AN OPEN SOURCE MOLECULAR GENETICS TEXTBOOK

GENETICS 1.1

Genetics 1.1
Michael Deyholos
30 November, 2010

The first edition of this book was produced in January, 2009 as instructional material for students in Biology 207 at the University of Alberta, and is released to the public for non-commercial use under the Creative Commons License (See below). Users are encouraged to make modifications and improvements to the book. All text in the original edition was written by Michael Deyholos, Ph.D. Photos and some diagrams were obtained from various, non-copyrighted sources, including Flickr, Wikipedia, Public Library of Science, and Wikimedia Commons. Photo attributions are listed at the end of the book.

This work is licensed under the Creative Commons Attribution-Noncommercial-Share Alike 2.5 Canada License. To view a copy of this license, visit http://creativecommons.org/licenses/by-nc-sa/2.5/ca/ or send a letter to Creative Commons, 171 Second Street, Suite 300, San Francisco, California, 94105, USA.

Under the terms of the license, you are free:

- **to Share —** to copy, distribute and transmit the work
- **to Remix —** to adapt the work

Under the following conditions:

- **Attribution**. You must include the name of the original author in books or excerpts derived from this work.
- **Noncommercial**. You may not use this work for commercial purposes.
- **Share Alike**. If you alter, transform, or build upon this work, you may distribute the resulting work only under the same or similar licence to this one.
- For any reuse or distribution, you must make clear to others the licence terms of this work.
- Any of the above conditions can be waived if you get permission from the copyright holder.
- The author's moral rights are retained in this licence.

TABLE OF CONTENTS

Chapter 1 INTRODUCTION AND OVERVIEW

Why do offspring look like their parents? Even ancient people were aware that the characteristics of an individual plant or animal could be passed between generations. They also knew that some heritable characteristics (such as the size of fruit) varied between individuals, and that they could select for the most favorable traits while breeding crops and animals. The once prevalent (but now discredited) concept of **blending inheritance** proposed that an undefined essence, in its entirety, contained all of the heritable information for an individual. Much like the mixing of two colors of paint, it was thought that mating combined the essences from each parent; once blended together, the individual characteristics of the parents could not be separated again.

GENES ARE UNITS OF INHERITANCE

Mendel was one of the first to take a scientific approach to the study of heredity. He started with well-characterized materials, repeated his experiments many times, and kept careful records of his observations. For example, working with peas, Mendel showed that white-flowered plants could be produced by crossing two purple-flowered plants, but only if the purple-flowered plants themselves had at least one white-flowered parent (Fig 1.2). This was evidence that the genetic factor that produced white-flowers had *not* blended irreversibly with the factor for purple-flowers. Mendel's observations helped to disprove blending inheritance, in favor of an alternative concept called **particulate inheritance**, in which heredity is the product of discrete factors that control independent traits. Each hereditary factor could exist in one or more different versions, or **alleles**.

Figure 1.1
Gregor Mendel

Mendel's discrete factors of heredity later became known as **genes**. In their narrowest definition, genes are abstract concepts: units of inheritance. The connection between genes and substances like DNA and chromosomes was established largely through the experiments described in the remainder of this chapter. However, it is worth noting that Mendel and many researchers who followed him were able to provide great insights into biology simply by observing the inheritance of specific traits.

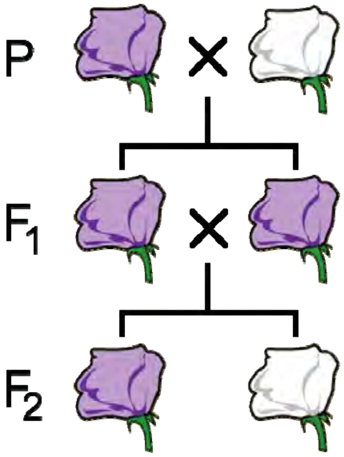


Figure 1.2 Inheritance of flower color in peas. Mendel observed that a cross between pure breeding,white and purple peas (generation P) produced only progeny (generation F_1) with purple flowers. However, white flowered plant reappeared among the F_2 generation progeny of a mating between two F_1 plants. The symbols P, F_1 and F_2 are abbreviations for parental, first filial, and second filial generations, respectively.

DNA IS THE GENETIC MATERIAL

By the early 1900's, biochemists had isolated hundreds of different chemicals from living cells. Which of these was the genetic material? Proteins seemed like promising candidates, since they were abundant and complex molecules. However, a few key experiments helped to prove that DNA, rather than protein, is the genetic material.

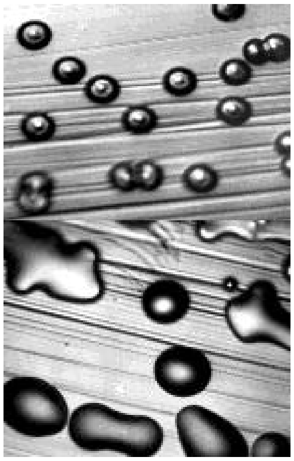


Figure 1.3 Colonies of Rough (top) and Smooth (bottom) strains of *S. pneumoniae.*

Microbiologists had identified two strains of the bacterium *Streptococcus pneumoniae*. One strain (R) produced rough colonies, while the other (S) was smooth (Fig. 1.3). More importantly, the S-type bacteria caused fatal infections, while the R-type did not. **Griffith** noticed that simply mixing the strains together could transform some R-type bacteria into smooth, pathogenic strains (Fig. 1.4). This transformation occurred even when the S-type strains were first killed by heat. Thus, some component of the S-type strains contained genetic information that could be transferred to the R-type strains.

What type of molecule from within the S-type cells was responsible for the transformation? To answer this, researchers named **Avery, MacLeod and McCarty** separated the S-type cells into various components, such as proteins, polysaccharides, lipids, and nucleic acids. Only the nucleic acids from S-type cells were able to make the R-strains smooth and fatal. Furthermore, when cellular extracts of S-type cells were treated with DNase (an enzyme that digests DNA), the transformation ability was lost. The researchers therefore concluded that DNA was the genetic material, which in this case controlled the appearance and pathogenicity of the bacteria.

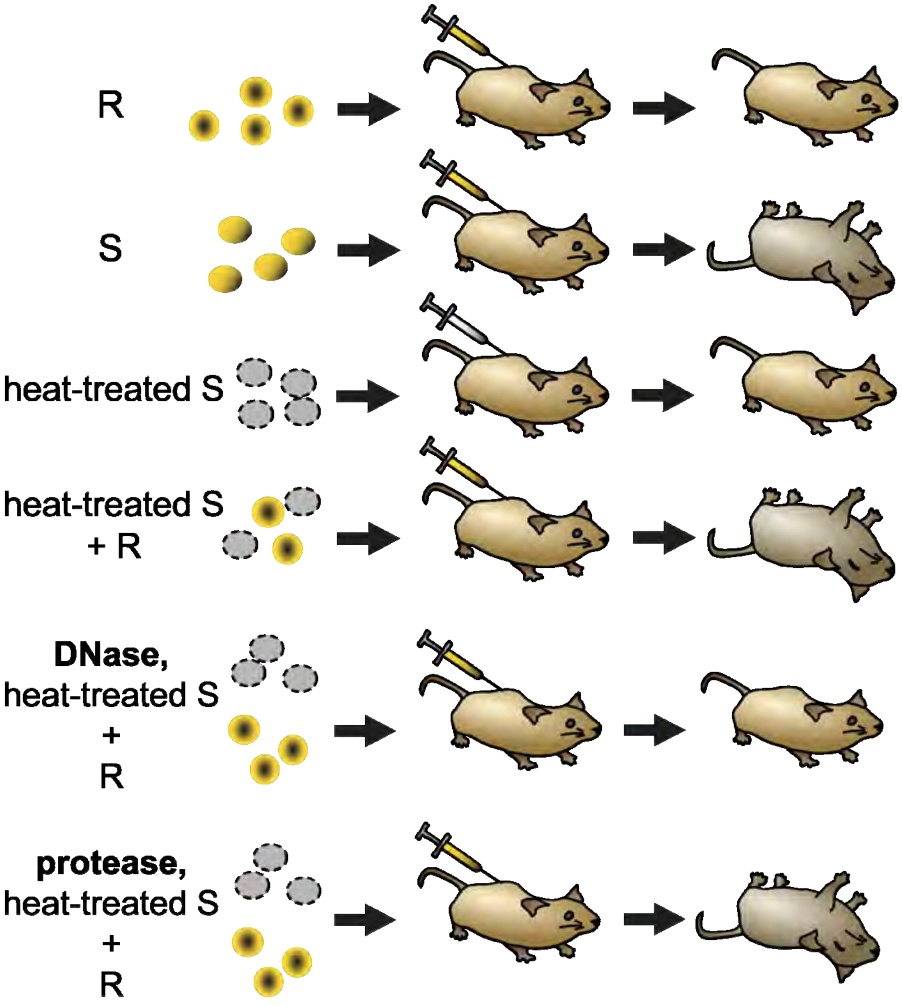


Figure 1.4 Experiments of Griffith and of Avery , MacLeod and McCarty. R strains of *S. pneumoniae* do not cause lethality. However, DNA-containing extracts from pathogenic S strains are sufficient to make R strains pathogenic.

Further evidence that DNA is the genetic material came from experiments conducted by **Hershey and Chase**. These researchers studied the transmission of genetic information from a virus called the T2 bacteriophage to its host bacterium, *Escherichia coli* (Fig. 1.5). Like all viruses, T2 hijacks the cellular machinery of its host to manufacture more viruses. Viruses contain both protein and DNA. To determine which of these types of molecules contained the genetic blueprint for the virus, Hershey and Chase grew viral cultures in the presence of radioactive isotopes of either phosphorus (^{32}P) or sulphur (^{35}S), which incorporated these isotopes into their DNA and proteins, respectively (Fig 1.6). The researchers then infected *E. coli* with the radiolabeled viruses, and looked to see whether ^{32}P or ^{35}S entered the bacteria. After ensuring that any viruses had been removed from the surface of the cells, the researchers observed that only infection with ^{32}P labeled viruses resulted in radioactive bacteria. This again indicated that DNA was the material that contained genetic instructions.

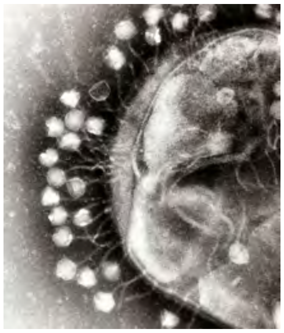


Figure 1.5 Electronmicrograph of T2 bacteriophage on surface of *E. coli*

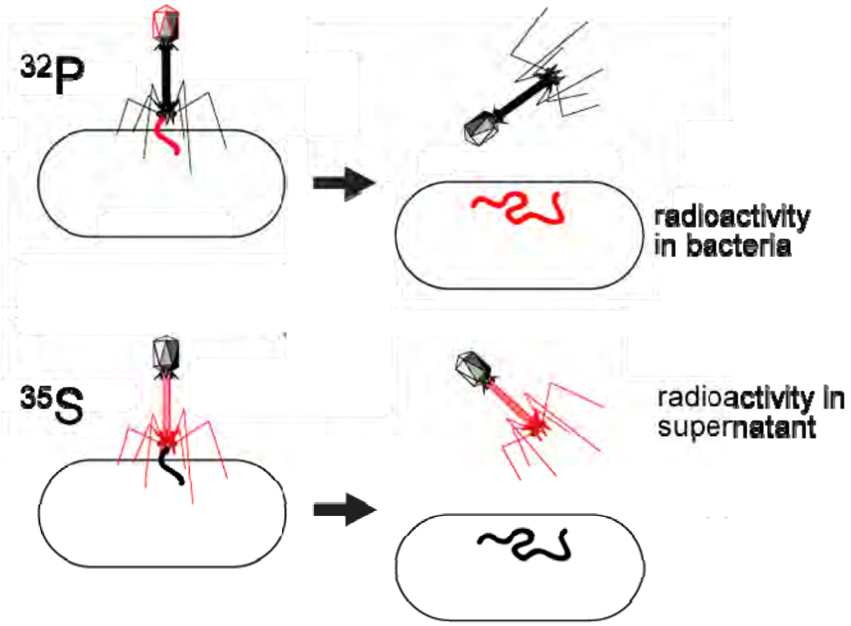


Figure 1.6 When ^{32}P-labeled phage infects *E. coli*, radioactivity is found only in the bacteria after the phage are removed by agitation and centrifugation. In contrast, after infection with ^{35}S-labeled phage, radioactivity is found only in the supernatant that remains after the bacteria are removed.

THE STRUCTURE OF DNA

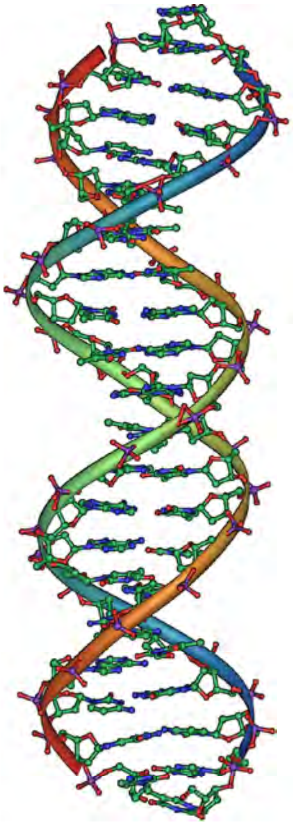


Figure 1.7 DNA structure

The experiments of Hershey and Chase, Avery, MacLeod, and McCarty, Griffith, and others proved that DNA was the genetic material, but very little was known about its structure. When **Watson and Crick** set out to determine the structure of DNA in the 1940's they knew that DNA was made up of four different types of molecules, called bases, or nucleotides: adenine (A), cytosine (C), thymine (T), guanine (G). They also knew of **Chargaff's Rules**, which was a set of observations about the relative amount of each nucleotide that was present in almost any extract of DNA. Chargaff had observed that for any given species, the abundance of A was the same as T, and G was the same as C. Using metal models of the individual nucleotides, Watson and Crick were able to deduce a structure for DNA that was consistent with Chargaff's Rules and with x-ray crystallography data that was obtained (with some controversy) from another researcher, Rosalind Franklin.

In Watson and Crick's famous double helix, each of the two strands contains DNA bases connected through covalent bonds to a sugar-phosphate backbone (Fig 1.7, 1.8). Because one side of each sugar molecule is always connected to the opposite side of the next sugar molecule, each strand of DNA has polarity: these are called the 5' (5-prime) end and the 3' (3-prime) end, in accordance with the nomenclature of the carbons in the sugars. The two strands of the double helix run in anti-parallel (i.e. opposite) directions, with the 5' end of one strand adjacent to the 3' end of the other strand. The double helix has a right-handed twist, (rather than the left-handed twist that is often represented in popular media). The DNA bases extend from the backbone towards the center of the helix, with a pair of bases from each strand forming hydrogen bonds that help to hold the two strands together. Under most conditions, the two strands are slightly offset, creating a major groove on one face of the double helix, and a minor groove on the other. Because of the structure of the bases, A can only form hydrogen bonds with T, and G can only form hydrogen bonds with C (hence, Chargaff's Rules). Each strand is therefore said to be complementary to the other, and each strand also contains enough information to replace the other. This redundancy is important in repairing DNA, and also in its replication.

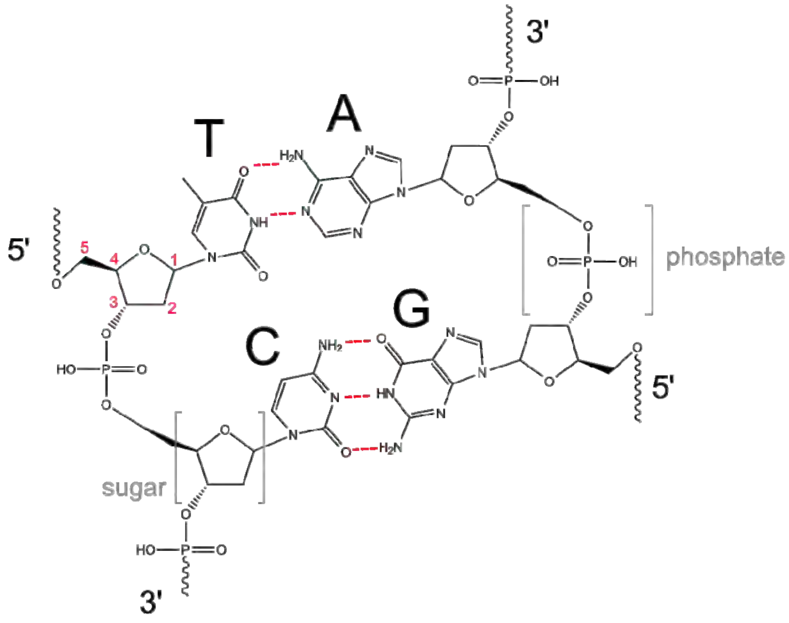


Figure 1.8 Chemical structure of two pairs of nucleotides in a fragment of double-stranded DNA. Sugar, phosphate, and bases A,C,G,T are labeled. Hydrogen bonds between bases on opposite strands are shown by dashed lines. Note that the G-C pair has more hydrogen bonds than A-T. The numbering of carbons within sugars is indicated by red numbers. Based on this numbering the polarity of each strand is indicated by the labels 5' and 3'.

THE FUNCTION OF DNA

How does the structure of DNA relate to inheritance of biological traits such as the flower color of Mendel's peas? The answer to this lies in what has become known as molecular biology's **Central Dogma**, which states that each gene is encoded in DNA, and then as needed, this genetic information is transcribed into RNA and then translated into protein. In certain circumstances, RNA may also be converted to DNA through a process called reverse transcription. The order of bases in DNA directly controls the order of amino acids that make up a protein. Proteins do most of the work in a cell, and catalyze the formation and breakdown of almost all of the molecules within an organism. By dictating the structure of each protein, DNA affects the function of that protein, and can thereby affect the entire organism. In the case of Mendel's peas, purple-flowered plants have a gene that encodes an enzyme that produces a purple pigment molecule. In the white-flowered plants, the DNA for this gene has been changed so that it no longer encodes a functional protein. This is an example of a natural mutation in a biochemical pathway.

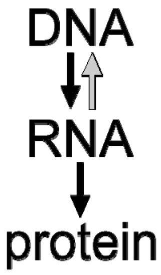
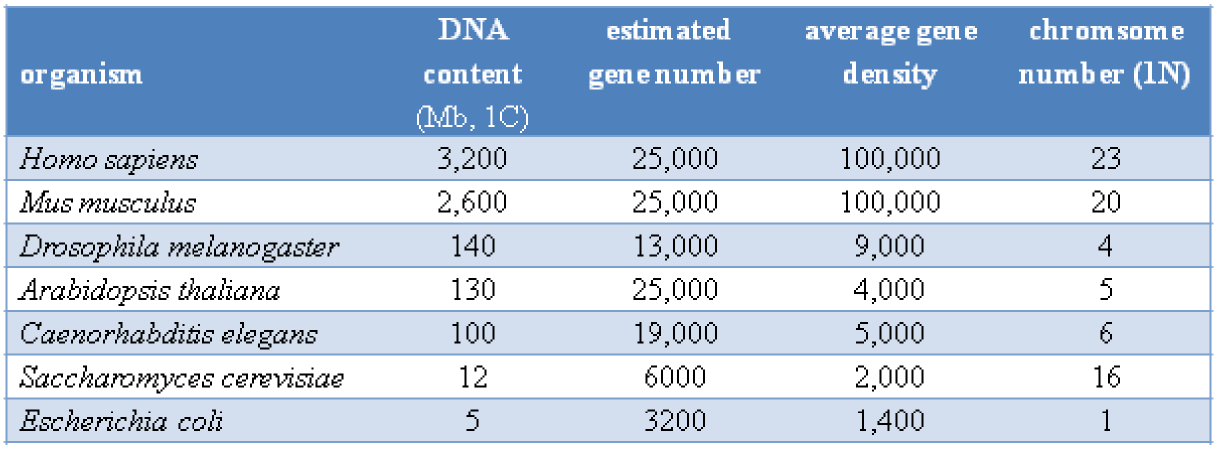


Figure 1.9 Central Dogma of molecular biology

The complete set of DNA within the nucleus of any organism is called its genome. Organelles such as mitochondria and chloroplasts also have their own genomes. There is surprisingly little correlation between the nuclear DNA content of a genome (i.e. the **c-value**) and the physical size or complexity of an organism. For example, a single copy of the human genome contains approximately 3 x 10^9 DNA bases, while a single wheat genome contains 17 x 10^9 DNA bases. This apparent paradox (called the c-value paradox) can be explained by the fact that not all DNA encodes genes. In fact, in many organisms, genes are separated from each other by long stretches of DNA that do not code for a protein. Some of this "non-coding" DNA may be transposons, which are an interesting class of self-replicating DNA elements discussed in more detail in a subsequent chapter.



organism	DNA content (Mb, 1C)	estimated gene number	average gene density	chromsome number (1N)
Homo sapiens	3,200	25,000	100,000	23
Mus musculus	2,600	25,000	100,000	20
Drosophila melanogaster	140	13,000	9,000	4
Arabidopsis thaliana	130	25,000	4,000	5
Caenorhabditis elegans	100	19,000	5,000	6
Saccharomyces cerevisiae	12	6000	2,000	16
Escherichia coli	5	3200	1,400	1

Table 1.1 Measures of genome size in selected organisms. The DNA content (1C) is shown in millions of basepairs (Mb). Average gene density is the mean number of non-coding bases (in bp) between genes in the genome. For eukaryotes, the chromosome number is the chromosomes counted in a gamete (1N) from each organism.

