10.

2.9 Every Sequence Has Three Possible Reading Frames

Key concept

- In general, only one reading frame is translated, and the other two are blocked by frequent termination signals.

If the genetic code is read in nonoverlapping triplets, there are three possible ways of translating any nucleotide sequence into protein, depending on the starting point. These are called **reading frames**. For the sequence

A C G A C G A C G A C G A C G A C G

the three possible reading frames are

ACG ACG ACG ACG ACG ACG ACG
CGA CGA CGA CGA CGA CGA CGA
GAC GAC GAC GAC GAC GAC GAC

A reading frame that consists exclusively of triplets representing amino acids is called an **open reading frame** or **ORF**. A sequence that is translated into protein has a reading frame that starts with a special **initiation codon** (**AUG**) and then extends through a series of triplets representing amino acids until it ends at one of three types of **termination codon** (see Chapter 25, *Using the Genetic Code*).

A reading frame that cannot be read into protein because termination codons occur frequently is said to be **closed** or **blocked**. If a sequence is blocked in all three reading frames, it cannot have the function of coding for protein.

When the sequence of a DNA region of unknown function is obtained, each possible reading frame is analyzed to determine whether it is open or blocked. Usually no more than one of the three possible frames of reading is open in any single stretch of DNA. **FIGURE 2.11** shows an example of a sequence that can be read in only one reading frame because the alternative reading frames are blocked by frequent termination codons. A long open reading frame is unlikely to exist by chance; if it were not translated into protein, there would have been no

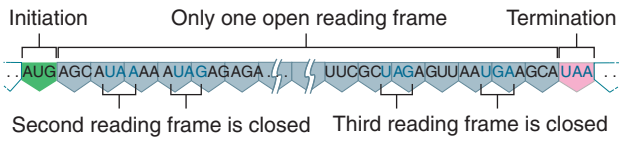


FIGURE 2.11 An open reading frame starts with AUG and continues in triplets to a termination codon. Blocked reading frames may be interrupted frequently by termination codons.

selective pressure to prevent the accumulation of termination codons. Thus the identification of a lengthy open reading frame is taken to be *prima facie* evidence that the sequence is translated into protein in that frame. An ORF for which no protein product has been identified is sometimes called an **unidentified reading frame** (**URF**).

2.10 Prokaryotic Genes Are Colinear with Their Proteins

Key concepts

- A prokaryotic gene consists of a continuous length of 3*N* nucleotides that encodes *N* amino acids.
- The gene, mRNA, and protein are all colinear.

By comparing the nucleotide sequence of a gene with the amino acid sequence of a protein, we can determine directly whether the gene and the protein are *colinear*; that is, whether the sequence of nucleotides in the gene corresponds exactly with the sequence of amino acids in the protein. In bacteria and their viruses, there is an exact equivalence. Each gene contains a continuous stretch of DNA whose length is directly related to the number of amino acids in the protein that it represents. A gene with an open reading frame of 3*N* bp is required to code for a protein of *N* amino acids, according to the genetic code.

The equivalence of the bacterial gene and its product means that a physical map of DNA will exactly match an amino acid map of the protein. How well do these maps fit with the recombination map?

The **colinearity** of gene and protein was originally investigated in the tryptophan synthetase gene of *E. coli*. Genetic distance was measured by the percent recombination between mutations; protein distance was measured by the number of amino acids separating sites of replacement. **FIGURE 2.12** compares the two maps. The order of seven sites of mutation is the same as the order of the corresponding sites of amino acid replacement, and the recombination distances are relatively similar to the actual distances in the protein. The recombination map expands the distances between some mutations, but otherwise there is little distortion of the recombination map relative to the physical map.

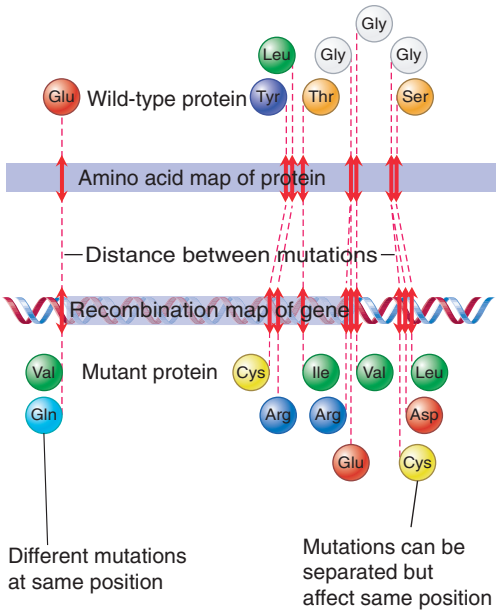


FIGURE 2.12 The recombination map of the tryptophan synthetase gene corresponds with the amino acid sequence of the protein.