MODEL ORGANISMS

We have seen already that many of the great advances in genetics were made using species that are not especially important from a medical, economic, or even ecological perspective. Geneticists, from Mendel onwards, have used model organisms for their experiments. Today, a small number of species are widely used a model genetic organisms. All of these species have characteristics that make them easy to grow in large numbers in laboratories: they are small, fast growing with a short generation time and produce lots of progeny from matings that can be easily controlled. Genetic model organisms also usually have small genomes (small c-value), and are diploid (i.e. chromosomes are present in pairs). Yeast (***Saccharomyces cerevisiae***) is a good general model for the basic functions of eukaryotic cells. The roundworm, ***Caenorhabditis elegans*** is a useful model for the development of multicellular organisms, in part because it is transparent throughout its life cycle, and its cells undergo a well-characterized series of divisions to produce the adult body. The fruit fly (***Drosophila melanogaster***) has been studied longer, and probably in more detail, than any of the other genetic model organisms still in use, and is a useful model for studying development as well as physiology and even behaviour. As a mammal, mouse (***Mus musculus***) is the model organism most closely related to humans, however some of the practical difficulties of working with mice led researchers more recently to develop zebrafish (***Danio rerio***) as a genetic model for vertebrates. Unlike mice, zebrafish embryos develop externally to their mothers and are transparent, making it easier to study their development. Finally, a small weed, ***Arabidopsis thaliana***, is the most widely studied plant genetic model organism.

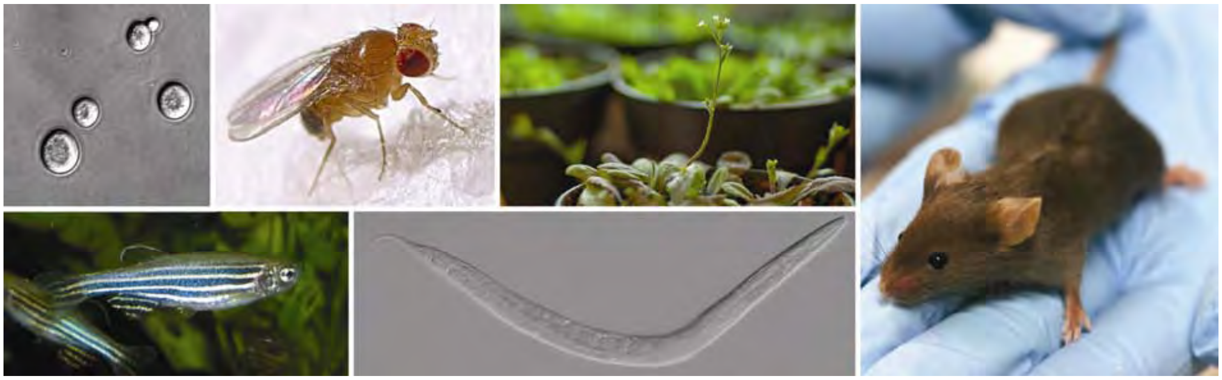


Figure 1.10 Some of the most important genetic model organisms in use today. Clockwise from top left: yeast, fruit fly, arabidopsis, mouse, roundworm, zebrafish.

The study of genetic model organisms has greatly increased our knowledge of genetics, and biology in general. Model organisms also have important implications in medical research. For example, at least 75% of the approximately 1,000 genes that have been associated with specific human diseases have highly similar sequences in both humans and ***D. melanogaster***. Information learned from model organisms about particular biochemical pathways can usually be applied to other species, since the main features of many biochemical pathways tend to be shared between species.

It is also possible, and sometimes necessary, to study biological processes in non-model organisms. Humans, for example, have none of the characteristics of a model organism, and there are some diseases or other traits for which no clear analog exists in other organisms. Some of the tools of genetic analysis can be applied to non-model organisms, especially with the development of new types of genetic mapping and whole genome sequencing.

SUMMARY

- Mendel demonstrated that heredity involved discrete, heritable factors that affected specific traits.
- A gene can be defined abstractly as unit of inheritance; many genetic experiments can be conducted without a knowledge of DNA.
- The ability of DNA from bacteria and viruses to transfer genetic information into bacteria helped prove that DNA is the genetic material.
- DNA is a double helix made of two anti-parallel strands of bases on a sugar-phosphate backbone.
- Specific bases on opposite strands pair through hydrogen bonding, ensuring complementarity of the strands.
- The Central Dogma explains how DNA affects heritable traits.
- Not all of the DNA in an organism contains genes.
- Model organisms accelerate the use of genetics in basic and applied research in biology, agriculture and medicine.

KEY TERMS

blending inheritance
particulate inheritance
Mendel
gene
allele
trait
P, F1, F2
Griffith
Avery, MacLeod, McCarty
Hershey and Chase
DNase
proteinase
^{35}S
^{32}P

bacteriophage
Chargaff's Rules
Watson and Crick
DNA bases
sugar-phosphate backbone
anti-parallel
complementary
hydrogen bond
minor groove
major groove
adenine
cytosine
thymine
guanine

Central Dogma
transcription
translation
RNA
genome
c-value paradox
model organism
Saccharomyces cerevisiae
Caenorhabditis elegans
Drosophila melanogaster
Mus musculus
Danio rerio
Arabidopsis thaliana
Escherichia coli

STUDY QUESTIONS

1.1 How would the results of the cross in Figure 1.2 have been different if heredity worked through blending inheritance rather than particulate inheritance?

1.2 Imagine that astronauts provide you with living samples of multicellular organisms discovered on another planet. These organisms reproduce with a short generation time, but nothing else is known about their genetics.

a) How could you define laws of heredity for these organisms?

b) How could you determine what molecules within these organisms contained genetic information?

c) Would the mechanisms of genetic inheritance likely be similar for all organisms from this planet?

d) Would the mechanisms of genetic inheritance likely be similar to organisms from earth?

1.3 It is relatively easy to extract DNA and protein from cells; biochemists had been doing this since at least the 1800's. Why then did Hershey and Chase need to use radioactivity to label DNA and proteins in their experiments?

1.4 Compare Watson and Crick's discovery with Avery, MacLeod and McCarty's discovery.

a) What did each discover, and what was the impact of these discoveries on biology?

b) How did Watson and Crick's approach generally differ from Avery, MacLeod and McCarty's?

c) Briefly research Rosalind Franklin on the internet. Why is her contribution to the structure of DNA controversial?

1.5 Starting with mice and R and S strains of *S. pneumoniae*, what experiments in additional to those shown in Figure 1.4 to demonstrate that DNA is the genetic material?

1.6 List the information that Watson and Crick used to deduce the structure of DNA.

1.7 **a)** List the defining characteristics of the structure of a DNA molecule.

b) Which of these characteristics are most important to replication?

c) Which characteristics are most important to the Central Dogma?

1.8 Refer to Table 1.1

a) What is the relationship between DNA content of a genome, number of genes, gene density, and chromosome number?

b) What feature of genomes explains the c-value paradox?

c) Do any of the numbers in Table 1.1 show a correlation with organismal complexity?

1.9 **a)** List the characteristics of an ideal model organism.

b) Which model organism can be used most efficiently to identify genes related to:

i) eye development
ii) skeletal development
iii) photosynthesis
iii) cell division
iv) cell differentiation
v) cancer

1.10 Refer to Figure 1.8

a) Identify the part of the DNA molecule that would be radioactively labeled in the manner used by Hershey & Chase

b) DNA helices that are rich in G-C base pairs are harder to separate (e.g. by heating) that A-T rich helices. Why?

Chapter 2 Chromosomes, Mitosis, and Meiosis

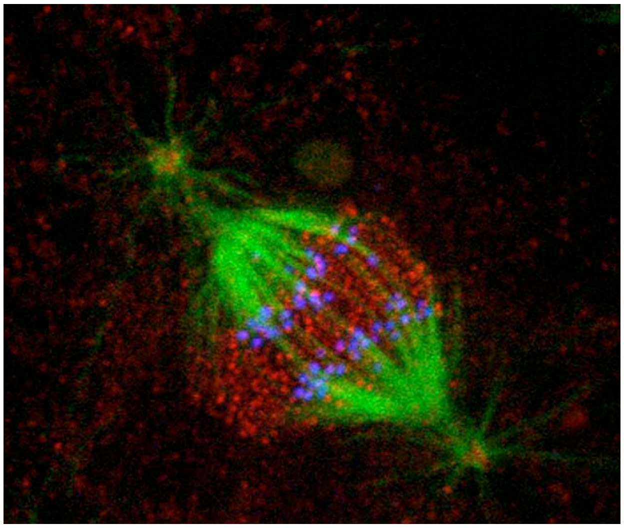


Figure 2.1 Moving chromosomes (blue) towards the poles at anaphase requires many proteins (red), all of which interact with microtubules (green).

Chromosomes contain genetic information. We often take this fact for granted, but just over a century ago, even the best biologists in the world were uncertain of the function of these rod-shaped structures. We now know that most chromosomes contain a single molecule of double-stranded DNA that is complexed with proteins. This arrangement allows very long DNA molecules to be compacted into a small volume that can more easily be moved during mitosis and meiosis (Fig 2.1). The compact structure also makes it easier for pairs of chromosomes to align with each other during meiosis. Finally, we shall see that chromosomal structure can affect whether genes are active or silent.

Chromosomes may be loose or compact

If stretched to its full length, the DNA molecule of the largest human chromosome would be 85mm. Yet during mitosis and meiosis, this DNA molecule is compacted into a chromosome approximately 5µm long. Although this compaction makes it easier to transport DNA within a dividing cell, it also makes DNA less accessible for other cellular functions such as DNA synthesis and transcription. Thus, chromosomes vary in how tightly DNA is packaged, depending on the stage of the cell cycle and also depending on the level of gene activity required in any particular region of the chromosome.

There are several different levels of structural organization in eukaryotic chromosomes, with each successive level contributing to the further compaction of DNA (Fig. 2.2). For more loosely compacted DNA, only the first few levels of organization may apply. Each level involves a specific set of proteins that associate with the DNA to compact it. First, proteins called the **core histones** act as spool around which DNA is coiled twice to form a structure called the **nucleosome**. Nucleosomes are formed at regular intervals along the DNA strand, giving the molecule the appearance of "beads on a string". At the next level of organization, **histone H1** helps to compact the DNA strand and its nucleosomes into a **30nm fibre**. Subsequent levels of organization involve the addition of **scaffold proteins** that wind the 30nm fibre into coils, which are in turn wound around other scaffold proteins.

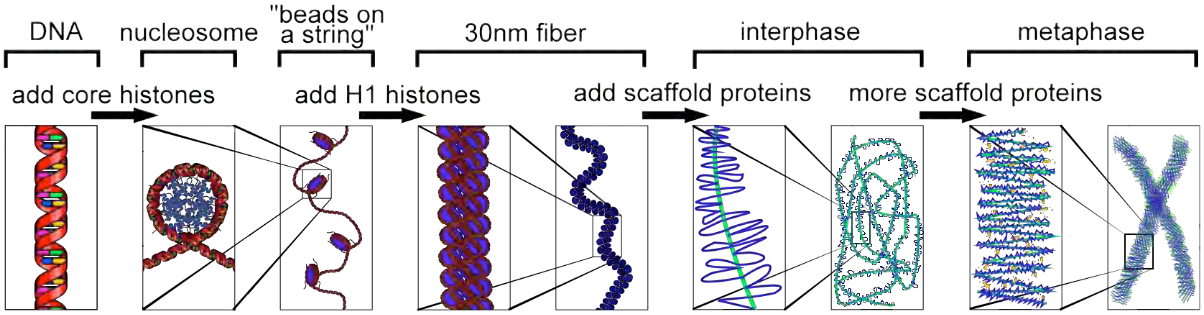


Figure 2.2 Successive stages of chromosome compaction depend on the introduction of additional proteins.

Chromosomes stain very intensely with some types of dyes, which is how they got their name (chromosome means "colored body"). Certain dyes stain some regions within a chromosome more intensely that others, giving some chromosomes a banded appearance. The material that makes up chromosomes, which we now know to be proteins and DNA, is called **chromatin**. There are two general types of chromatin. **Euchromatin** is more loosely packed, and tends to contain more genes that are being transcribed, as compared to the more densely compacted **heterochromatin** which is rich in short, repetitive sequences called **microsatellites**.

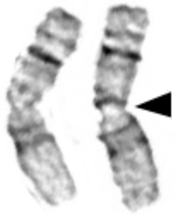
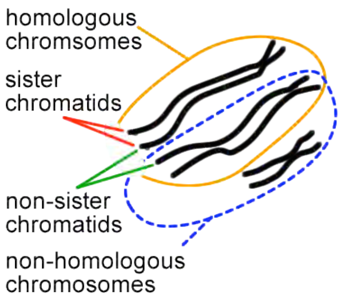


Figure 2.3 A pair of metacentric chromosomes. The arrow shows a centromeric region.

Chromosomes also contain other distinctive features such as centromeres and telomeres. Both of these are heterochromatic. In most cases, each chromosome contains one **centromere**. These sequences are bound by centromeric proteins that link the centromere to microtubules that transport chromosomes during cell division. Under the microscope, centromeres can sometimes appear as constrictions in the body of the chromosome (Fig. 2.3). If a centromere is located near the middle of a chromosome, it is said to be **metacentric**, while an **acrocentric** centromere is closer to one end of a chromosome, and a **telocentric** chromsome is at the very end. More rarely, in a **holocentric** centromere, no single centromere can be defined and the entire chromsome acts as the centromere. **Telomeres** are repetitive sequences near the ends of linear chromosomes, and are important in maintaining the length of the chromosomes during replication, and protecting the ends of the chromosomes from alterations.

Figure 2.4
Relationships between chromosomes and chromatids.



It is useful to describe the similarity between chromosomes using appropriate terminology (Fig 2.4). **Homologous** chromosomes are typically pairs of similar, but non-identical, chromosomes in which one member of the pair comes from the male parent, and the other comes from the female parent. Homologs contain the same genes but not necessarily the same alleles. **Non-homologous** chromosomes contain different sets of genes, and may or may not be distinguishable based on cytological features such as length and centromere position. Within a chromosomes that has undergone replication, there are **sister chromatids,** which are physically connected to each other at the centromere and remain joined until cell division. Because a pair of sister chromatids is produced by the replication of a single DNA molecule, their sequences are essentially identical. On the other hand, **non-sister chromatids** come from two separate, but homologous chromosomes, and therefore usually contain the same genes in the same order, but do not necessarily have identical DNA sequences.

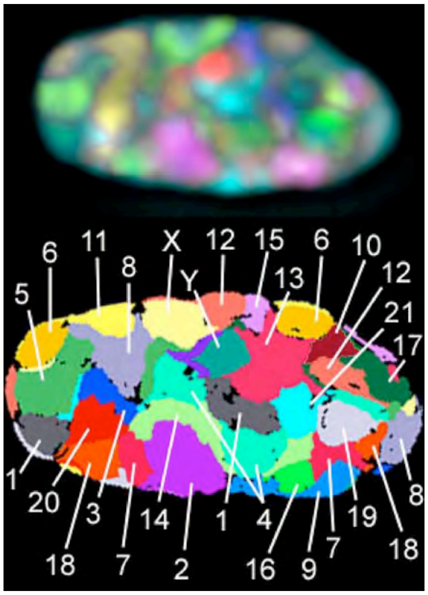


Figure 2.5

Top: FISH (Fluorescence in situ hybridization) labeling of all 24 different human chromosomes (1 - 22, X, and Y) in a fibroblast nucleus, each with a different combination of in total seven fluorochromes.
Bottom: False color representation of all chromosome territories visible in this mid-section after computer classification.

Mitosis

Cell division is essential to asexual reproduction and the development of multicellular organisms. Accordingly, the primary function of mitosis is to ensure that each daughter cell inherits identical genetic material, i.e. exactly one copy of each chromosome. To make this happen, replicated chromosomes condense (**prophase**), and are positioned near the middle of the dividing cell (**metaphase**), and then one sister chromatid from each chromosome migrates towards opposite poles of the dividing cell (**anaphase**), until the identical sets of chromosomes are completely separated from each other within the newly formed nuclei of each daughter cell (**telophase**) (see Figs. 2.5-2.7 for diagrams of the process). This is followed by the completion of the division of the cytoplasm (**cytokinesis**). The movement of chromosomes is aided by microtubules that attach to the chromosomes at centromere.

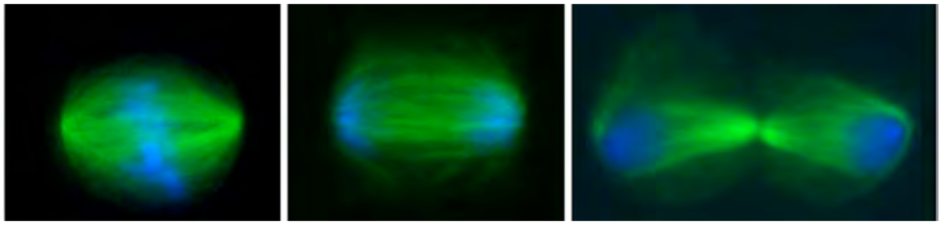


Figure 2.6 Mitosis in arabidopsis showing fluorescently labeled chromosomes (blue) and microtubules (green) at metaphase, anaphase and telophase (from left to right).

Meiosis

Meiosis, like mitosis, is also a necessary part of cell division. However, in meiosis not only do sister chromatids separate from each other, homologous chromosomes also separate from each other. This extra, reductional step of meiosis is essential to sexual reproduction. Without meiosis, the chromosome number would double in each generation of a species and would quickly become too large to be viable.

Meiosis is divided into two stages designated by the roman numerals I and II. Meiosis I is called a **reductional** division, because it reduces the number of chromosomes inherited by each of the daughter cells. Meiosis I is further divided into Prophase I, Metaphase I, Anaphase I, and Telophase I, which are roughly similar to the corresponding stages of mitosis, except that in Prophase I and Metaphase I, homologous chromosomes pair with each other in transient structures called **bivalents** (Figs. 2.7, 2.8). This is an important difference between mitosis and meiosis, because it affects the segregation of alleles, and also allows for recombination to occur through crossing-over, as described later in the course. During Anaphase I, one member of each pair of homologous chromosomes migrates into a daughter cell. Meiosis II is essentially the same as mitosis, with one sister chromatid from each chromosome separating to produce two identical cells. Because Meiosis II, like mitosis, results in products that contain identical sequences, Meiosis II is called an **equational** division.

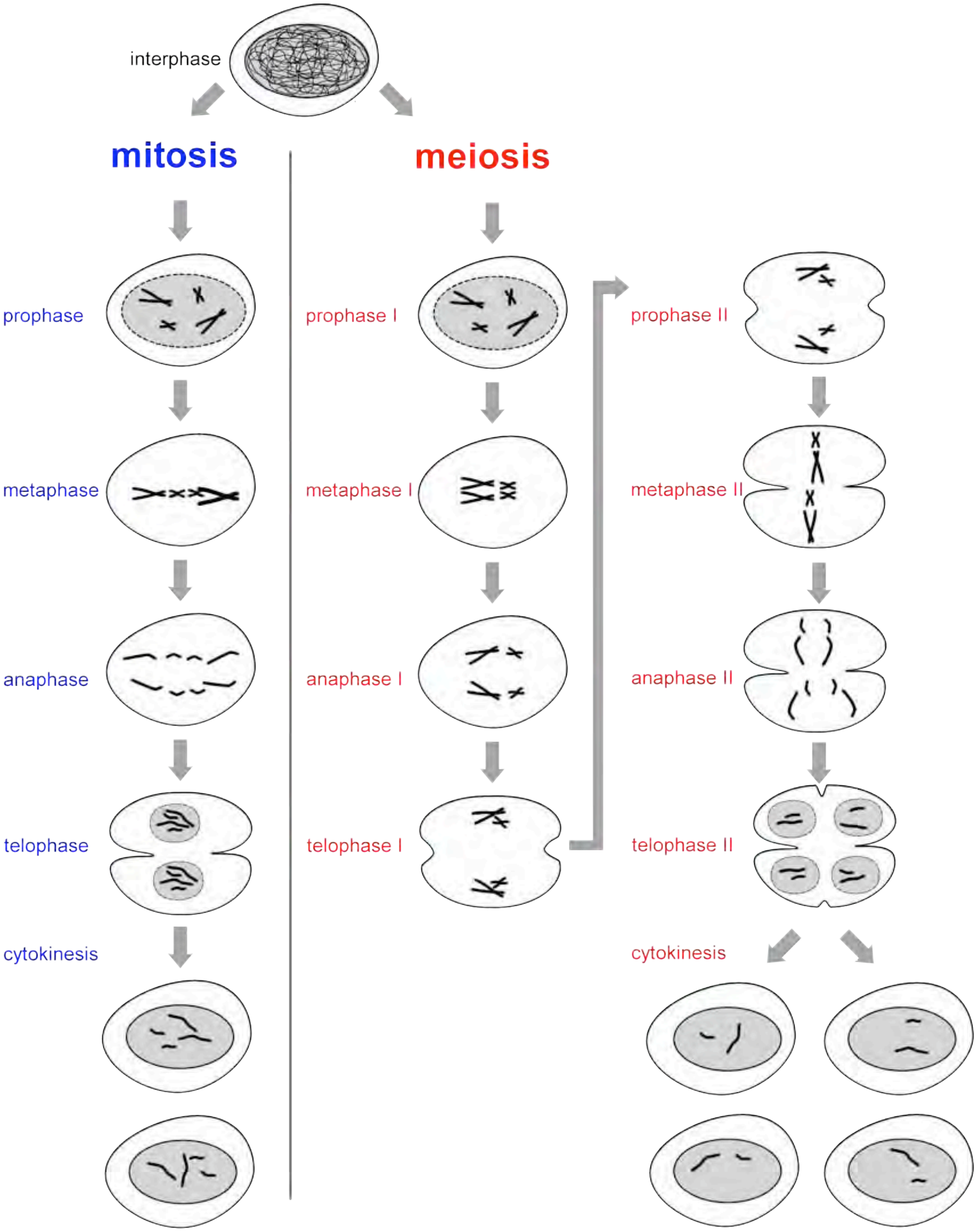
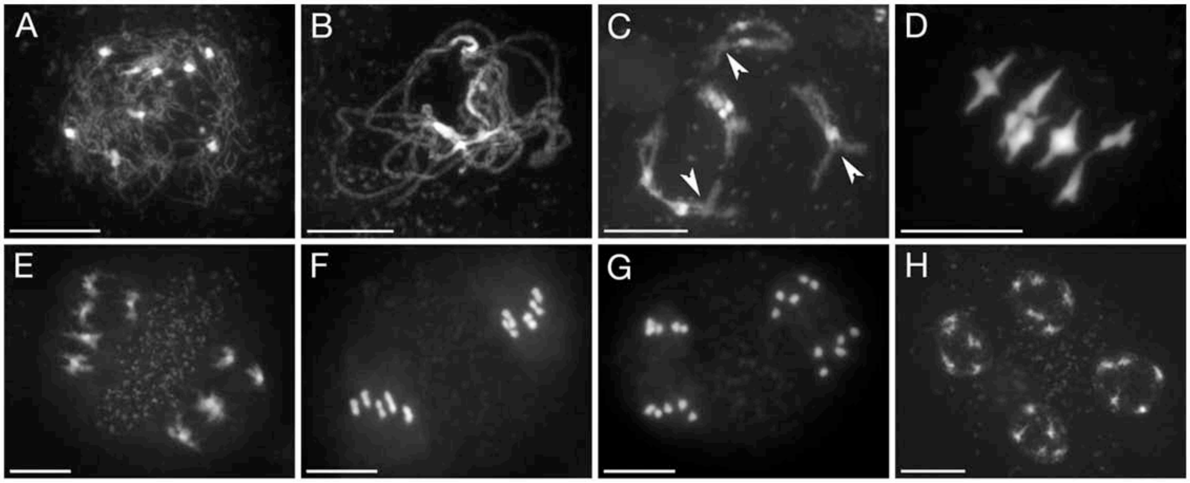


Figure 2.7 Mitosis and meiosis. Note the similarities and differences between metaphase in mitosis and metaphase I and II of meiosis.

Figure 2.8 Meiosis in Arabidopsis (n=5). Panels A-C show different stages of prophase I, each with an increasing degree of chromosome condensation. Subsequent phases are shown: metaphase I (D), telophase I (E), metaphase II (F), anaphase II (G), and telophase II (H).



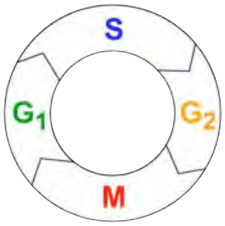


Figure 2.9 A typical eukaryotic cell cycle.

THE CELL CYCLE AND CHANGES IN DNA CONTENT

The life cycle of an eukaryotic cell can generally be divided into at least four stages (Fig. 2.9). When a cell is produced through fertilization or cell division, there is usually a lag before it undergoes DNA synthesis. This lag period is called Gap 1 (**G_1**), and ends with the onset of the DNA synthesis (**S**) phase, during which each chromosome is replicated. Following replication, there may be another lag, called Gap 2 (**G_2**), before mitosis (**M**). Cells undergoing meiosis do not usually have a G_2 phase. Interphase is as term used to describe all phases of the cell cycle excluding mitosis or meiosis. A typical cell cycle is shown in Fig. 2.9. My variants of this generalized cell cycle also exist. Some cells never leave G_1 phase, and are said to enter a permanent, non-dividing stage called **G_0**. On the other hand, some cells undergo many rounds of DNA synthesis (S) without any mitosis or cell division. These endoreduplicated cells are described later in this chapter. Understanding the control of the cell cycle is an active area of research, particularly because of the relationship between cell division and cancer.

The amount of DNA within a cell changes following each of the following events: fertilization, DNA synthesis, mitosis, and meiosis (Fig 2.10). We use "**c**" to represent the DNA content in a cell, and "**n**" to represent the number of complete sets of chromosomes. In a gamete (i.e. sperm or egg), the amount of DNA is 1c, and the number of chromosomes is 1n. Upon fertilization, both the DNA content and the number of chromosomes doubles to 2c and 2n, respectively. Following DNA synthesis, the DNA content doubles again to 4c, but each pair of sister chromatids is still counted as a single chromosome, so the number of chromosomes remains unchanged at 2n. If the cell undergoes mitosis, each daughter cell will be 2c and 2n, because it will receive half of the DNA, and one of each pair of sister chromatids. In contrast, the cells that are produced from the meiosis of a 2n, 4c cell are 1c and 1n, since each pair of sister chromatids, and each pair of homologous chromosomes divides during meiosis.

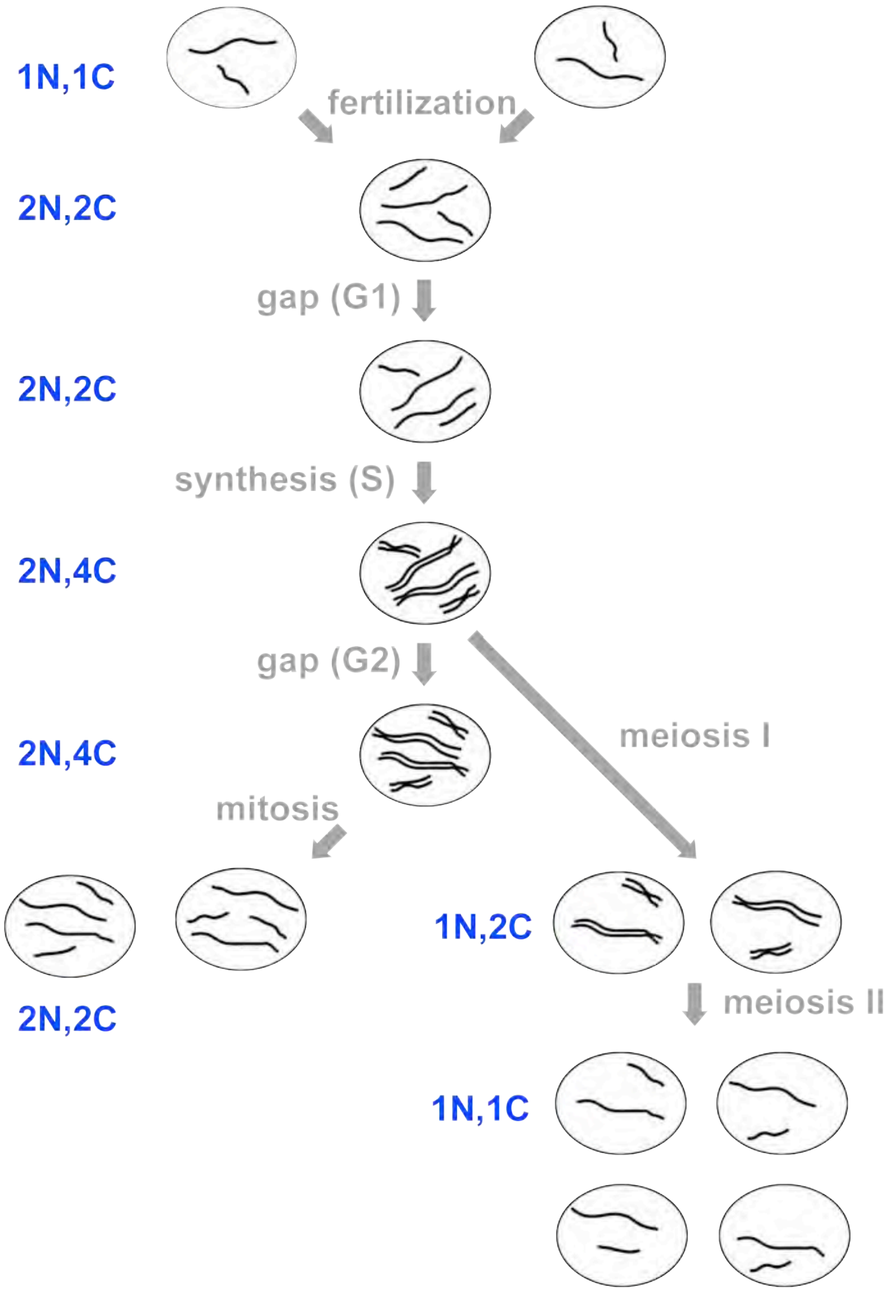


Figure 2.10 Changes in DNA and chromosome content during the cell cycle. For simplicity, nuclear membranes are not shown, and all chromosomes are represented in a similar stage of condensation.

KARYOTYPES SHOW CHROMOSOME NUMBER AND STRUCTURE

Each eukaryotic species has its total nuclear genome divided among a number of chromosomes that is characteristic of that species. For example, a haploid human nucleus (i.e. sperm or egg) normally has 23 chromosomes (n=23), and a diploid human nucleus has 23 pairs of chromosomes (2n=46). Various stains and fluorescent dyes produce characteristic banding patterns on some chromosomes. This can make it easier to identify specific chromosomes. The number of chromosomes varies between species (see Table 1.1), but there appears to be very little correlation between chromosome number and either the complexity of an organism or its total amount genomic DNA.

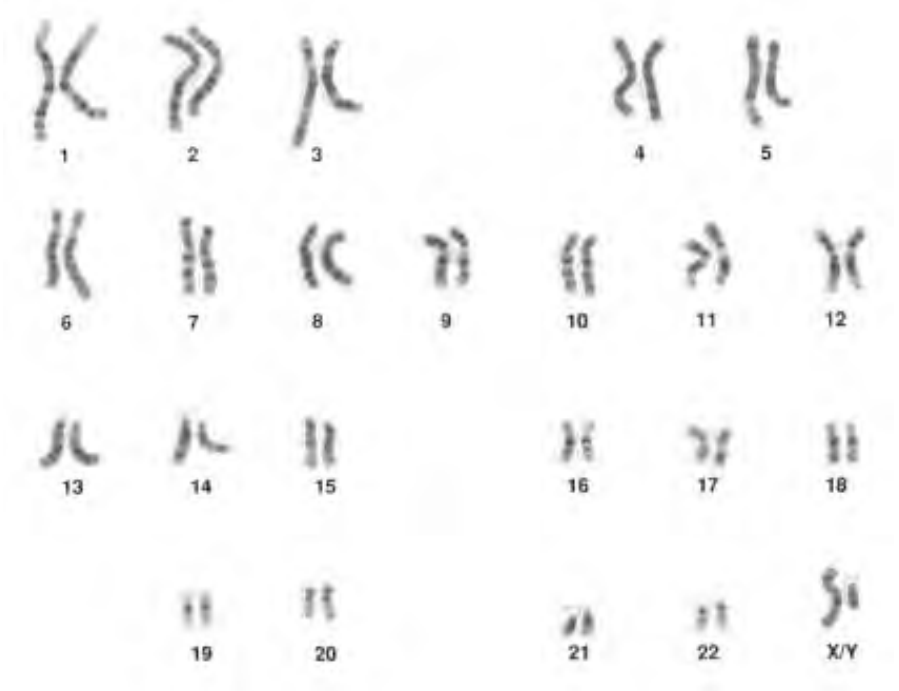


Figure 2.11 Karyotype of a normal human male

A **karyotype** shows the complete set of chromosomes of an individual (Fig. 2.11). Analysis of karyotypes can identify of chromosomal abnormalities, including **aneuploidy**, which is the addition or subtraction of a chromosome from a pair of homologs. More specifically, the absence of one member of a pair of homologous chromosomes is called **monosomy**. On the other hand, in a **trisomy**, there are three, rather than two homologs of a particular chromosome. Different types of aneuploidy are sometimes represented symbolically; if **2n** symbolizes the normal number of chromosomes in a cell, then **2n-1** indicates monosomy and **2n+1** represents trisomy.

The most familiar human aneuploidy is trisomy-21 (i.e. three copies of chromosome 21), which is one cause of **Down's syndrome**. Most (but not all) other human aneuploidies are lethal at an early stage of embryonic development. Note that aneuploidy usually affects only one set of homologs within a karyotype, and is therefore distinct from **polyploidy**, in which the entire karyotype is duplicated (see below). Aneuploidy is almost always deleterious, whereas polyploidy appears to be beneficial in some organisms, particularly some species of plants.

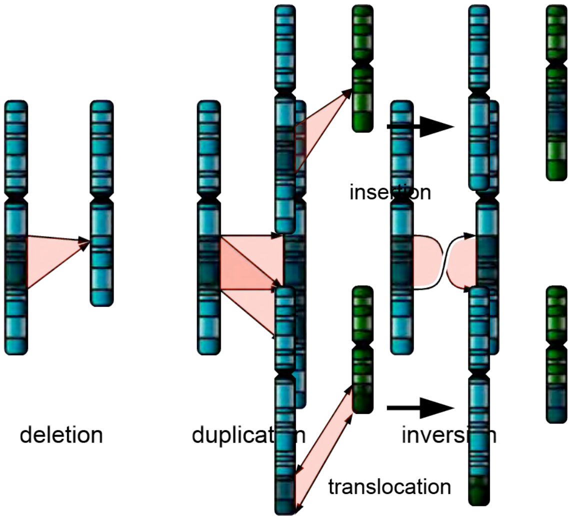


Figure 2.12

Structural abberations in chromosomes.

Structural defects in chromosomes are another type of abnormality that can be detected in karyotypes (Fig 2.12). These defects include **deletions, duplications**, and **inversions**, which all involve changes in a segment of a single chromosome. **Insertions** and **translocations** involve two non-homologous chromosomes. In an insertion, DNA from one chromosome is unidirectional, while in translocation, the transfer of chromosomal segments is bidirectional. Structural defects affect only part of a chromosome, and so tend to be less harmful than aneuploidy. In fact, there are many examples of ancient chromosomal rearrangements in the genomes of species including our own. Duplications of some small chromosomal segments, in particular, may have some evolutionary advantage by providing extra copies of some genes, which can then evolve in new ways.

Chromosomal abnormalities arise in many different ways. Many of these can be traced to rare errors in natural cellular processes. **Non-disjunction** is the failure of at least one pair of chromosomes or chromatids to separate during mitosis or meiosis. **Chromosome breakage** also occurs infrequently as the result of physical damage (such as radiation), movement of some types of transposons, and other factors. During the repair of a broken chromosome, deletions, insertions, translocations and even inversions can be introduced.

POLYPLOIDY

Humans, like most animals and all of the eukaryotic genetic model organisms in wide use, are diploid. This means that most of their cells have two homologous copies of each chromosome. In contrast, many plant species and even a few animal species are **polyploids**. This means there are more than two homologs of each chromosome in each cell.

When describing polyploids, we use the letter "**x**" to define the level of ploidy. A diploid is 2x, because there are two basic sets of chromosomes, and a tetraploid is 4x, because it contains four copies of each chromosome. For clarity when discussing polyploids, geneticists will often combine the "x" notation with the "n" notation already defined previously in this chapter. Thus for both diploids and polyploids, "n" is the number of chromosomes in a gamete, and "2n" is the number of chromosomes following fertilization. For a diploid, therefore, n=x, and 2n=2x. For a tetraploid, n=2x, and 2n=4x.

Like diploids (2n=2x), stable polyploids generally have an even number of copies of each chromosome: tetraploid (2n=4x), hexaploid (2n=6x), and so on. The reason for this is clear from a consideration of meiosis. Remembering that the purpose of meiosis is to reduce the sum of the genetic material by half, meiosis can equally divide an even number of chromosome sets, but not an odd number. Thus, polyploids with an uneven number of chromosomes (e.g. triploids, 2n=3x) tend to be sterile, even if they are otherwise healthy. The mechanism of meiosis in stable polyploids is essentially the same as in diploids: during metaphase I, homologous chromosomes pair with each other. Depending on the species, all of the homologs may be aligned together at metaphase, or in multiple separate pairs. For example, in a tetraploid, some species may form **tetravalents** in which the four homologs from each chromosome align together, or alternatively, two pairs of homologs may form two bivalents. Note that because that mitosis does not involve any pairing of homologous chromosomes, mitosis is equally effective in diploids, even-number polyploids, and odd-number polyploids.

Triploidy is used in the production of seedless fruits, such as watermelon, grapes and bananas. All of the tissues of these fruit are triploid. Because almost all of the cells of the plant, including its fruit, are produced through mitosis, the uneven number of chromosome sets does not affect their development. However, cells that contribute to the production of gametes are produced through meiosis, and because the triploids are unable to complete normal meioses, their gametes fail to develop, so no zygotes are formed, and the seeds (which normally contain embryos that develop from the zygotes) are aborted.

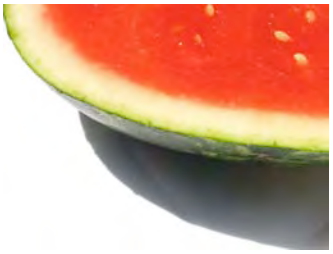


Figure 2.13. Part of a triploid watermelon, showing white, aborted seeds within the flesh

If triploids cannot make seeds, how do we obtain enough triploid individuals for cultivation? The answer depends on the plant species involved. In some cases, such as banana, it is possible to propagate the plant asexually; new progeny can simply be grown from cuttings from a triploid plant. On the other hand, seeds for seedless watermelon are produced sexually: a tetraploid watermelon plant is crossed with a diploid watermelon plant. Both the tetraploid and the diploid are fully fertile, and produce gametes with two (1n=2x) or one (1n=1x) sets of chromosomes, respectively. These gametes fuse to produce a zygote (2n=3x), that is able to develop normally into an adult plant through multiple rounds of mitosis, but is unable to compete normal meiosis or produce seeds.

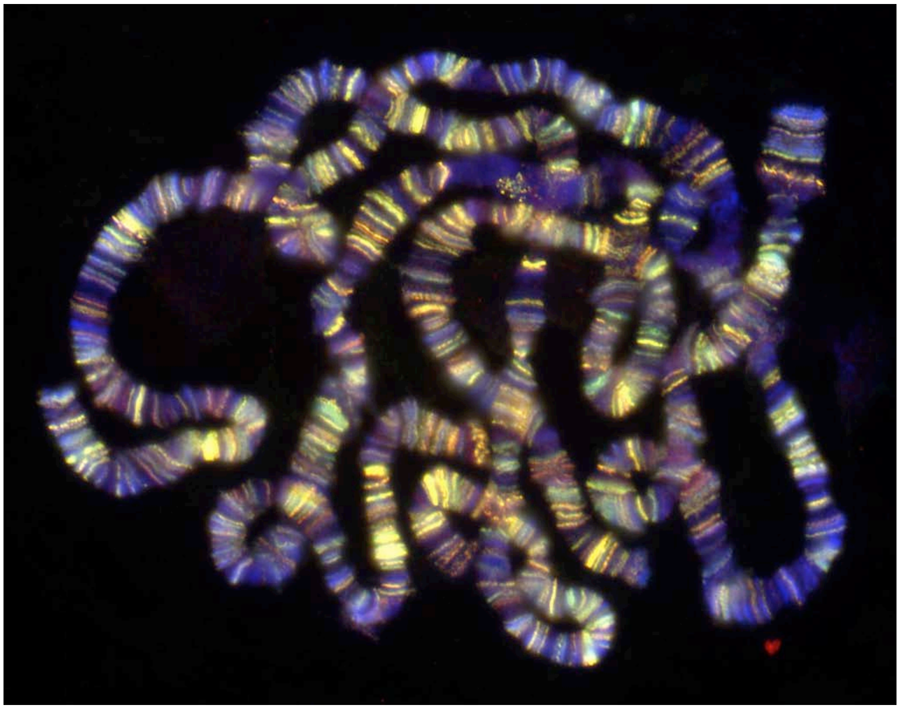


Figure 2.14. Endoreduplicated chromosomes from an insect salivary gland. The banding pattern is produced with fluorescent labels.

ENDOREDUPLICATION

Endoreduplication, also known as **endopolyploidy**, is a special type of tissue-specific genome amplification that occurs in many types of plant cells and in specialized cells of some animals including humans. Endoreduplication does not affect the germline or gametes, so species with endopolyploidy are not considered polyploids. Endopolyploidy occurs when a cell undergoes multiple rounds of DNA synthesis (S-phase) without any mitosis. This produces multiple chromatids of each chromosome. Endopolyploidy seems to be associated with cells that are metabolically very active, and produce a lot of enzymes and other proteins in a short amount of time. The highly endoreduplicated salivary gland chromosomes of *D. melanogaster* have been useful research models in genetics, since their relatively large size makes them easy to study under the microscope.