The recombination map makes two further general points about the organization of the gene. Different mutations may cause a wild-type amino acid to be replaced with different substituents. If two such mutations cannot recombine, they must involve different point mutations at the same position in DNA. If the mutations can be separated on the genetic map, but affect the same amino acid on the upper map (the connecting lines converge in the figure), they must involve point mutations at different positions that affect the same amino acid. This happens because the unit of genetic recombination (1 bp) is smaller than the unit coding for the amino acid (3 bp).

2.11 Several Processes Are Required to Express the Protein Product of a Gene

Key concepts

- A prokaryotic gene is expressed by transcription into mRNA and then translation of the mRNA into protein.
- In eukaryotes, a gene may contain internal regions that are not represented in protein.
- Internal regions are removed from the mRNA transcript by RNA splicing to give an mRNA that is colinear with the protein product.
- Each mRNA consists of an untranslated 5′ region, a coding region, and an untranslated 3′ trailer.

In comparing gene and protein, we are restricted to dealing with the sequence of DNA stretching between the points corresponding to the ends of the protein. A gene is not directly translated into protein, though, but instead is expressed via the production of a **messenger RNA** (abbreviated to **mRNA**), a nucleic acid intermediate actually used to synthesize a protein (as we see in detail in Chapter 22, *mRNA Stability and Localization*).

Messenger RNA is synthesized by the same process of complementary base pairing used to replicate DNA, with the important difference that it corresponds to only one strand of the DNA double helix. FIGURE 2.13 shows that the sequence of mRNA is complementary with the sequence of one strand of DNA and is identical (apart from the replacement of T with U) with the other strand of DNA. The convention for writing DNA sequences is that the top strand runs 5′→3′, with the sequence that is the same as RNA.

The process by which a gene gives rise to a protein is called **gene expression**. In bacteria, it consists of two stages. The first stage is **transcription**, when an mRNA copy of one strand of the DNA is produced. The second stage is **translation** of the mRNA into protein. This is the process by which the sequence of an mRNA is read in triplets to give the series of amino acids that make the corresponding protein.

An mRNA includes a sequence of nucleotides that corresponds with the sequence of amino acids in the protein. This part of the nucleic acid is called the **coding region**. Note, however, that the mRNA includes additional sequences on either end; these sequences do not directly encode polypeptide. The 5′ untranslated region is called the **leader** or **5′ UTR**, and the 3′ untranslated region is called the **trailer** or **3′ UTR**.

The *gene* includes the entire sequence represented in messenger RNA. Sometimes mutations impeding gene function are found in the additional, noncoding regions, confirming the view that these comprise a legitimate part of the genetic unit.

FIGURE 2.14 illustrates this situation, in which the gene is considered to comprise a continuous stretch of DNA that is needed to produce a particular protein. It includes the sequence coding for that protein, but also includes sequences on either side of the coding region.

A bacterium consists of only a single compartment, so transcription and translation occur in the same place, as illustrated in FIGURE 2.15.

In eukaryotes transcription occurs in the nucleus, but the mRNA product must be *transported* to the cytoplasm in order to be translated. For the simplest eukaryotic genes (just like in bacteria) the translated RNA is in fact the transcribed copy of the gene. For more complex genes, however, the immediate transcript of the gene is a **pre-mRNA** that requires **RNA processing** to generate the mature mRNA. The basic stages of gene expression in a eukaryote are outlined in FIGURE 2.16. This results in a spa-

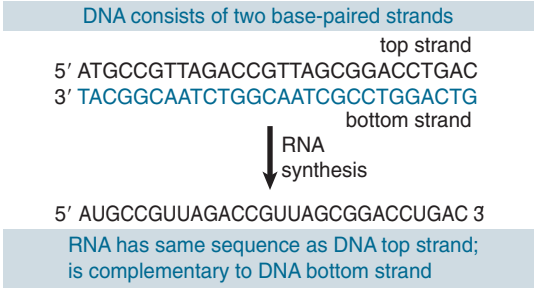


FIGURE 2.13 RNA is synthesized by using one strand of DNA as a template for complementary base pairing.

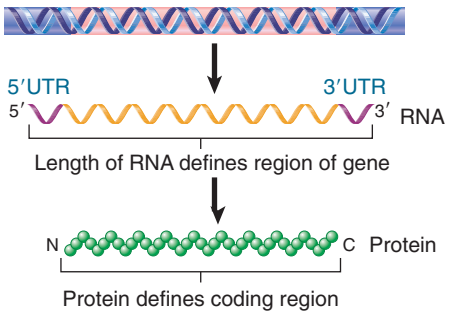


FIGURE 2.14 The gene may be longer than the sequence coding for protein.

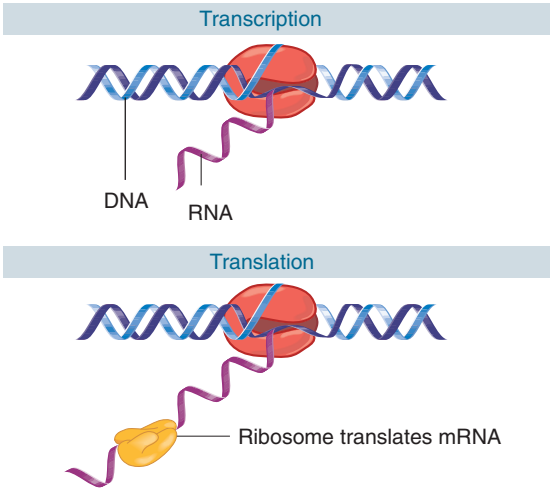


FIGURE 2.15 Transcription and translation take place in the same compartment in bacteria.

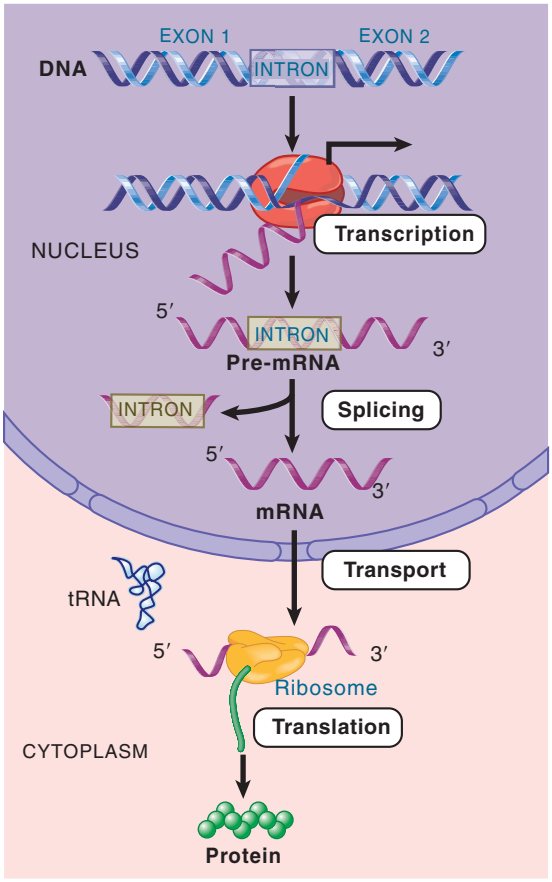


FIGURE 2.16 Gene expression is a multistage process.

tial separation between transcription (in the nucleus) and translation (in the cytoplasm).

The most important stage in processing is **splicing**. Many genes in eukaryotes (and a majority in multicellular eukaryotes) contain internal regions called **introns** that do not code for protein. The process of splicing removes these regions from the pre-mRNA to generate an RNA that has a continuous open reading frame (see Figure 4.1). (The remaining, expressed regions of the mRNA are called **exons**.) Other processing events that occur at this stage involve the modification of the 5′ and 3′ ends of the pre-mRNA (see Figure 21.1).

Translation is accomplished by a complex apparatus that includes both protein and RNA components. The actual "machine" that undertakes the process is the **ribosome**, a large complex that includes some large RNAs (**ribosomal RNAs**, abbreviated to **rRNAs**) and many small proteins. The process of recognizing which amino acid corresponds to a particular nucleotide triplet requires an intermediate **transfer RNA** (abbreviated to **tRNA**); there is at least one tRNA species for every amino acid. Many ancillary proteins are involved. We describe translation in Chapter 24, *Translation*, but note for now that the ribosomes are the large structures in Figure 2.14 that move along the mRNA.