GENE BALANCE

Why do trisomies, duplications, and other chromosomal abnormalities that increase gene copy number sometimes have a negative effect on the normal development or physiology of an organism? This is particularly intriguing because in many species, aneuploidy is detrimental or lethal, while polyploidy is tolerated or even beneficial. The answer is probably related to the concept of **gene balance**, which can be summarized as follows: genes, and the proteins they produce, have evolved to be part of complex metabolic and regulatory networks. Some of these networks function best when certain enzymes and regulators are present in specific ratios to each other. Increasing the gene copy number for just one part of the network may throw the network out of balance, leading to increases or decreases of certain metabolites, which may be toxic in high concentrations or which may be limiting in other important processes in the cell. The activity of genes and metabolic networks is regulated in many different ways besides changes in gene copy number, so duplication of just a few genes will usually not be harmful. However, trisomy and large segmental duplications of chromosomes affect the dosage of so many genes that cellular networks are unable to adjust to the changes.

ORGANELLAR GENOMES

Chromosomes also exist outside of the nucleus, within both the chloroplast and mitochondria. These organelles are likely the remnants of a prokaryotic endosymbionts that entered the cytoplasm of ancient progenitors of today's eukaryotes. These endosymbionts had their own, circular chromosomes, like most bacteria that exist today. Likewise, chloroplasts and mitochondria also have circular chromosomes that behave more like bacterial chromosomes than eukaryotic chromosomes, i.e. these organellar genomes do not undergo mitosis or meiosis. Organellar genomes are also often present in multiple copies within each organelle, and in most species are inherited maternally.

SUMMARY

- Chromosomes are complex and dynamic structures consisting of DNA and proteins (chromatin).

- The degree of chromatin compaction varies between heterochromatic and euchromatic regions and between stages of the cell cycle.

- Some chromosomes can be distinguished cytologicaly based on their length, centromere position, and banding patterns when stained dyes or labelled with sequence-specific probes.

- Homologous chromosomes contain the same genes, but not necessarily the same alleles. Sister chromatids usually contain the same genes and the same alleles.

- Mitosis reduces the c-number, but not the n-number. Meiosis reduces both c and n.

- Homologous chromosomes associate with each other during meiosis, but not mitosis.

- Several types of structural defects in chromosomes occur naturally, and can affect cellular function and even evolution.

- Aneuploidy results from the addition or subtraction of one or more chromosomes from a group of homologs, and is usually deleterious to the cell.

- Polyploidy is the presence of more than two complete sets of chromosomes in a genome. Even-numbered multiple sets of chromosomes can be stably inherited in some species, especially plants.

- Endopolyploidy is tissue-specific type of polyploidy observed in some species, including diploids.

- Both aneuploidy and structural defects such as duplications can affect gene balance.

- Organelles also contain chromosomes, but these are much more like prokaryotic chromosomes than the nuclear chromosomes of eukaryotes.

KEY TERMS

chromosome
core histones
nucleosome
30nm fiber
histone H1
scaffold proteins
heterochromatin
euchromatin
microsatellite
chromatid
centromere
metacentric
acrocentric
telocentric
holocentric
telomere
homolog
non-homologous
chromatid
sister chromatid
non-sister chromatid
interphase

mitosis
prophase
metaphase
anaphase
telophase
prophase
meiosis
prophase (I, II)
metaphase (I, II)
anaphase (I, II)
telophase (I, II)
cytokinesis
bivalent
reductional division
equational division
G_1
G_2
S
M
G_0
interphase
n

c
karyotype
aneuploidy
monsomic
trisomic
Down's syndrome
deletion
duplication
insertion
inversion
translocation
non-disjunction
chromosome breakage
polyploid
x
tetravalent
endoreduplication
endopolyploidy
gene balance
cellular network
endosymbiont
organellarchromo

STUDY QUESTIONS

2.1 Define chromatin. What is the difference between DNA, chromatin and chromosomes?

2.2 Species A has n=4 chromosomes and Species B has n=6 chromosomes. Can you tell from this information which species has more DNA? Can you tell which species has more genes?

2.3 The answer to question 2 implies that not all DNA within a chromosome encodes genes. Can you name any examples of chromosomal regions that contain relatively few genes?

2.4
a) How many centromeres does a typical chromosome have?
b) What would happen if there was more than one centromere per chromosome?
c) What if a chromosome had zero centromeres?

2.5 For a diploid with 2n=16 chromosomes, how many chromosomes and chromatids are per cell present in the gamete, and zygote and immediately following G_1, S, G_2, mitosis, and meiosis?

2.6 Bread wheat (*Triticum aestivum*) is a hexaploid. Using the nomenclature presented in class, an egg cell of wheat has n=21 chromosomes. How many chromosomes in a zygote of bread wheat?

2.7 For a given gene:
a) What is the maximum number of alleles that can exist in a 2n cell of a given diploid individual?
b) What is the maximum number of alleles that can exist in a 1n cell of a tetraploid individual?
c) What is the maximum number of alleles that can exist in a 2n cell of a tetraploid individual?
d) What is the maximum number of alleles that can exist in a population?

2.8
a) Why is aneuploidy more often lethal than polyploidy?
b) Which is more likely to disrupt gene balance: polyploidy or duplication?

2.9 For a diploid organism with 2n=4 chromosomes, draw a diagram of all of the possible configurations of chromosomes during normal anaphase I, with the maternally and paternally derived chromosomes labelled.

2.10 For a triploid organism with 2n=3x=6 chromosomes, draw a diagram of all of the possible configurations of chromosomes at anaphase I (it is not necessary label maternal and paternal chromosomes).

2.11 For a tetraploid organism with 2n=4x=8 chromosomes, draw all of the possible configurations of chromosomes during a normal metaphase.

Chapter 3 GENETICS OF ONE LOCUS

Figure 3.1 Pea plants were used in the discovery of some fundamental laws of genetics.

Before Mendel, the basic rules of heredity were not understood. For example, it was known that green-seeded pea plants occasionally produced offspring that had yellow seeds; but were the hereditary factors that controlled seed color somehow changing from one generation to the next, or were certain factors disappearing and reappearing? And did the same factors that controlled seed color also control things like plant height?

MENDEL'S FIRST LAW

Through careful study of patterns of inheritance, Mendel recognized that a single gene could exist in one or more versions, or **alleles**, even within an individual plant or animal. For example, he found two alleles of a gene for seed color: one allele gave green seeds, and the other gave yellow seeds. Mendel also observed that although different alleles could influence a single trait, they remained independent and could be inherited separately. This is the basis of **Mendel's First Law**, also called **The Law of Equal Segregation**, which states: during gamete formation, the two members of a gene pair segregate from each other;

each gamete has an equal probability of containing either member of the gene pair.

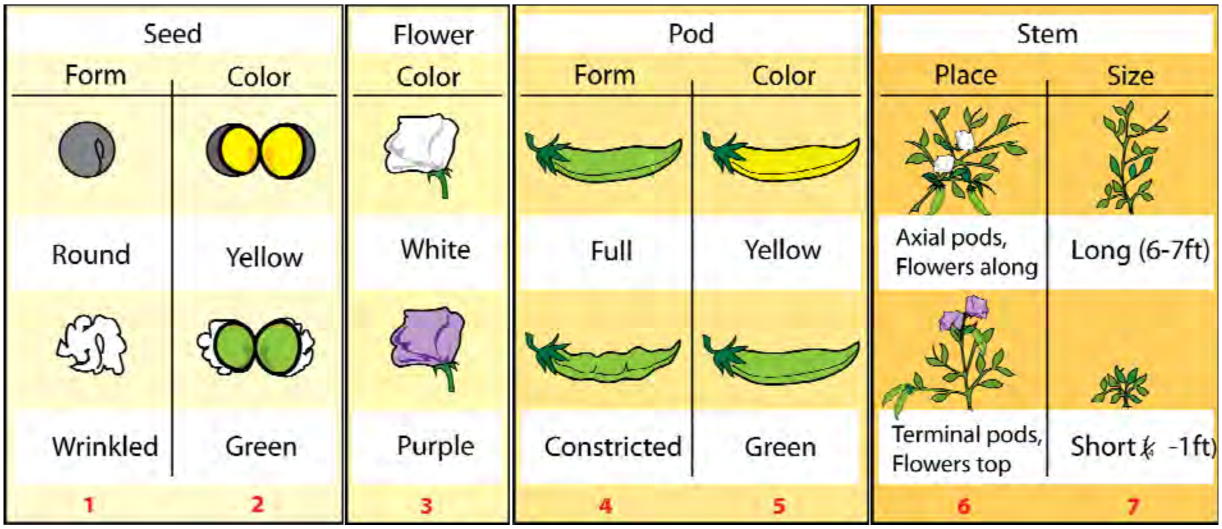


Figure 3.2 Seven traits Mendel studied in peas.

Mendel's First Law is especially remarkable because he made his observations without knowing about the relationships between genes, chromosomes, and DNA. We now know that the reason that more than one allele of a gene can be present in an individual is that most eukaryotic organisms have at least two sets of homologous chromosomes. For organisms that are predominantly diploid, such as humans or Mendel's peas, chromosomes exist in pairs, with one homolog inherited from each parent. Diploid cells can therefore hold up to two different alleles of each gene, with one allele on each member of a pair of homologous chromosomes. If both alleles of a particular gene are identical, the individual is said to be **homozygous** for that gene. On the other hand, if the alleles are different from each other, the genotype is **heterozygous**. In some cases, such as when describing genes on the sex chromosomes of a human male, we use the term **hemizygous**. This is because the cell contains only one X and one Y chromosome, so there is only one possible allele for each gene. Although a typical diploid individual can have at most two different alleles of a particular gene, more than two alleles can exist in a population of individuals. The most common allele in a natural population is called the **wild-type** allele.

RELATIONSHIPS BETWEEN ALLELES AND PHENOTYPES

A specific position on a chromosome is called a **locus**. Most of the loci geneticists discuss are occupied by genes, and the terms locus and gene are often used interchangeably. The complete set of alleles at any loci of interest in an individual define its **genotype**. The detectable effect of these alleles on the structure or function of that individual is called its

phenotype. The phenotype studied in any particular genetic experiment may range from simple, visible traits such as hair color, to more complex phenotypes including disease susceptibility or behavior.

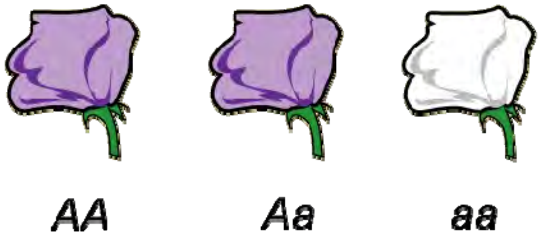


Figure 3.3 Relationship between genotype and phenotype for a dominant allele.

Let us return to an example of a simple phenotype: flower color in Mendel's peas. We have already said that one allele produces purple flowers, while the other allele produces white flowers (Figure 3.3). But what does this tell us about the flower color of an individual that has one purple allele and one white allele, in other words, what is the **phenotype** of an individual whose genotype is heterozygous? We know from experimental data that individuals heterozygous for the purple and white alleles of the flower color gene have purple flowers. The allele associated with purple color is therefore said to be **dominant** to the allele that produces the white color. The other allele, whose phenotype is masked by a dominant allele in a heterozygote, is called **recessive**. Often, a dominant allele will be represented by a capital letter (e.g. *A*) while a recessive allele will be represented in lower case (e.g. *a*). However, many different systems of genetic symbols are in use (Table 3.1), so it is important to understand the different types of notation and to use them consistently.

Table 3.1 Examples of symbols used to represent alleles, and their dominance relationships

examples of genotypes of heterozygotes	typical interpretation
Aa	Uppercase letter(s) most often represents dominant allele, and lowercase letter(s) indicates recessive allele.
a^+aw^{c2}	Lowercase letters with + superscript represent wild-type allele, which is often dominant. A lower case letter with other superscripts (or with no superscripts) represents other alleles, which are often recessive.
A_1A_2	The name of the locus is in uppercase or lower case letters, with subscripts or superscripts used to indicate different alleles. Generally, neither allele is completely dominant, or else the dominance relationships are unknown.

Besides dominance and recessivity, other relationships can exist between alleles. In **semi-dominance** (also called **incomplete dominance,** Figure 3.4), both alleles affect the trait additively, and the phenotype of the heterozygote is intermediate between either of the homozygotes. For example, alleles of for red color in carnations and some other species exhibit semi-dominance, so that a heterozygote has pink petals, while a homozygote for one allele has red petals and the other has white petals. Thus in semi-dominance, the dosage of the alleles affects the phenotype, e.g. the phenotype of the heterozygote is approximately half as strong as the phenotype of a homozygote.

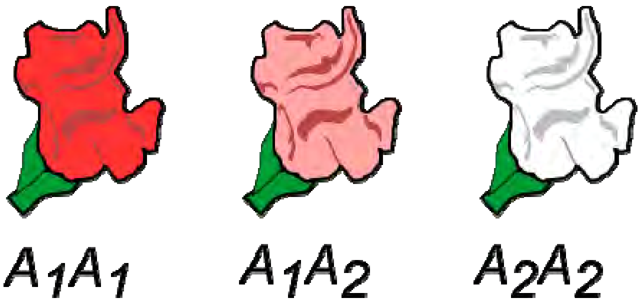


Figure 3.4
Relationship between genotype and phenotype for semi-dominant alleles.

Co-dominance is another type of allelic relationship, in which a heterozygous individual expresses the phenotype of both alleles simultaneously. An example of co-dominance is found within the ABO blood group of humans, which is controlled by a gene called *I* (Figure 3.5). There are three possible alleles at this locus: I^A, I^B, and *i*. Homozygous individuals for I^A or I^B produce only A or B type antigens, respectively, on the surface of their blood cells, and therefore have either type A or type B blood. Heterozygous I^AI^B individuals have both A and B antigens on their cell surface, and so have type AB blood. Notice that the heterozygote expresses both alleles simultaneously, and is not some kind of novel intermediate between A and B. Co-dominance is therefore distinct from semi-dominance, although they are sometimes confused. Many types of molecular markers, which we will discuss in a later chapter, display a co-dominant relationship among alleles. It is also important to note that the third allele, *i*, does not make any antigens and is recessive to all other alleles. Homozygous recessive individuals (*ii*) have type O blood. This is a useful reminder that different types of dominance relationships can exist, even for alleles of the same gene.

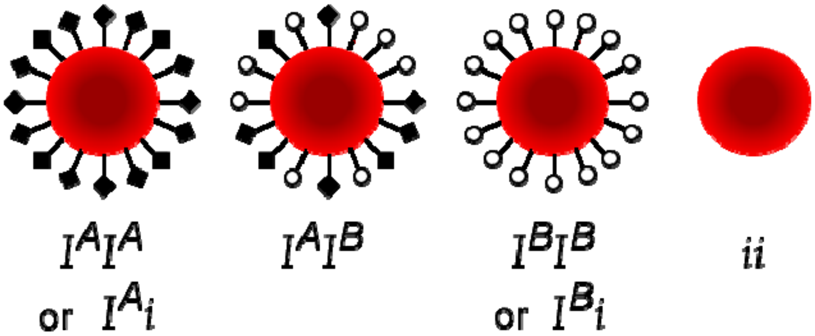


Figure 3.5 Relationship between genotype and phenotype for co-dominant alleles (I^A, I^B), and a recessive allele.

BIOCHEMICAL BASIS OF DOMINANCE

We cannot predict whether an allele will be dominant based simply on the phenotype of homozygotes; we must also know the phenotype of a heterozygote. What, then, causes dominance? There are many different biochemical conditions that may make one allele dominant to another, but one of the most common is **haplosufficiency**. Most genes produce more than enough protein to accomplish their normal biological function. If an allele is haplosufficient, then just one copy of that allele produces enough protein to have the same effect as the amount of protein produced by two copies of that allele. For the majority of genes studied, the normal (i.e. **wild-type**) alleles are haplosufficient, so even if a mutation causes a complete loss of function in one allele, the wild-type allele will be dominant and the normal function of the biochemical pathway will be retained. On the other hand, in some biochemical pathways, a single wild-type allele may be **haploinsufficient**, i.e. it may not produce enough protein to result in a normal phenotype if the other allele in a heterozygote is not functional. In this case, a non-functional allele will be dominant to a wild-type allele.

THE PUNNETT SQUARE AND MONOHYBRID CROSSES

Geneticists, including Mendel, make use of **true breeding lines** (Figure 3.6 a). These are populations of plants or animals in which all parents and their offspring have identical phenotypes over many generations with respect to a particular trait. True breeding lines are useful, because they can be assumed to be homozygous for the alleles that affect a trait of interest. When two individuals that are homozygous for the same alleles are crossed, all of their offspring will all also be homozygous.

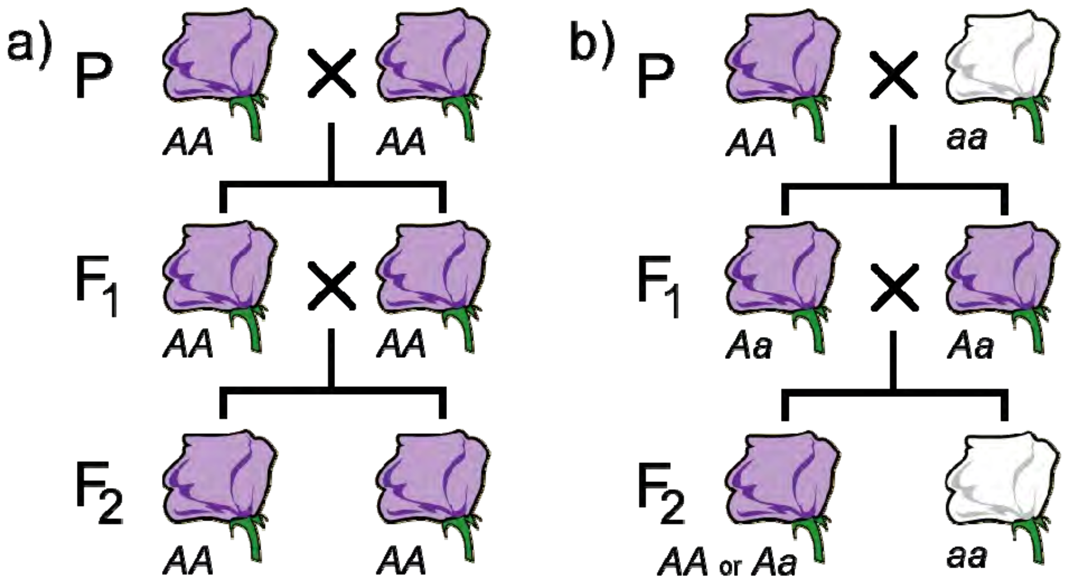


Figure 3.6 a) a true-breeding line **b)** a monohybrid cross produced by mating two pure-breeding lines

Given the genotypes of any two parents, we can predict all of the possible genotypes of the offspring. Furthermore, if we also know the dominance relationships for all of the alleles, we can predict the phenotypes of the offspring. A convenient method for calculating the expected genotypic and phenotypic ratios from a cross is the use of Punnett Square, which is a matrix in which all of the possible gametes produced by one parent are listed along one axis, and the gametes from the other parent are listed along the other axis. Each possible combination of gametes is listed at the intersection of each row and column.

When two heterozygotes are crossed, what genotypes are produced, and what is the expected frequency of each genotype? If we use the symbols *A* and *a* to represent each the two alleles of the heterozygote, then from the **Punnett Square** (Figure 3.7) we see that three genotypes are produced in a ratio of 1:2:1. If we know something about the phenotypes and dominance relationships of these alleles (as implied by the symbols used), we can predict the expected phenotypic ratio of the progeny from this cross is 3:1. A cross between two individuals that are both heterozygous at a single locus is so common in genetics that this cross is given its own name: the **monohybrid cross** (Figure 3.6 b).

	A	***a***
A	*AA*	*Aa*
a	*Aa*	*aa*

Figure 3.7 A Punnett Square showing a monohybrid cross

TEST CROSSES CAN BE USED TO DETERMINE GENOTYPES

Knowing the genotypes of an individual is usually an important part of a genetic experiment. However, genotypes cannot be observed directly; they must be inferred based on phenotypes. Because of dominance, it is often not possible to distinguish between a heterozygote and a homozgyote based on phenotype alone (e.g. see the purple-flowered F_2 plants in Figure 3.6b). To determine the genotype, a **test cross** can be performed, in which an individual with an ambiguous genotype is crossed with an individual that is homozygous recessive for all of the loci being tested. For example, if you were given a pea plant with purple flowers (and no other information), you would not know whether the plant was a homozygote (*AA*) or heterozygote (*Aa*). You could cross this purple-flowered plant to a white-flowered plant as a **tester**, since you know the genotype of the tester is *aa*. You will observe different ratios in the F_1 generation, depending on the genotype of the purple-flowered parent. If the purple-flowered parent was a homozgyote, all of the F_1 progeny will be purple. If the purple-flowered parent was a heterozygote, the F_1 progeny should segregate purple-flowered and white-flowered plants in a 1:1 ratio.

	A	***A***
a	*Aa*	*Aa*
a	*Aa*	*Aa*

	A	***a***
a	*Aa*	*aa*
a	*Aa*	*aa*

Figure 3.8 A Punnett Squares showing examples of test crosses.