The important point to note at this stage is that the process of gene expression involves RNA not only as the essential substrate, but also in providing components of the apparatus. The rRNA and tRNA components are coded by genes and are generated by the process of transcription (just like mRNA, except that there is no subsequent stage of translation).

2.12 Proteins Are *trans*-acting, but Sites on DNA Are *cis*-acting

Key concepts

- All gene products (RNA or proteins) are *trans*-acting. They can act on any copy of a gene in the cell.
- *cis*-acting mutations identify sequences of DNA that are targets for recognition by *trans*-acting products. They are not expressed as RNA or protein and affect only the contiguous stretch of DNA.

A crucial step in the definition of the gene was the realization that all its parts must be present on one contiguous stretch of DNA. In genetic terminology, sites that are located on the same DNA are said to be in *cis*. Sites that are located on two different molecules of DNA are described as being in *trans*. So two mutations may be in *cis* (on the same DNA) or in *trans* (on different DNAs). The complementation test uses this concept to determine whether two

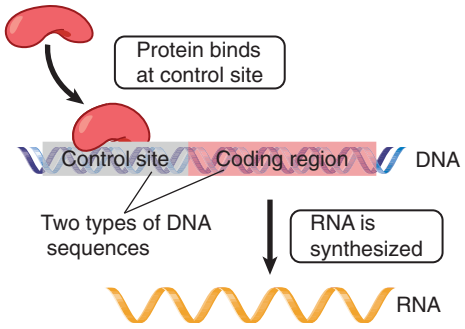


FIGURE 2.17 Control sites in DNA provide binding sites for proteins; coding regions are expressed via the synthesis of RNA.

mutations are in the same gene (see Figure 2.3). We may now extend the concept of the difference between *cis* and *trans* effects from defining the coding region of a gene to describing the interaction between and a gene and its regulatory elements.

Suppose that the ability of a gene to be expressed is controlled by a protein that binds to the DNA close to the coding region. In the example depicted in **FIGURE 2.17**, mRNA can be synthesized only when the protein is bound to the DNA. Now suppose that a mutation occurs in the DNA sequence to which this protein binds, so that the protein can no longer recognize the DNA. As a result, the DNA can no longer be expressed.

So a gene can be inactivated either by a mutation in a control site or by a mutation in a coding region. The mutations cannot be distinguished genetically, because both have the property of acting only on the DNA sequence of the single allele in which they occur. They have identical properties in the complementation test, and a mutation in a control region is therefore defined as comprising part of the gene in the same way as a mutation in the coding region.

FIGURE 2.18 shows that a deficiency in the control site *affects only the coding region to which it is connected; it does not affect the ability of the other allele to be expressed*. A mutation that acts solely by affecting the properties of the contiguous sequence of DNA is called a ***cis*-acting sequence**. It should be noted that in many eukaryotes the control region can influence the expression of DNA at some distance, but nonetheless the control region resides in the same DNA molecule as the coding sequence.

We may contrast the behavior of the *cis*-acting mutation shown in Figure 2.17 with the result of a mutation in the gene coding for the regulator protein. **FIGURE 2.19** shows that the

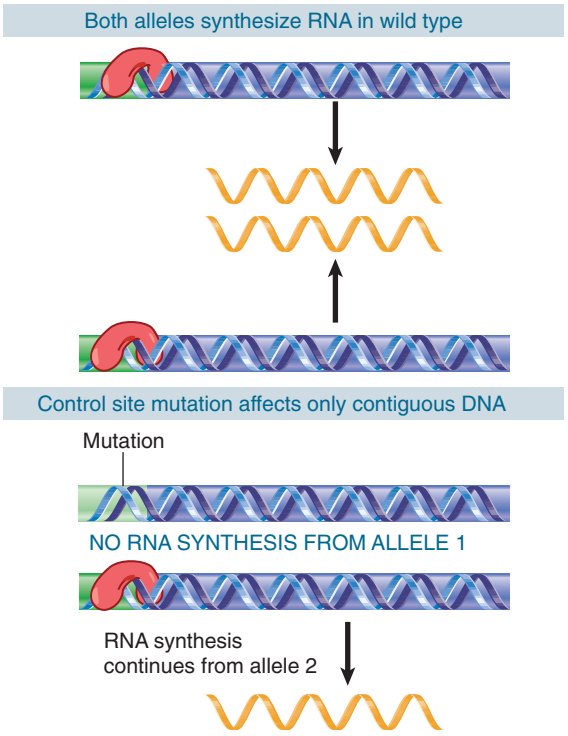


FIGURE 2.18 A *cis*-acting site controls expression of the adjacent DNA but does not influence the other allele.

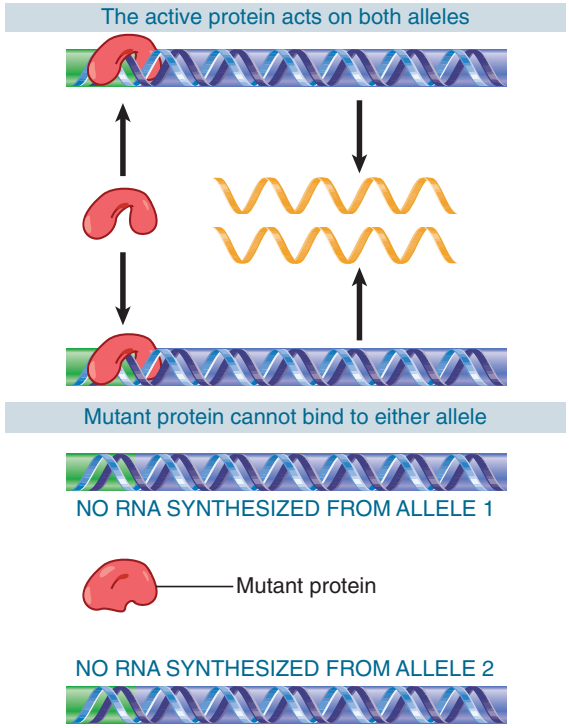


FIGURE 2.19 A *trans*-acting mutation in a protein affects both alleles of a gene that it controls.

absence of regulator protein would prevent *both* alleles from being expressed. A mutation of this sort is said to be in a ***trans*-acting sequence**.

Reversing the argument, if a mutation is *trans*-acting, we know that its effects must be exerted through some diffusible product (either a protein or a regulatory RNA) that acts on multiple targets within a cell. If a mutation is *cis*-acting, though, it must function via affecting directly the properties of the contiguous DNA, which means that it is *not expressed in the form of RNA or protein*.

2.13 Summary

A chromosome consists of an uninterrupted length of duplex DNA that contains many genes. Each gene (or cistron) is transcribed into an RNA product, which in turn is translated into a polypeptide sequence if the gene codes for protein. An RNA or protein product of a gene is said to be *trans*-acting. A gene is defined as a unit of a single stretch of DNA by the complementation test. A site on DNA that regulates the activity of an adjacent gene is said to be *cis*-acting.

When a gene codes for protein, the relationship between the sequence of DNA and sequence of the protein is given by the genetic code. Only one of the two strands of DNA codes for protein. A codon consists of three nucleotides that represent a single amino acid. A coding sequence of DNA consists of a series of codons, read from a fixed starting point and nonoverlapping. Usually one of the three possible reading frames can be translated into protein.

A gene may have multiple alleles. Recessive alleles are caused by loss-of-function mutations that interfere with the function of the protein. A null allele has total loss-of-function. Dominant alleles are caused by gain-of-function mutations that create a new property in the protein.

References

2.8 The Genetic Code Is Triplet

Review

Roth, J. R. (1974). Frameshift mutations. *Annu. Rev. Genet.* 8, 319–346.

Research

Benzer, S. and Champe, S. P. (1961). Ambivalent rII mutants of phage T4. *Proc. Natl. Acad. Sci. USA* 47, 403–416.

Crick, F. H. C., Barnett, L., Brenner, S., and Watts-Tobin, R. J. (1961). General nature of the genetic code for proteins. *Nature* 192, 1227–1232.

2.10 Prokaryotic Genes Are Colinear with Their Proteins

Research

Yanofsky, C., Drapeau, G. R., Guest, J. R., and Carlton, B. C. (1967). The complete amino acid sequence of the tryptophan synthetase A protein (μ subunit) and its colinear relationship with the genetic map of the A gene. *Proc. Natl. Acad. Sci. USA* 57, 2966–2968.

Yanofsky, C. et al. (1964). On the colinearity of gene structure and protein structure. *Proc. Natl. Acad. Sci. USA*, 51, 266–272.