SEX-LINKAGE IS AN IMPORTANT EXCEPTION TO MENDEL'S FIRST LAW

In mammals, Drosophila, and many other organisms, males have one copy of each of two different sex chromosomes (XY), while females have two copies of the same chromosome (XX) (but note that although the chromosomes have the same names, the mechanism of sex determination is very different in mammals and flies). The X and Y chromosomes do not carry the same loci, so some genes are present on the X, but not Y chromosomes. Such genes are said to be X-linked (there are very few examples of Y-linked genes, except those involved specifically in sexual differentiation). A cross involving an X-linked gene with a heterozygous female and a hemizygous male therefore produces progeny in phenotypic ratios very different from the 3:1 ratio obtained in a monohybrid cross for an **autosomal** (i.e. not sex-linked) locus. Other species with sex chromosomes may also be subject to sex-linkage, although the chromosomes may be called something other than X and Y.

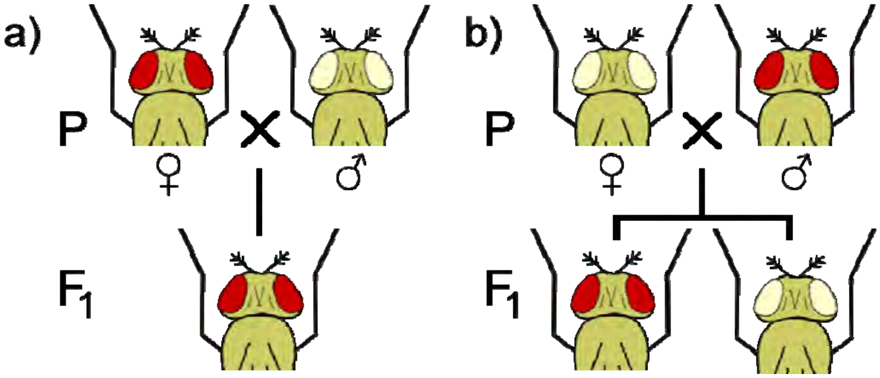


Figure 3.7 Reciprocal crosses involving a sex-linked trait.

A researcher may not know beforehand whether a novel allele is sex-linked. The definitive method to test for sex-linkage (and other forms of sex-dependent inheritance) is to conduct **reciprocal crosses** (Figure 3.7). This means to cross a male and a female that have different phenotypes, and then conduct a second set of crosses, in which the phenotypes are reversed relatively to the sex of the parents in the first cross. For example, cross a white-eyed female with a red-eyed male fly, then separately cross a red-eyed female with a white-eyed male. Note how in reciprocal crosses, the phenotypes among the progeny of sex-linked genes are different from what is expected for autosomal genes (Figure 3.8).

Figure 3.8 Reciprocal crosses involving a sex-linked trait.

	X^{w+}	X^{w+}
X^{w}	$X^{w+}X^{w}$	$X^{w+}X^{w}$
Y	$X^{w+}Y$	$X^{w+}Y$

	X^{w}	X^{w}
X^{w+}	$X^{w}X^{w+}$	$X^{w}X^{w+}$
Y	$X^{w}Y$	$X^{w}Y$

PHENOTYPE MAY BE INFLUENCES BY OTHER FACTORS BESIDES GENOTYPE

The phenotypes described in the examples used in this chapter all have nearly perfect correlation with their associated genotypes, in otherwords an individual with a particular genotype always has the expected phenotype. However. many phenotypes are not determined entirely by genetics; these are instead determined by an interaction between genotype and non-genetic, environmental factors. This **genotype-by-environment** (**G × E**) interaction is especially is especially relevant in the study of economically important phenotypes, such as human diseases or agricultural productivity. For example, a particular genotype may pre-dispose an individual to cancer, but cancer may only develop if the individual is exposed to certain DNA-damaging chemicals. Therefore, not all individuals with the particular genotype will develop the cancer phenotype.

The terms penetrance and expressivity are useful when describing the relationship between certain genotypes and their phenotypes. **Penetrance** is the proportion of individuals with a particular genotype that display a corresponding phenotype. Because all pea plants that are homozygous for the allele for white flowers (e.g. *aa* in Figure 3.3) actually have white flowers, this allele is 100% penetrant. In contrast, many human genetic diseases are incompletely penetrant, since less than 100% of individuals with the disease genotype actually develop any symptoms associated with the disease. **Expressivity** describes the range of variation in phenotypes observed in individuals with a particular genotype. Many human genetic diseases provide examples of variable expressivity, since individuals with the same genotypes may vary greatly in the severity of their symptoms. Both incomplete penetrance and variable expressivity may be due to environmental or other genetic or non-genetic factors.

DEVIATIONS FROM EXPECTED RATIOS

For a variety of reasons, the phenotypic ratios observed from real crosses rarely match the ratios expected based on a Punnett Square or other prediction techniques. There are many possible explanations for deviations from expected ratios. Sometimes these deviations are due to **sampling effects**, in other words, the random selection of a non-representative subset of individuals for observation. On the other hand, it may be that the expected ratios were calculated using incorrect assumptions. For example, a particular allele may actually be semi-dominant rather than dominant, or maybe the genotype of one parent was really heterozygous, rather than homozygous, or perhaps a gene that was assumed to be autosomal is really sex-linked.

A statistical procedure called the **chi-square** test (χ^2) can be used to help a geneticist decide whether the deviation between observed and expected ratios is due to sampling effects, or whether the difference is so large that some other explanation must be sought by re-examining the assumptions used to calculate the expected ratio. The procedure for performing a chi-square test is explained in Appendix VII of the lab manual. Essentially, the procedure calculates the difference between observed and expected frequencies of each phenotypic class, squares the difference (to remove negative values), and then adjusts the scale of this difference as a proportion of the expected frequency. The chi-square statistic is the sum of all of these values, and is therefore a standardized measure of the total deviation between observed and expected phenotypic ratios; the larger the chi-square statistic, the larger the difference between observed and expected ratios.

How do we decide whether a chi-square statistic is likely too large to be due to sampling effects alone? To do this, we compare the chi-square value for our experiment to a previously calculated probability distribution for all possible chi-square values. This distribution shows the probability of obtaining any particular chi-square value due to sampling effects. As you might expect, the distribution shows that low chi-square values are fairly common, while larger chi-square values are rare (Figure 3.9). From this distribution (or more conveniently, from a table of chi-square values), you can find the probability (or **p-value**) that matches the chi-square statistic calculated for your experiment. For example, a p-value of 0.05 means that in only one or fewer of 20 experiments will such a large difference be observed between observed and expected results, if the expected values were calculated accurately . As a general rule, geneticists use a p-value of 0.05 as a cut-off to determine whether deviations between observed and expected ratios can be attributed to sampling effects, or whether the underlying assumptions must be re-examined. Finally, it must be noted the chi-square distribution depends on the number of degrees of freedom, which is a statistical concept that in the context of genetics is usually one less than the number of phenotypic classes (n-1).

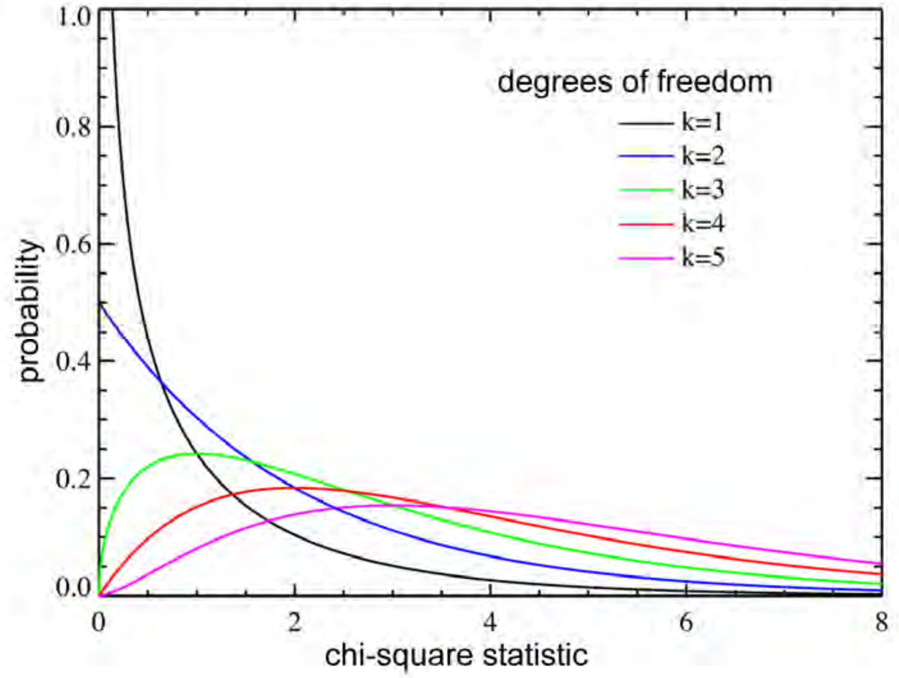


Figure 3.9 Probability distribution of the chi-square statistic for five different degrees of freedom

SUMMARY

- A diploid can have up to two different alleles at a single locus. The alleles are distributed equally between gametes during meiosis.
- Phenotype depends on the alleles that are present, their dominance relationships, and sometimes also interactions with the environment and other factors.
- The expected ratios of genotypes and phenotypes can be calculated for the progeny of any cross, if the mode of inheritance and the dominance relationships are known.
- Deviations between expected and observed phenotypic ratios may result from sampling effects, or from additional genetic or non-genetic factors not initially considered when calculating the expected ratios.
- The chi-square test is useful when deciding whether to investigate the underlying assumptions of the expected ratios.
- Sex-linkage is an exception to some definitions of Mendel's First Law, and can be best demonstrated through reciprocal crosses.

KEY TERMS

allele
Mendel's First Law
Law of Equal Segregation
homozygous
heterozygous
hemizygous
wild-type
locus
genotype
phenotype
dominant
recessive
semi-dominance
incomplete dominance
co-dominance
ABO
haplosufficiency
haploinsufficiency
true breeding lines
Punnett Square
monohybrid cross
test cross
tester
sex-linked
autosomal
reciprocal cross
G × E
penetrance
expressivity
sampling effects
chi-square, χ^2
p-value
degrees of freedom

QUESTIONS

3.1 What is the maximum number of alleles for a given locus in a normal gamete of a diploid species?

3.2 Wirey hair (*W*) is dominant to smooth hair (*w*) in dogs.

a) If you cross a homozygous, wirey-haired dog with a smooth-haired dog, what will be the genotype and phenotype of the F_1 generation?

b) If two dogs from the F_1 generation mated, what would be the most likely ratio of hair
phenotypes among their progeny?

c) When two wirey-haired *Ww* dogs actually mated, they had alitter of three puppies, which all had smooth hair. How do you explain this observation?

d) Someone left a wirey-haired dog on your doorstep. Without extracting DNA, what would be the easiest way to determine the genotype of this dog?

e) Based on the information provided in question 1, can you tell which, if either, of the alleles is wild-type?

3.3 An important part of Mendel's experiments was the use of homozygous lines as parents for his crosses. How did he know they were homozygous, and why was the use of the lines important?

3.4 In the table below, match the mouse hair color phenotypes with the term from the list that best explains the observed phenotype, given the genotypes shown. In this case, the allele symbols do not imply anything about the dominance relationships between the alleles. List of terms: haplosufficiency, haploinsufficiency, pleiotropy, semi-dominance, co-dominance, incomplete penetrance, variable expressivity.

3.5 Does equal segregation of alleles into daughter cells happen during mitosis, meiosis, or both?

3.6 If your blood type is B, what are the possible genotypes of your parents at the locus that controls ABO blood types?

3.7 A rare dominant mutation causes a neurological disease that appears late in life in all people that carry the mutation. If a father has this disease, what is the probability that his daughter will also have the disease?

3.8 The recessive *w* allele of the *white* gene in fruit flies produces white eyes.

a) If a heterozygous female and a white-eyed male are crossed, what are their phenotypic ratios among their F_1 progeny with respect to eye color?

b) What are the phenotypic ratios among their F_1 progeny with respect to both sex and eye color?

3.9 A particular mutant allele of the Hairy wing (*Hw*) gene is dominant and causes hairy wings, while the wild-type allele (Hw^+) is recessive.

a) How could you test whether this gene is sex-linked?

b)What would be the expected genotypic and phenotypic ratios in the F_2 generation?

3.10 Almost all of the examples of sex-linked inheritance refer to genes on the X chromosome. Imagine a gene with a non-sexual phenotype that was located exclusively on the Y-chromosome. What would be inheritance pattern of this gene?

3.11 Mendel's First Law (as stated in class) does not apply to alleles of most genes located on sex chromosomes. Does the law apply to the chromosomes themselves?

3.12 You calculate χ2 value of 2.7 for a particular experiment in which you expected to see
two phenotypes in a particular ratio. Use Table 3-1 (9th ed.) or Table 2-2 (8th ed.) to explain what this result means.

3.13 When can you have more confidence in your assumptions about an expected pattern of inheritance :

a) with a bigger χ^2 value or a smaller χ^2 value

b)with a bigger p-value or a smaller p-value?

c) with a bigger df or a smaller df?

3.14 Determine the appropriate degrees of freedom to use when scoring progeny from each of the following crosses.

a) *Aa* x *aa*, where *A* is dominant

b) *Aa* x *Aa*, where *A* is dominant

c) *Aa* x *Aa*, where *A* is semi-dominant

Table for Question 3.4

	A_1A_1	A_1A_2	A_2A_2
1	all hairs black	on the same individual: 50% of hairs are all black 50% of hairs are all white	all hairs white
2	all hairs black	all hairs are the same shade of grey	all hairs white
3	all hairs black	all hairs black	50% of individuals have all white hairs 50% of individuals have all black hairs
4	all hairs black	all hairs black	mice have no hair
5	all hairs black	all hairs white	all hairs white
6	all hairs black	all hairs black	all hairs white
7	all hairs black	all hairs black	hairs are a wide range of shades of grey

Chapter 4 MUTATION AND VARIATION

Figure 4.1 The difference in appearance between blue and white peacocks is due to mutation

The techniques of genetic analysis that we discussed in the previous chapters depend on the availability of two or more alleles for a gene of interest. Where do these alleles come from? The short answer is **mutation**. We have previously noted that an important property of DNA is its fidelity: most of the time it accurately passes the same information from one generation to the next. However, DNA sequences can also change. Changes in DNA sequences that we purposefully induce are called **mutations**. If a mutation changes the phenotype of an individual, the individual is said to be a **mutant**. Naturally occurring, but rare, sequence variants that are clearly different from a normal, wild-type sequence are also called mutations. On the other hand, many naturally occurring variants exist for traits for which no clearly normal type can be defined; thus, we use the term **polymorphism** to refer to variants of DNA sequences (and other phenotypes) that co-exist in a population at relatively high frequencies (>1%). Polymorphisms and mutations arise through similar biochemical processes, but the use of the word “polymorphism” avoids implying that any particular allele is more normal or abnormal. For example, a change in a person’s DNA sequence that leads to a disease such as cancer is appropriately called a mutation, but a difference in DNA sequence that explains whether a person has red hair rather than brown hair is an example of polymorphism. Molecular markers, which we will discuss in Chapter 9, are a particularly useful type of polymorphism for some areas of genetics research.

Humans have an interesting relationship with mutation. From our perspective, mutations can be extraordinarily useful, since mutations allow evolution to proceed. Mutation is also the basis for the domestication and improvement of almost all of our food. On the other hand, as the cause of many cancers and other diseases, mutation can be devastating to individuals. Yet, the vast majority of mutations probably go undetected. In this section, we will examine some of the causes and effects of mutations.

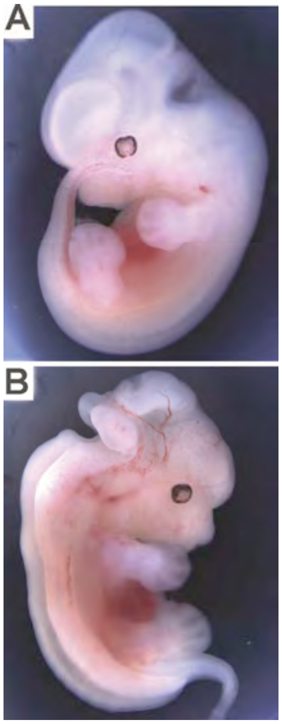


Figure 4.2 Examples of wild-type (A) and mutant (B) mouse embryos observed while screening for genes affecting cranium development.

ORIGINS OF MUTATIONS

Mutation may involve the **insertion** or **deletion** of one or more bases, or else the **substitution** of one DNA sequence with another DNA sequence of equal length. These changes in DNA sequence can arise in many ways, some of which arise spontaneously from natural sources and some of which are induced intentionally or unintentionally by human actions. There are many ways to classify **mutagens**, which are the agents that cause mutation. We will classify mutagens here as being either biological, chemical, or physical.

BIOLOGICAL

A major source of spontaneous mutation is errors that arise during DNA replication. DNA polymerases are usually very accurate in adding a base to the growing strand that is the exact complement of the base on the template strand. However, occasionally, an incorrect base is inserted. In some cases, this is because rare and minor shifts in the chemical structure of nucleotides (called **tautomerism**) allow for unusual base pairs to form, such as G-T and C-A (Figure 4.3). Usually, the machinery of DNA replication will recognize and repair mispaired bases, but nevertheless, some errors become permanently incorporated in a daughter strand, and so become mutations that will be inherited by the cell's descendents (Figure 4.4).

Another source of error introduced during replication is caused by a temporary misalignment of a few bases between the template strand and daughter strand (Figure 4.5). This causes one or more bases on either strand to be temporarily displaced in a **loop** that is not paired with the opposite strand. If this loop forms on the template strand, the bases in the loop may not be replicated, and a deletion will be introduced in the growing daughter strand. Conversely, if a region of the daughter strand that has just been replicated becomes displaced, this region may be replicated again, leading to an insertion of additional sequence in the daughter strand, as compared to the template strand. Regions of DNA that have several repeats of the same few nucleotides in a row are especially prone to this type of error during replication. These short-sequence repeats (**SSRs**) are also called microsatellites and tend to be highly polymorphic, and are therefore particularly useful in genetics.

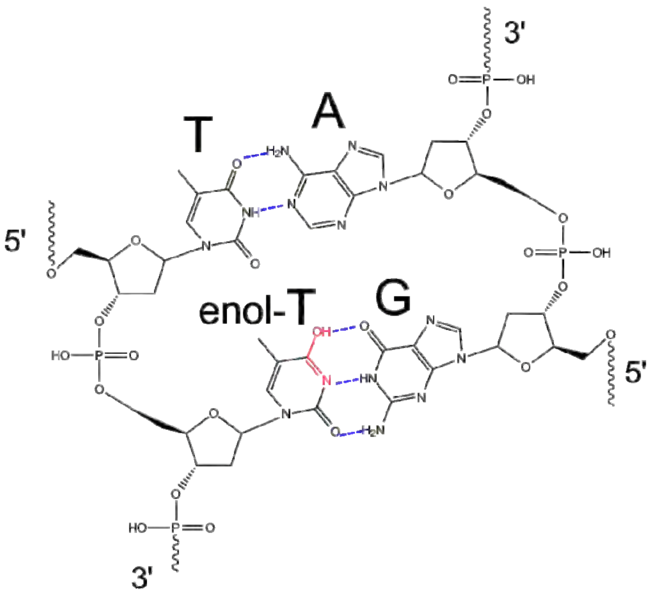


Figure 4.3 A rare, enol tautomer of thymine (enol-T) can base pair with guanine (G) rather than adenine (A). The enol tautomer can form spontaneously and is also stabilized by some types of mutagens.

Mutations can also be introduced by viruses, transposable elements, and other types of DNA that are naturally inserted at more or less random positions in chromosomes. The insertion may disrupt the coding sequence of a gene, or have other consequences including the fusion of part of one gene with another. These insertions occur spontaneously, and can also be intentionally stimulated in the laboratory as a method of mutagenesis. For example, a type of transposable element called a **P-element** is widely used as a biological mutagen in Drosophila, and **T-DNA**, which is an insertional element modified from a bacterial pathogen is used as a mutagen in some plant species.

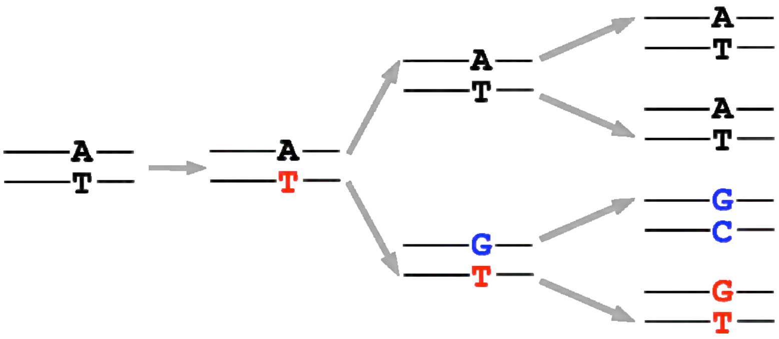


Figure 4.4 Mispairing of bases (e.g. G with T) can occur due to tautomerism, alkylating agents, or other effects. As a result, in this example the AT base pair in the original DNA strand will become permanently substituted by a GC based pair in some progeny. The mispaired GT basepair will likely be repaired or eliminated before further rounds of replication.

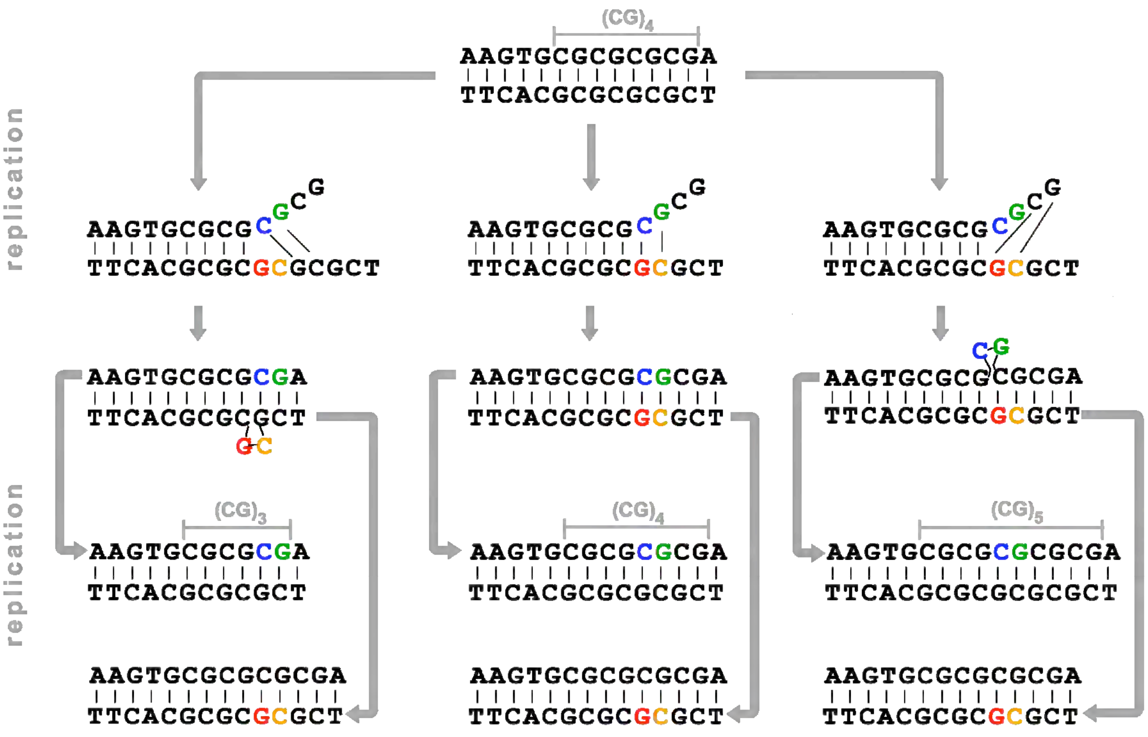


Figure 4.5 Improperly aligned basepairing during replication can occur occasionally, especially in regions with short, repeated sequences. This can lead to either deletion (left) or insertion (right) of sequences compared to the products of normal replication (center), depending on whether the template strand or daughter strand "loops-out" during replication.

CHEMICAL

Many chemical compounds, whether natural or synthetic, can react with DNA. These reactions may change the chemical structure of particular bases, so that they are misread during replication. Other chemical mutagens distort the double helix causing it to be replicated inaccurately, while still others may cause breaks in chromosomes that lead to deletions and other types of aberrations. Following are examples of two classes of chemical mutagens, **alkylating agents**, and **intercalating agents**, that are important in genetics or medicine.

Ethane methyl sulfonate (**EMS**), is an example of an alkylating agent that is commonly used by geneticists to induce mutations in a wide range of both prokaryotes and eukaryotes. The organism is fed or otherwise exposed to a solution of EMS, which reacts with some of the G bases it encounters in a process called alkylation. The addition of an alkyl group to G changes the base pairing properties so that the next time the alkylated DNA strand is replicated, a T instead of a C will be inserted opposite to the alkylated-G in the daughter strand (Figure 4.6).

Box 4-1 Transposable Elements

Transposable elements (TEs) occur naturally throughout the chromosomes of almost all organisms. These DNA sequences have a unique ability to be inserted into new locations in the genome, after being cut or copied from their original location. TEs are also known as mobile genetic elements, or more informally as jumping genes. The locations into which TEs are inserted are not entirely random, but TEs can in principle be inserted into almost any region of the genome. TEs can therefore move into other genes, causing mutations. Researchers have methods of artificially increasing the rate of transposition, making TEs a useful type of mutagen. However the biological importance of TEs extends far beyond their use in mutant screening; TEs are also important causes of disease and phenotypic instability, and they are a major force in evolution.

There are two major classes of TEs in eukaryotes (Figure 4-B1). **Class I** elements include **retroposons** and **retrotransposons**; these are copied by means of an RNA intermediate. The transcript is reverse transcribed into DNA before being inserted elsewhere in the genome through the action of enzymes such as **integrase**. **Class II** elements are known also as **transposons**; these do not use reverse transcriptase or an RNA intermediate for transposition. Using an enzyme called **transposase**, most transposons are cut from their original location and then this excised dsDNA fragment is inserted into a new location. Note that the name transposon is sometimes used incorrectly to refer to any type of TEs, but in this book we use transposon to refer only to Class II elements.

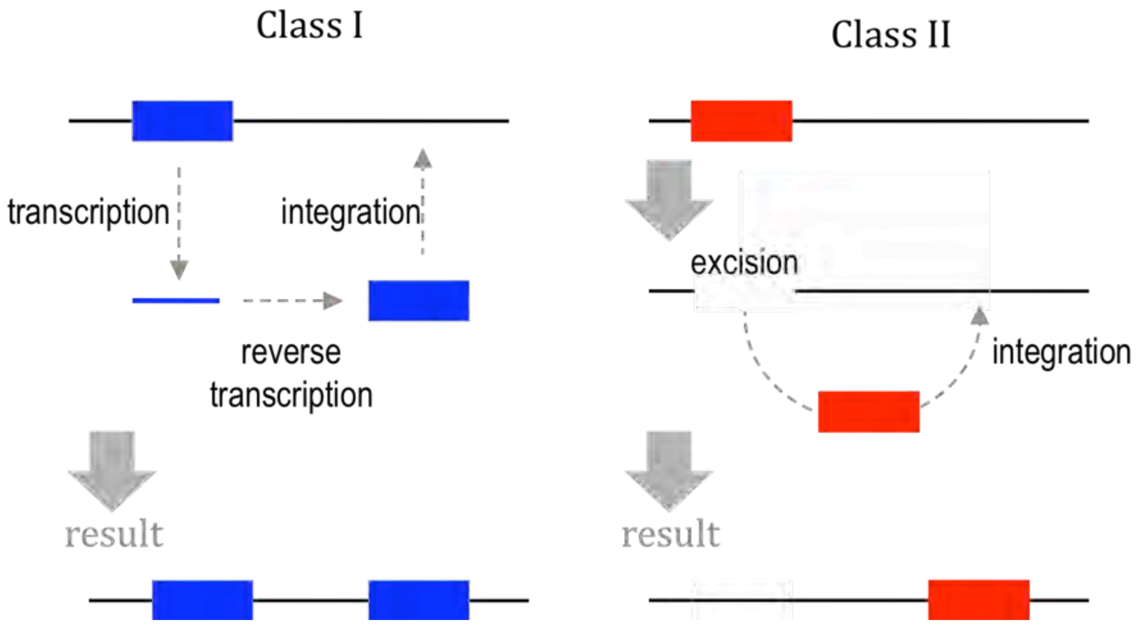


Figure 4-B1 . Representative examples of the two main types of transposable elements. (TEs) Class I elements transpose via an ssRNA intermediate, which is reverse transcribed to dsDNA prior to insertion of this copy in a new site in the genome. Class II elements do not involve an RNA intermediate; most Class II elements are cut from their original location as dsDNA, prior to being inserted into a new site in the genome. Although the diagram shows TEs being inserted on the same chromosome as they originated from, TEs can also move to other chromosomes within the same cell.

TEs are relatively short DNA sequences (between 100bp and 10kb), and encode no more than a few proteins (if any). Normally, the protein-coding genes within a TE are all related to the TE's own transposition functions, e.g. **reverse transcriptase**, transposase, and integrase. However, some TEs (of either Class I or II) do not encode any proteins at all. These **non-autonomous** TEs can only transpose if they are supplied with enzymes produced by other, **autonomous** TEs located elsewhere in the genome. In all cases, enzymes for transposition recognize conserved nucleotide sequences within the TE, which show the enzymes where to begin cutting or copying, and re-insertion.

The human genome is nearly 45% TEs, the vast majority of which are families of Class I elements called **LINEs** and **SINEs**. The short, ***Alu*** type of SINE occurs in more than one million copies in the human genome (compare this to the approximately 30,000, non-TE, protein-coding genes in humans). Indeed, TEs make up a significant portion of the genomes of almost all eukaryotes. Class I elements, which usually transpose via a copy-and-paste mechanism, tend to be more abundant than Class II elements, which mostly use a cut-and-paste mechanism. But even the cut-paste mechanism can lead to an increase in TE copy number, in some circumstances (for example, if the site vacated by the excised transposon is repaired with a DNA template from a homologous chromosome that itself contains a copy of a transposon).

Besides greatly expanding the DNA content of genomes, TEs contribute to genome evolution in many other ways. As already mentioned, they may disrupt gene function by insertion into a gene's coding region or regulatory region. More interestingly adjacent regions of chromosomal DNA are sometimes mistakenly transposed along with the TE; this can lead to gene duplication. The duplicated genes are then free to evolve independently, leading in some cases to the development of new functions. The breakage of strands by TE excision and integration can disrupt genes, and can lead to chromosome rearrangement or deletion if errors are made during strand rejoining. Furthermore, having so many similar TE sequences distributed throughout a chromosome sometimes allows mispairing of regions of homologous chromosomes at meiosis, which can cause unequal crossing-over, resulting in deletion or duplication of large segments of chromosomes. Thus, TEs are an important evolutionary force, and are not merely "junk DNA" as they were once called.

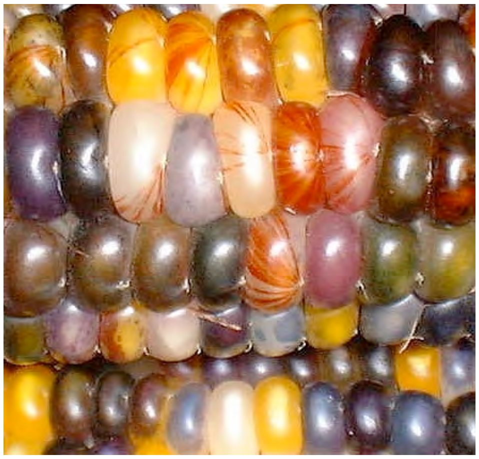


Figure 4-B2. Barbara McClintock won a Nobel Prize for her discovery of TEs. She did so by studying pigment variegation in maize kernels, which is caused by the movement of TEs in and out of pigmentation genes. This is an example of phenotypic instability.

The new strand therefore bears a mutation, which will be inherited, in all the strands that are subsequently replicated from it.

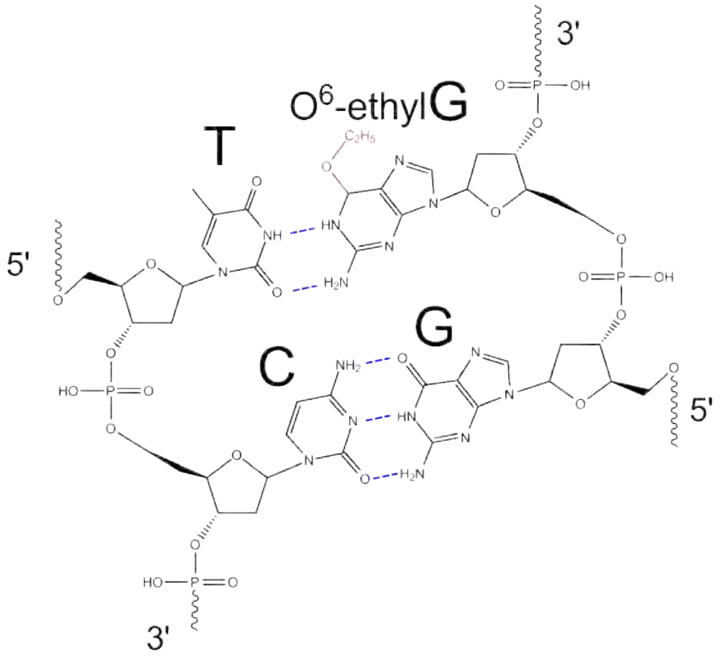


Figure 4.6 Alkylation of guanine (shown in red) allows G to bond with thymine rather than cytosine at replication.

Intercalating agents are another type of chemical mutagen. These induce mutations by inserting between the stacked bases at the center of the DNA helix (Figure 4.7). This intercalation distorts the shape of the DNA helix, which can cause the wrong bases to be added to a growing DNA strand during DNA synthesis. Intercalating agents tend to be flat, planar molecules such as **benzopyrene**, a component of wood and tobacco smoke. Another important intercalating agent is thalidomide, an anti-nausea drug whose harmful effects were unknown until its consumption by thousands of pregnant woman resulted in birth defects. Finally **ethidium bromide**, the dye that fluorescently stains DNA in laboratory assays, is also an intercalating agent. For this reason, molecular biologists are trained to handle this chemical carefully.

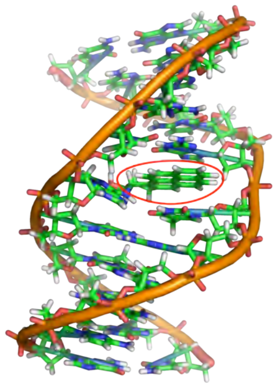


Figure 4.7 Benzopyrene (circled in red) is an example of an intercalating agent

PHYSICAL

Anything that damages DNA by transferring energy to it can be considered a physical mutagen. Usually this involves radioactive particles, x-rays, or UV light. Smaller, fast moving particles may substitute or delete a single base, while larger, slightly slower particles induce larger deletions by breaking the double stranded helix. Physical mutagens can also create unusual structures in DNA, such as the thymine dimers formed by UV light (Figure 4.8). Thymine dimers disrupt normal base-pairing in the double helix, and may block replication altogether if not repaired by the cell.

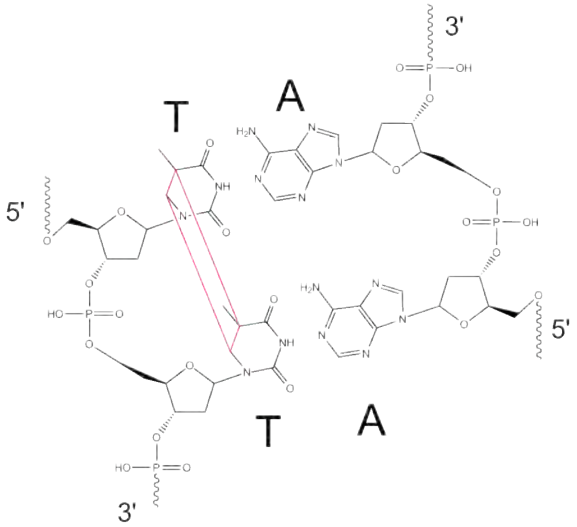


Figure 4.8 Thymine dimers are formed when adjacent thymine bases on the same DNA strand become covalently linked (red bonds) follow exposure to mutagens such as UV light. The dimers distort base pairing and can interrupt processes such as replication.

Mutant screening: Forward genetics

One way to identify genes that affect a particular biological process is to induce random mutations in a population, and then look for individuals with phenotypes that might be caused by a disruption of a particular biochemical pathway. This strategy of **mutant screening** has been used very effectively to better understand the molecular components of hundreds of different biological processes. For example, to better understand processes of memory and learning, researchers have screened mutagenized populations of *Drosophila* to identify flies (or larvae) that lack the normal ability to learn to associate a particular odor with an electric shock. Some of the genes identified by this mutant screen may be relevant to learning and memory in other animals, including conditions such as Alzheimer's disease in humans.

Exposure of an organism to a mutagen causes mutations in essentially random positions along the chromosomes. Most of the mutant phenotypes recovered from a genetic screen are caused by **loss-of-function** mutations. These are changes in the sequence of an allele that cause it to no longer produce the same level of active protein as the wild-type allele. Loss-of-function alleles tend to be recessive because the wild-type allele is haplosufficient (see Chapter 6). A loss-of-function allele that produces no active protein is called an **amorph**, or **null**. On the other hand, alleles with only a partial loss-of-function are called hypomorphic. More rarely, a mutant allele may have a **gain-of-function**, producing either more of the active protein (**hypermorph**) or producing an active protein with a new function (**neomorph**).

In a typical mutant screen, researchers will treat a parental population with a mutagen. This may, for example, involve soaking seeds in EMS, or mixing this mutagen with the food fed to flies. The individuals that are directly exposed with the mutagen are called the M_0 generation. No phenotypes will be visible among the M_0 generation, in part because not all somatic (i.e. non-reproductive) cells of the M_1 individuals will be affected in the same way. More important in the M_1 generation are the germline cells: gametes and any of their developmental precursors. Since a single gamete contributes half of the chromosomes to every cell of the next (M_2) generation, mutagenizing gametes or the cells that produce them is the easiest way to generate individuals in which every cell bears the same mutation. In most cases, however, the M_1 individual will be heterozygous for any induced mutation; this is because it would be very unlikely for the same gene to have been mutated in the other gamete that contributed to the M_1 individual. Because most induced mutations are recessive, the M_1 generation must therefore be selfed (i.e. crossed to siblings, or self-fertilization if the species is a hermaphrodite like worms and most plants), and the following generations (e.g. M_2, in which mutant alleles could potentially become homozygous) examined for the presence of novel phenotypes. Once a relevant mutant has been identified, geneticists can begin to make inferences about what the normal function of the mutated gene is, based on its mutant phenotype.

SOME MUTATIONS MAY NOT HAVE DETECTABLE PHENOTYPES

The vast majority of mutations (especially substitutions) have no effect on the phenotype. Often, this is because the mutation is **silent**; it changes the DNA sequence of a non-coding region of the DNA, or else the changes a base within a codon without changing the amino acid that it encodes (recall that the genetic code is degenerate; for example, GCT, GCC, GCA, and GCG all encode alanine). There are also cases where a mutation can cause a complete loss-of-function of a gene, yet not produce a phenotype even when the mutant allele is homozygous. This can often be attributed to genetic **redundancy**, i.e. the encoding of similar genes at more than one locus in the genome. It is important therefore to remember this important limitation of mutational analysis: genes with redundant functions cannot be easily identified by mutant screening.

Some phenotypes require individuals to reach a particular developmental stage before they can be scored. For example, flower color can only be scored in plants that are mature enough to make flowers, and eye color can only be scored in flies that have developed eyes. However, some alleles may not develop sufficiently to be included among the progeny that are scored according to their phenotype. Lethal alleles that arrest the development of an individual at an embryonic stage may therefore go unnoticed in a typical mutant screen. Furthermore, the progeny of a monohybrid cross involving an embryonic lethal recessive allele may therefore all be of a single phenotypic class, giving a phenotypic ratio of 1:0 (which is the same as 3:0).

Many genes are first identified in mutant screens, and so they tend to be named after their mutant phenotypes. This can cause some confusion for students of genetics. For example, we have already encountered an X-linked gene named *white* in fruit flies. Null mutants of white have white eyes, but the normal function of the white gene is actually the production of a red pigment.

COMPLEMENTATION TESTING

Mutations in different genes can produce the same phenotype. For example, in the biochemical pathway shown in Figure 4.9, a plant that lacks the function of gene A (genotype *aa*) would produce mutant, white flowers that looked just like the flowers of a plant that lacked the function of gene B (genotype *bb*). The genetics of two loci are discussed more in the following chapters.

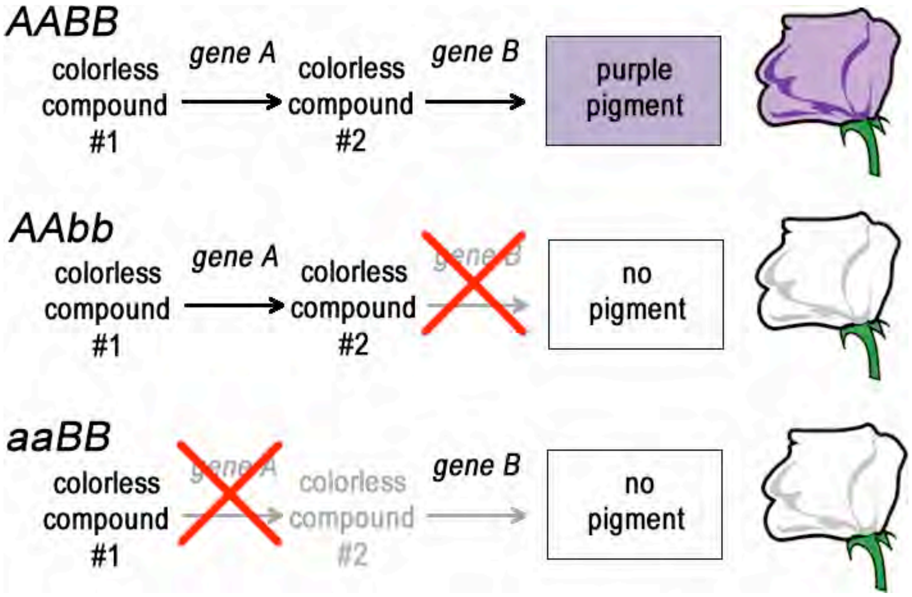


Figure 4.9 In this simplified biochemical pathway, two enzymes encoded by two different genes modify chemical compounds in two sequential reactions to produce a purple pigment. Loss of either of the enzymes disrupts the pathway and no pigment is produced.

As explained earlier in this chapter, mutant screening is one of the main activities of geneticists. When geneticists find two mutants with similar phenotypes, either in natural populations or during a mutant screen, an immediate question is whether or not the mutants have lost the function of the same gene. The other possibility is that each mutant has lost the function of a different gene. If they are both mutants of the same gene, then their genotypes can both be represented as *aa*. If each has a mutation in a different gene, then the genotype of one individual can be represented as *aa*, and the other individual as *bb* (or more completely as *aaBB* and *AAbb*, respectively).

A technique called **complementation testing** can be used to determine whether two individuals with the same phenotype carry mutations in the same gene or different genes. The only requirement of this test is two pure-breeding individuals with the same phenotype; no prior knowledge of the genes or biochemical pathways is required. To perform a complementation test, two homozygous individuals with similar mutant phenotypes are crossed (Figure 4.10). If the F_1 progeny all have the mutant phenotype, then we infer that same gene is mutated in each parents. If the F_1 progeny all appear to be wild-type, then each of the parents most likely carries a mutation in a different gene (Figure 4.11).

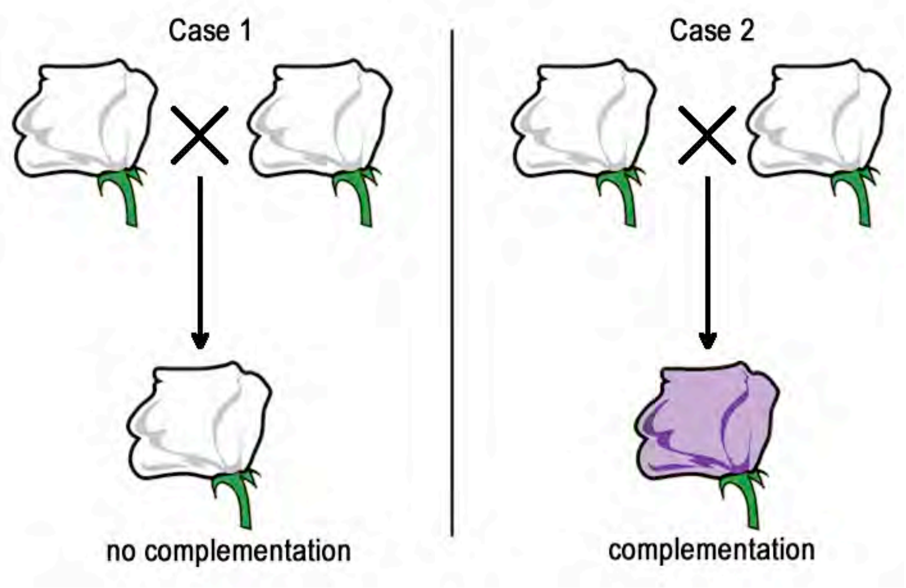
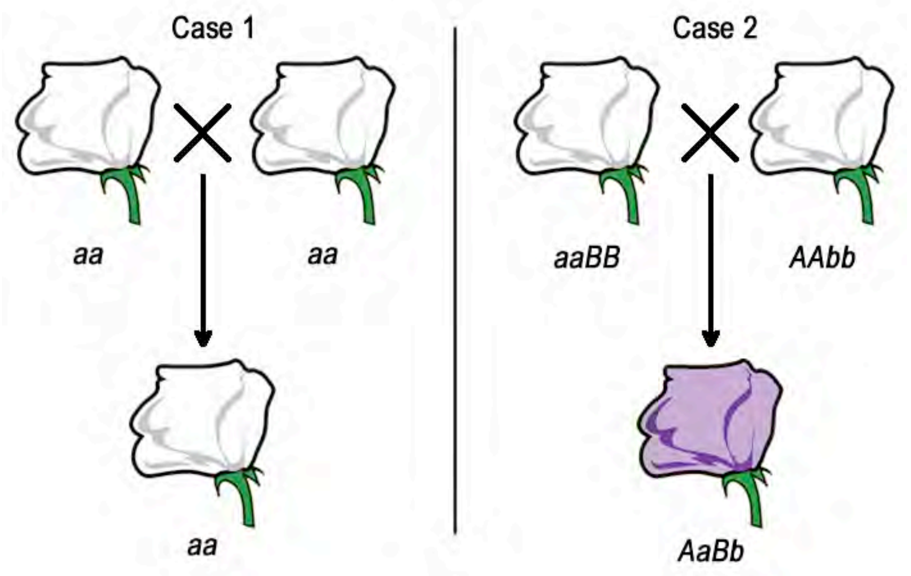


Figure 4.10

In a typical complementation test, the genotypes of two parents are unknown (although they must be pure breeding, homozygous mutants). If the F1 progeny all have a mutant phenotype there (Case 1), there is no complementation. If the F1 progeny are all wild-type, the mutations are said to have complemented each other. See Figure 4.11 for interpretation.

Figure 4.11

The pure breeding, homozygous mutants parents had unknown genotypes before the complementation test, but it could be assumed that they were either mutations in the same genes (Case 1) or different genes (Case 2). In Case 1, all of the progeny would have a mutant phenotype, because they would all have the same, homozygous genotype as the parents. In Case 2, each parent has a mutation in a different gene, therefore none of the F_1 progeny will be homozygous mutant at any one locus. Note that the genotype in Case 1 could be written as either *aa* or *aaBB*.



Thus, if two homozygous mutants produce F_1 progeny that have a wild-type phenotype, **complementation** is said to have occurred, because the wild-type alleles of each of the genes were able to compensate for the recessive, mutant alleles that were also inherited (Case 2, Figure 4.11). On the other hand, if the F_1 progeny all have a mutant phenotype, (case 1, Figure 4.11), the mutant genotypes fail to complement each other, because they contain mutations of the same gene. These could be either the same mutant alleles, or different mutant alleles of the same gene. If the two mutants are independent (e.g. they came from different natural populations or from independently mutagenized individuals), the mutations are probably different alleles of the same gene. All mutants that fail to complement each other are said to be in the same **complementation group**.

SUMMARY

- When a variation in DNA sequence originated recently, and is rare in a population, we call that change a mutation.

- When variations in DNA sequence co-exist in a population, and neither one can be meaningfully defined as wild-type, we call the variations polymorphisms.

- Mutations may either occur spontaneously, or may be induced by exposure to mutagens.

- Mutations may result in either substitutions, deletions, or insertions.

- Mutation usually causes either a partial or complete loss of function, but sometimes results in a gain of function, including new functions.

- Spontaneous mutations arise from many sources including natural errors in DNA replication, usually associated with base mispairing, or else insertion deletion especially within repetitive sequences.

- Induced mutations result from mispairing, DNA damage, or sequence interruptions caused by chemical, biological, or physical mutagens.

- By randomly inducing mutations, then screening for a specific phenotype, it is possible to identify genes associated with specific biological pathways.

- Transposable elements are dynamic, abundant components of eukaryotic genomes and important forces in evolution.

- Transposable elements are dynamic, abundant components of eukaryotic genomes and important forces in evolution.

- Mutation of different genes can produce a similar phenotype.

- Complementation testing determines whether two mutants are the result of mutation of the same gene, or if each mutant is caused by mutation of a different gene.

- The efficiency of mutant screening is limited by silent mutations, redundancy, and embyronic lethality.

KEY TERMS

mutation
mutant
polymorphism
insertion
deletion
substitution
mutagen
biological mutagen
chemical mutagen
physical mutagen
tautomer
mispairing
loop
SSR
insertional mutagen
Class I, Class II
transposon
retrotransposon
reverse transcriptase
transposase
non-autonomous
autonomous
SINE, LINE, Alu
P-element
T-DNA
alkylation agent
EMS
intercalating agent
benzopyrene
ethidium bromide
thymine dimer
mutant screen
loss-of-function
gain-of-function
amorph
null
hypomorph
hypermorph
neopmorph
somatic
germline
M_0, M_1, M_2
silent mutation
redundancy
lethality
complementation group

STUDY QUESTIONS

4.1 How are polymorphisms and mutations alike? How are they different?

4.2 What are all of the ways a substitution can occur in a DNA sequence?

4.3 What are all of the ways a deletion can occur in a DNA sequence?

4.4 What are all of the ways an insertion can occur in a DNA sequence?

4.5 In the context of this chapter, explain the health hazards of smoking tobacco.

4.6 You have exposed a female fruit fly to a mutagen. Mating this fly with a non-mutagenized male produces offspring that appear to be completely normal. However there are twice as many females as males in the F_1 progeny of this cross.
a) Propose a hypothesis to explain these observations.
b) How could you test your hypothesis?

4.7 You decide to use genetics to investigate how your favourite plant makes its flowers smell good.
a) What steps will you take to identify some genes that are required for production of the sweet floral scent? Assume that this plant is a self-pollinating diploid.
b) One of the recessive mutants you identified has fishy-smelling flowers, so you name the mutant (and the mutated gene) *fishy*. What do you hypothesize about the normal function of the wild-type *fishy* gene?

c) Another recessive mutant lacks floral scent altogether, so you call it *nosmell*. What could you hypothesize about the normal function of this gene?

4.8 Suppose you are only interested in finding dominant mutations that affect floral scent.
a) What do you expect to be the relative frequency of dominant mutations, as compared to recessive mutations, and why?
b) How will you design your screen differently than in the previous question, in order to detect dominant mutations specifically?
c) Which kind of mutagen is most likely to produce dominant mutations, a mutagen that produces point mutations, or a mutagen that produces large deletions?

4.9 Which types of transposable elements are transcribed?

4.10 You are interested in finding genes involved in synthesis of proline (Pro), an amino acid that is normally synthesizes by a particular model organism.
a) How would you design a mutant screen to identify genes required for Pro synthesis?
b) Imagine that your screen identified ten mutants (#1 through #4) that grew poorly unless supplemented with Pro. How could you determine the number of different genes represented by these mutants?
c) If each of the four mutants represents a different gene, what will be the phenotype of the F1 progeny if any pair of the four mutants are crossed?
d) If each of the four mutants represents the same gene, what will be the phenotype of the F1 progeny if any pair of the four mutants are crossed?

Chapter 5 PEDIGREES AND POPULATIONS

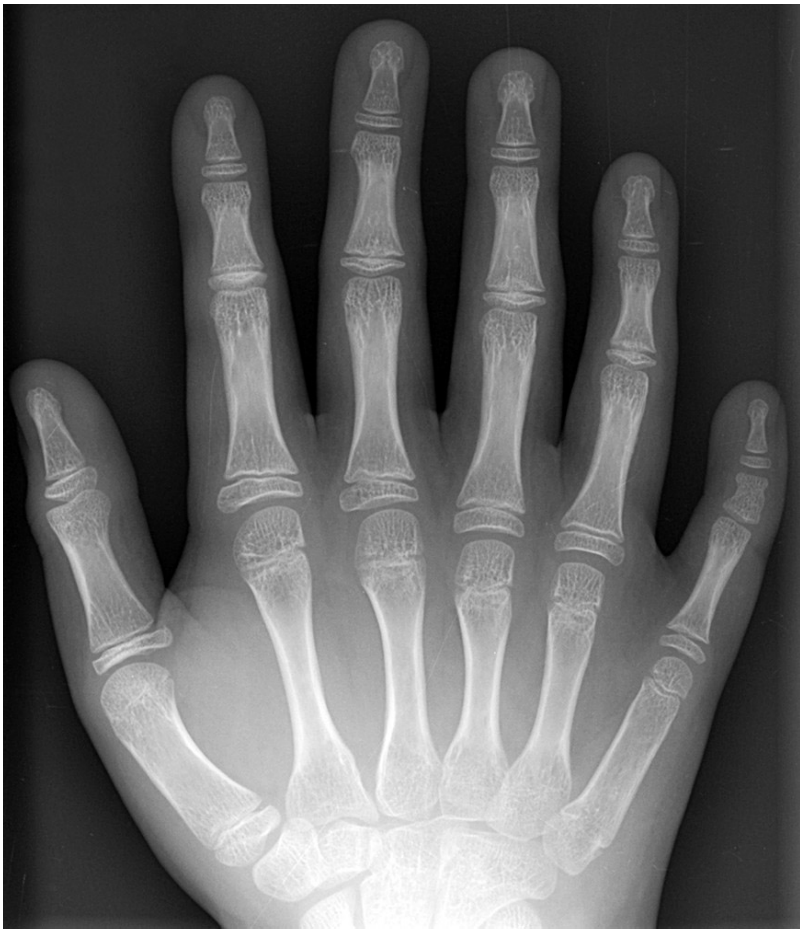


Figure 5.1 Polydactyly is an example of a human trait that can be studied by pedigree analysis

The basic concepts of genetics described in the preceding chapters can be applied to almost any eukaryotic organism. However, some techniques, such as test crosses, can only be performed with model organisms or other species that can be experimentally manipulated. To study the inheritance patterns of genes in humans and other species for which controlled matings are not possible, geneticists use techniques including the analysis of pedigrees and populations

PEDIGREE ANALYSIS

Pedigrees use a standardized set of symbols to represent an individual's sex, family relationships and phenotype. These diagrams are used to determine the **mode of inheritance** of a particular disease or trait, and to predict the probability of its appearance among offspring. Pedigree analysis is therefore an important tool in both basic research and **genetic counseling**.

Each pedigree represents all of the available information about the inheritance of a single trait (most often a disease) within a family. The pedigree is therefore drawn from factual information rather than theoretical predictions, but there is always some possibility of errors in this information, especially when relying on family members' recollections or even clinical diagnoses. In real pedigrees, further complications can arise due to incomplete penetrance (including age of onset) and variable expressivity of disease alleles, but for the examples presented in this book, we will presume complete accuracy of the pedigrees. A pedigree may be drawn when trying to determine the nature of a newly discovered disease, or when an individual with a family history of a disease wants to know the probability of passing the disease on to their children. In either case, a tree is drawn, as shown in Figure 5.2, with circles to represent females, and squares to represent males. If an individual is known to have symptoms of the disease, the symbol is filled in. Sometimes a half-filled in symbol is used to indicate a **carrier** of a disease; this is someone who does not have any symptoms of the disease, but who passed the disease on to subsequent generations. Note that when a pedigree is constructed, it is often unknown whether a particular individual is a carrier or not, so not all carriers are always explicitly indicated in a pedigree. For simplicity, we will assume that the pedigrees presented in this chapter are accurate, and represent fully penetrant traits.

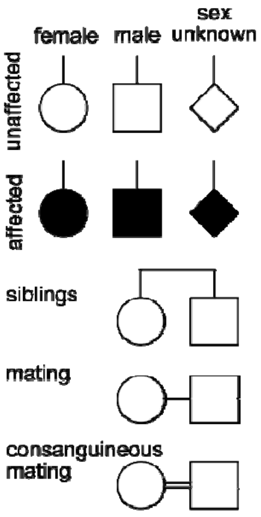


Figure 5.2 Symbols used in drawing a pedigree.

INFERRING THE MODE OF INHERITANCE

Given a pedigree of an uncharacterized disease or trait, one of the first tasks is to determine the mode of inheritance, as this information is essential in calculating the probability that the trait will be inherited in offspring. We will consider four major types of inheritance: autosomal dominant (AD), autosomal recessive (AR), X-linked dominant (XD), X-linked recessive (XR).

AUTOSOMAL DOMINANT (AD)

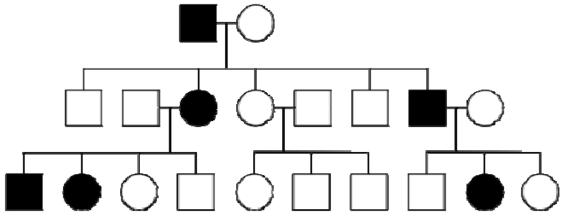


Figure 5.3 A pedigree consistent with AD inheritance.

When a disease is caused by a dominant allele of a gene, every person with that allele will show symptoms of the disease (assuming complete penetrance), and only one disease allele needs to be inherited for an individual to be affected. Thus, every affected individual must have an affected parent. A pedigree with affected individuals in every generation is typical of AD diseases. However, beware that other modes of inheritance can also show the disease in every generation, as described below. It is also possible for an affected individual with an

AD disease to have a family without any affected children, if the affected parent is a heterozygote. This is particularly true in small families, where the probability of every child inheriting the normal, rather than disease allele is not extremely small. Note that AD diseases are usually rare within populations, therefore affected individuals with AD diseases tend to be heterozygotes (otherwise, both parents would have had to been affected with the same rare disease). Achondroplastic dwarfism, and polydactyly are both examples of human conditions that may follow an AD mode of inheritance.

X-LINKED DOMINANT (XD)

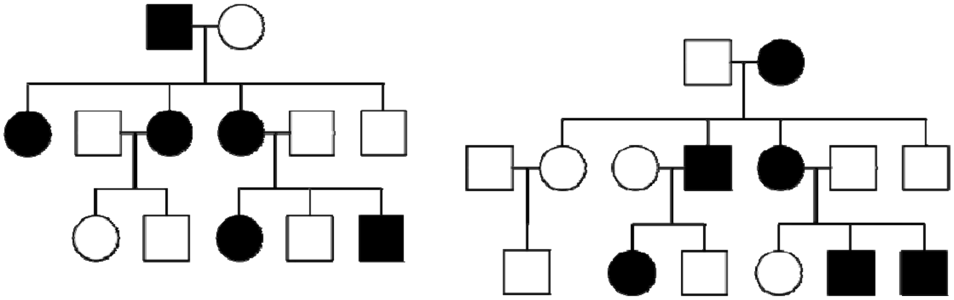


Figure 5.4 Two pedigrees consistent with XD inheritance.

In X-linked dominant inheritance, the gene responsible for the disease is located on the X-chromosome, and the allele that causes the disease is dominant to the normal allele. Because females have twice as many X-chromosomes as males, females tend to be affected more frequently than males in XD pedigrees. However, not all pedigrees provide sufficient information to distinguish XD and AD. One definitive indication that a trait is inherited as AD rather than XD is that an affected father passes the disease to a son; this type of transmission is <u>not</u> possible with XD, since males inherit their X chromosome from their mothers.

Figure 5.5 Some types of rickets may follow an XD mode of inheritance.

AUTOSOMAL RECESSIVE (AR)

Diseases that are inherited in an autosomal recessive pattern require that both parents of an affected individual carry at least one copy of the disease allele. With AR traits, many individuals in a pedigree can be carriers, probably without knowing it. Compared to pedigrees of dominant traits, AR pedigrees tend to show fewer affected individuals and are more likely than AD or XD to "skip a generation". Thus, the major feature that distinguishes AR from AD or XD is that unaffected individuals can have affected offspring.

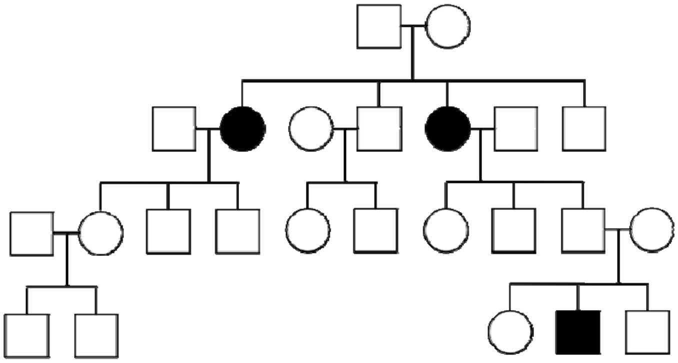


Figure 5.6 A pedigree consistent with AR inheritance.

Figure 5.7 Many inborn errors of metabolism, such as phenylketonuria (PKU) are inherited as AR. Newborns are often tested for a few of the most common metabolic diseases.

X-LINKED RECESSIVE (XR)

Because males have only one X-chromosome, any male that inherits an X-linked recessive disease allele will be affected by it (assuming complete penetrance). Therefore, in XR modes of inheritance, males tend to be affected more frequently than females. This is in contrast to AR and AD, where both sexes tend to be affected equally, and XD, in which females are affected more frequently. Note, however, that in the small sample sizes typical of human families, it may not be possible to accurately determine whether one sex is affected more frequently than others. On the other hand, one feature of a pedigree that can be used to definitively establish that an inheritance pattern is not XR is the presence of an affected daughter from unaffected parents; because she would have had to inherit one X-chromosome from her father, he would also have been affected in XR.

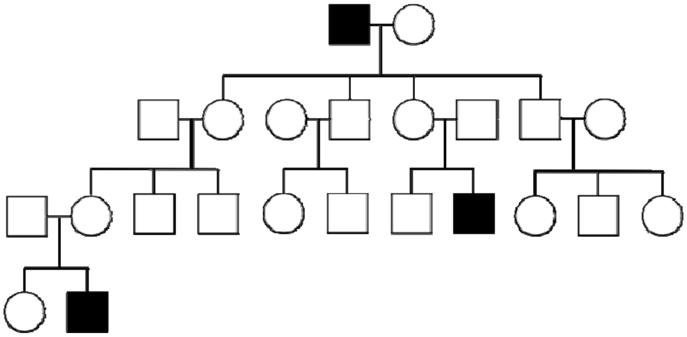


Figure 5.8 A pedigree consistent with XR inheritance.

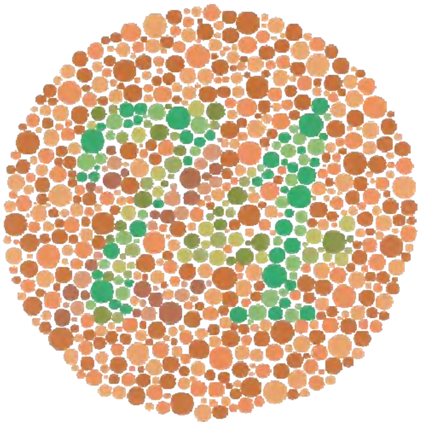


Figure 5.9 Some forms of colour blindness are inherited as XR-traits. Colour blindness is diagnosed using tests such as this Ishihara Test.

SPORADIC AND NON-HERITABLE DISEASES

Only a fraction of all of the known human traits and diseases have been proven to be caused by alleles of a single gene. Many other diseases have a heritable component, but have more complex inheritance patterns due to the involvement of multiple genes and environmental factors. On the other hand, some diseases may appear to be heritable, because they affect multiple members of the same family, but this could also be because the family members are exposed to the same toxins or other environmental factors in their homes. Finally, diseases with similar symptoms may have different causes, so of which may be genetic while others are not. One example of this is ALS (amyotrophic lateral sclerosis); approximately 5-10% of cases are inherited in an AD pattern, while the majority of the remaining cases appear to be **sporadic**, in other words, not caused by a mutation inherited from a parent. We now know that different genes or proteins are affected in the inherited and sporadic forms of ALS. The physicist Stephen Hawking and baseball player Lou Gehrig both suffered from sporadic ALS.

Figure 5.10
Stephen Hawking

CALCULATING PROBABILITIES

Once the mode of inheritance of a disease or trait is known, some inferences about the genotype of individuals in a pedigree can be made, based on their phenotypes. Given these genotypes, it is possible to calculate the probability of a particular genotype being inherited in subsequent generations. This can be useful in genetic counseling, for example when prospective parents wish to know the likelihood of their offspring inheriting a disease for which they have a family history.

Probabilities in pedigrees are calculated using the same basic methods as are used in other fields. The first formula is the **product rule**: the joint probability of two independent events is the product of their individual probabilities; this is the probability of one event AND another event occurring. For example, the probability of a rolling a "five" with a single throw of a single six-sided die is 1/6, and the probability of rolling "five" in each of three successive rolls is 1/6 x 1/6 x 1/6 = 1/216. The second useful formula is the **sum rule**, which states that the combined probability of two independent events is the sum of their individual probabilities. This is the probability of one event OR another event occurring. For example, the probability of rolling a five or six in a single throw of a dice is 1/6 + 1/6 = 1/3.

With these rules in mind, we can calculate the probability that two carriers (i.e. heterozygotes) of an AR disease will have a child affected with the disease as ½ x ½ = ¼, since for each parent, the probability of any gametes carrying the disease allele is ½. This is consistent with what we already know from calculating probabilities using a Punnett Square (e.g. in a monohybrid cross *Aa* x *Aa*, ¼ of the offspring are *aa*).

We can likewise calculate probabilities in the more complex pedigree shown in Figure 5.11.

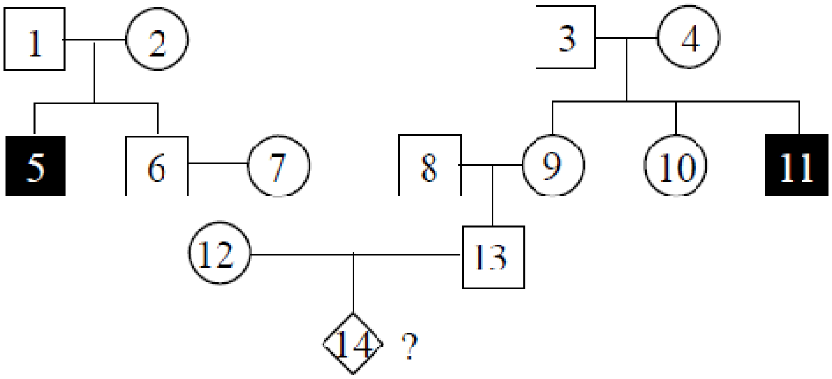


Figure 5.11

Individuals in this pedigree are labeled with numbers to make discussion easier

Assuming the disease has an AR pattern of inheritance, what is the probability that individual 14 will be affected? We can assume that individuals #1, #2, #3 and #4 are heterozygotes (*Aa*), because they each had at least one affected (*aa*) child, but they are not affected themselves. This means that there is a 2/3 chance that individual #6 is also *Aa*. This is because according to Mendelian inheritance, when two heterozygotes mate, there is a 1:2:1 distribution of genotypes *AA*:*Aa*:*aa*. However, because #6 is unaffected, he can't be *aa*, so he is either *Aa* or *AA*, but the probability of him being *Aa* is twice as likely as *AA*. By the same reasoning, there is likewise a 2/3 chance that #9 is a heterozygous carrier of the disease allele.

If individual 6 is a heterozygous for the disease allele, then there is a ½ chance that #12 will also be a heterozygote (i.e. if the mating of #6 and #7 is *Aa* × *AA*, half of the progeny will be *Aa*; we are also assuming that #7, who is unrelated, does not carry any disease alleles). Therefore, the combined probability that #12 is also a heterozygote is 2/3 x 1/2 = 1/3. This reasoning also applies to individual #13, i.e. there is a 1/3 probability that he is a heterozygote for the disease. Thus, the overall probability that both individual #12 and #13 are heterozygous, and that a particular offspring of theirs will be homozygous for the disease alleles is 1/3 x 1/3 x 1/4 = 1/36.

POPULATION GENETICS

Figure 5.12

Allele frequencies may also be studied on the population level.

A **population** is a large group of individuals of the same species, who are capable of mating with each other. It is useful to know the frequency of particular alleles within a population, since this information can be used to calculate disease risks. Population genetics is also important in ecology and evolution, since changes in allele frequencies may be associated with migration or natural selection.

When calculating the frequency of two alleles of the same locus (e.g. *A*,*a*), we use the symbol **p** to represent the frequency of the dominant allele within the population, and **q** for the frequency of the recessive allele. Because there are only two possible alleles, we can say that the frequency of p and q together represent 100% of the alleles in the population (**p+q=1**).

If we know the genotypes of a representative sample of individuals in a population, we can calculate the values of p and q by simply counting the alleles and dividing by the total number of alleles examined; for a give allele, homozygotes will count for twice as much as heterozygotes. For example, given:

genotype	number of individuals
AA	320
Aa	160
aa	20

p = 2 (AA) + Aa / total alleles counted = 2(320) + 160 / 2(320) + 2(160) + 2(20) = 0.8

q= 2 (aa) + Aa / total alleles counted = 2(20) + 160 / 2(320) + 2(160) + 2(20) = 0.2

Given the allele frequencies within a population we can use an extension of the Punnett Square, and the product rule, to calculate the expected frequency of each genotype following matings of the entire population. This is the basis of the **Hardy-Weinberg formula**: **$p^2 + 2pq + q^2=1$**, where p^2 is the frequency of homozygotes *AA*, where 2pq is the frequency of the heterozygotes, and q^2 is the frequency of homozygotes *aa*.

	***A* (p)**	***a* (q)**
***A* (p)**	***p^2***	***pq***
***a* (q)**	***pq***	***q^2***

Notice that if we substitute the allele frequencies we calculated above (p=0.8, q=0.2) into the formula **$p^2 + 2pq + q^2=1$,** we obtain expected probabilities for each of the genotypes that exactly match our original observations:

$p^2=0.8^2=0.64$	0.64 x 500 = 320
2pq= 2(0.8)(0.2)=0.32	0.32 x 500 = 160
$q^2=0.2^2=0.04$	0.04 x 500 = 20

This is a demonstration of the **Hardy-Weinberg equilibrium** , which states that both the genotype frequencies and allele frequencies in a population remain unchanged following successive matings within a population, *if* certain conditions are met. These conditions are listed in Table 5.1. Few natural populations actually satisfy all of these conditions. Nevertheless, large populations of many species, including humans, appear to approach Hardy-Weinberg equilibrium for many loci. In these situations, deviations of a particular gene from Hardy-Weinberg equilibirum can be an indication that one of the alleles affects the reproductive success of organism, for example through natural selection or **assortative mating**.

Table 5.1. Conditions for the Hardy-Weinberg equilibrium

1. Random mating: Individuals of all genotypes mate together with equal frequency. Assortative mating, in which certain genotypes preferentially mate together, is a type of non-random mating.
2. No natural selection: All genotypes have equal fitness.
3. No migration: Individuals do not leave or enter the population.
4. No mutation: The allele frequencies do not change due to mutation.
5. Large population: Random sampling effects in mating (i.e. genetic drift) are insignificant in large populations.

The Hardy-Weinberg formula can also be used to estimate allele frequencies, when only the frequency of one of the genotypic classes is known. For example, if 0.04% of the population is affected by a particular genetic condition, and all of the affected individuals have the genotype aa, then we assume that $q^2 = 0.0004$ and we can calculate p, q, and 2pq as follows:

$$q^2 = 0.04\% = 0.0004$$

$$q= \sqrt{0.0004} = 0.02$$

$$p= 1-q = 0.98$$

$$2pq = 2(0.98)(0.02) = 0.04$$

Thus, approximately 4% of the population is expected to be heterozygous (i.e. a carrier) of this genetic condition. Note that while we recognize that the population is probably not exactly in Hardy-Weinberg equilibrium for this locus, application of the Hardy-Weinberg formula nevertheless can give a reasonable estimate of allele frequencies, in the absence of any other information.

SUMMARY

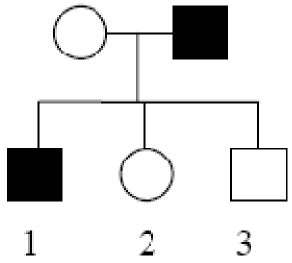
- Pedigree analysis can be used to determine the mode of inheritance of specific traits such as diseases.
- If the mode of inheritance is known, a pedigree can be used to calculate the probability of inheritance of a particular genotype by an individual.
- The frequencies of all alleles and genotypes remain unchanged through successive generations of a population in Hardy-Weinberg equilibrium.
- Populations in true Hardy-Weinberg equilibrium have random mating, and no genetic drift, no migration, no mutation, and no selection with respect to the gene of interest.
- The Hardy-Weinberg formula can be used to estimate allele and genotype frequencies given only limited information about a population.

KEY TERMS

mode of inheritance
genetic counseling
carrier
autosomal dominant
autosomal recessive
X-linked dominant
X-linked recessive
product rule
sum rule
population
p+q=1
Hardy-Weinberg formula
$p^2 + 2pq + q^2=1$
Hardy-Weinberg equilibrium
assortative mating
random mating
migration
genetic drift

STUDY QUESTIONS

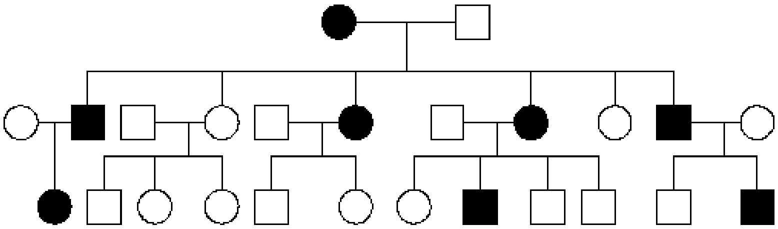
1. What are some of the modes of inheritance that are consistent with this pedigree?



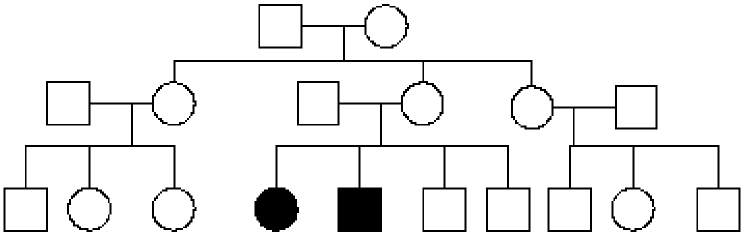
2. In this pedigree in question 1, the mode of inheritance cannot be determined unamibguously. What are some examples of data (e.g. from other generations) that, if added to the pedigree would help determine the mode of inheritance?

3. For each of the following pedigrees, name the most likely mode of inheritance (AR=autosomal recessive, AD=autosomal dominant, XR=X-linked recessive, XD=X-linked dominant). (These pedigrees were obtained from various external sources).

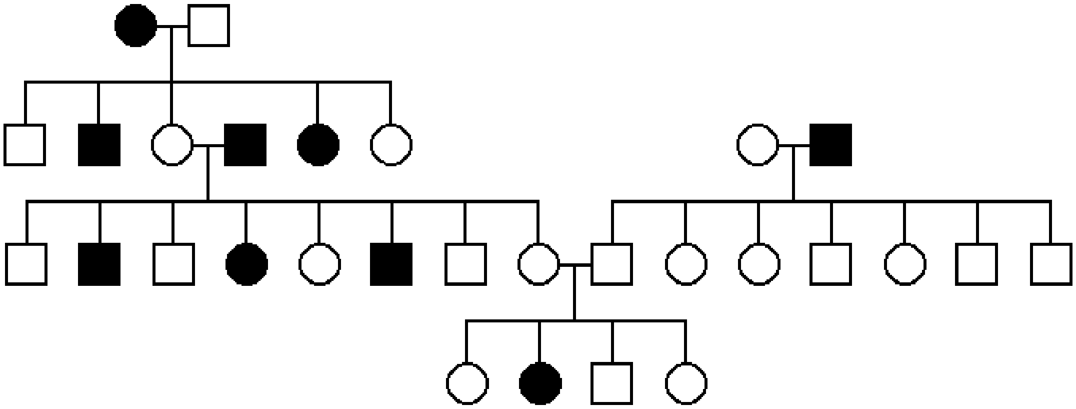
a)



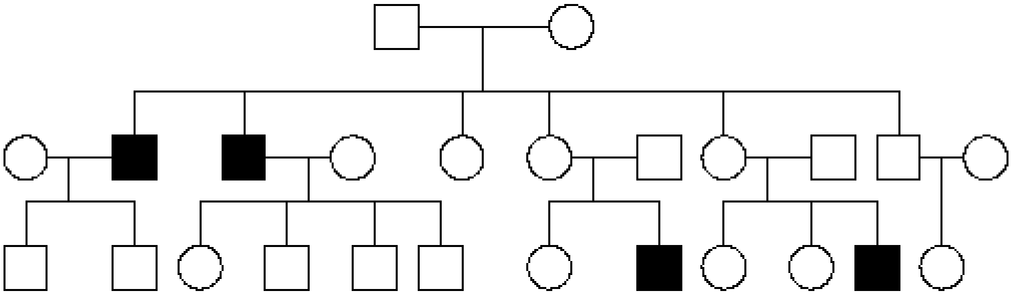
b)



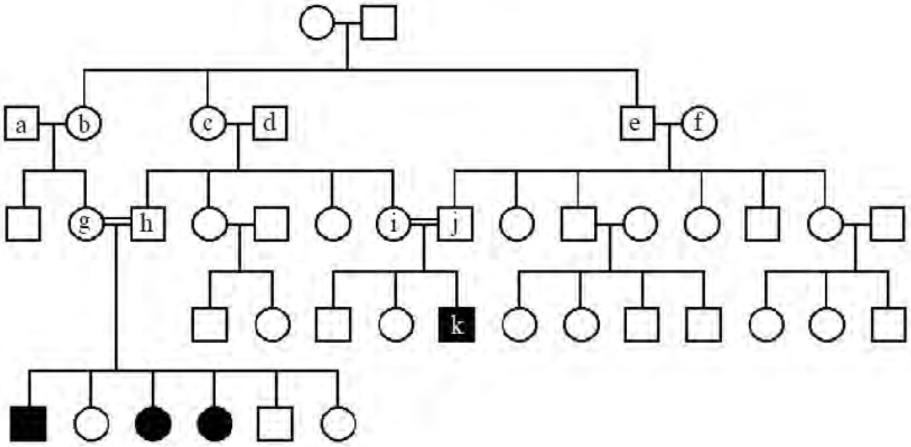
c)



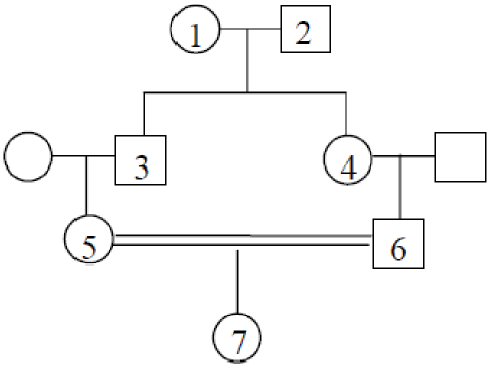
d)



4. The following pedigree represents a rare, autosomal recessive disease. What are the genotypes of the individuals who are indicated by letters?



5. If individual #1 in the following pedigree is a heterozygote for a rare, AR disease, what is the probability that individual #7 will be affected by the disease? Assume that #2 and the spouses of #3 and #4 are not carriers.



6. You are studying a population in which the frequency of individuals with a recessive homozygous genotype is 1%. Assuming the population is in Hardy-Weinberg equilibrium, calculate:

a) The frequency of the recessive allele.
b) The frequency of dominant allele.
c) The frequency of the heterozygous phenotype.
d) The frequency of the homozygous dominant phenotype.

7. Determine whether the following population is in Hardy-Weinberg equilibrium.

genotype	number of individuals
AA	432
Aa	676
aa	92

8. Out of 1200 individuals examined, 432 are homozygous dominant (*AA*)for a particular gene. What numbers of individuals of the other two genotypic classes (*Aa, aa*) would be expected if the population is in Hardy-Weinberg equilibrium?

9. Propose an explanation for the deviation between the genotypic frequencies calculated in question 8 and those observed in the table in question 7.

Chapter 6 GENETIC ANALYSIS OF TWO LOCI

Figure 6.1 Coat color in animals is an example of a trait that affected by more than one locus.

The principles of genetic analysis that we have described for a single locus can be extended to the study of two genes simultaneously. Analysis of two genes in parallel is required for genetic mapping and can also reveal gene interactions. These techniques can be very useful for both basic and applied research. Before discussing these techniques in detail, we will first revisit Mendel and his groundbreaking experiments.

MENDEL'S SECOND LAW

Before Mendel, it had not yet been established that heritable traits were controlled by discrete factors. An important question was therefore whether distinct traits could be shown to be controlled by factors that were inherited independently. To answer this, Mendel took two apparently unrelated traits, such as seed shape and seed color, and studied their inheritance in the same individuals. He studied two variants of each trait: seed color was either green or yellow, and seed shape was either round or wrinkled. When either of these traits was studied alone, the phenotypes segregated in a 3:1 ratio among the progeny of a monohybrid cross (Figure 6.2), with ¾ of the seeds green and ¼ yellow in one cross, and ¾ round and ¼ wrinkled in the other cross. To analyze the segregation of the two traits at the same time, he crossed a pure breeding line of green, wrinkled peas with a pure breeding line of yellow, round peas to produce F$_1$ progeny that were all green and round, and which were also **dihybrids**; they carried two alleles at each of two loci (Figure 6.3),.

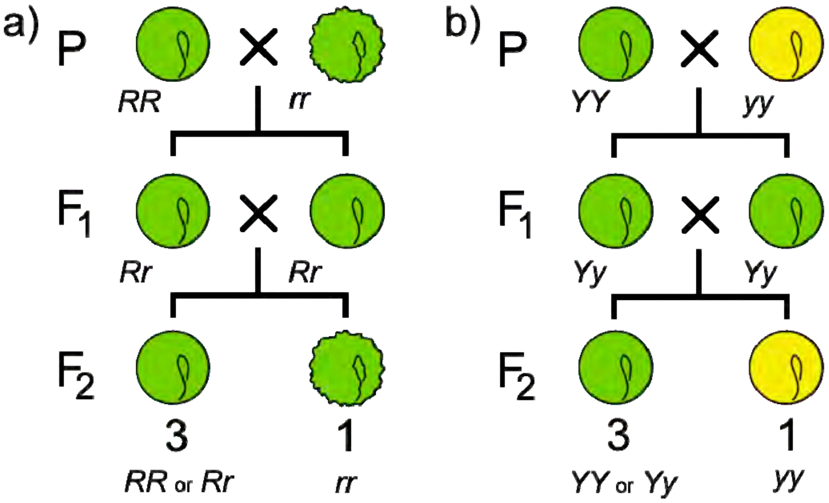


Figure 6.2
Monohybrid crosses involving two distinct traits in peas.

If the inheritance of seed color was truly independent of seed shape, then when the F$_1$ dihybrids were crossed to each other, a 3:1 ratio of one trait should be observed within each phenotypic class of the other trait (Figure 6.3). Using the product law, we would therefore predict that if ¾ of the progeny were green, and ¾ of the progeny were round, then ¾ × ¾ = 9/16 of the progeny would be both round and green (Table 6.1). Likewise, ¾ × ¼ = 3/16 of the progeny would be both round and yellow, and so on. By applying the product rule to all of these combinations of phenotypes, we can predict a **9:3:3:1** phenotypic ratio among the progeny of a dihybrid cross, if certain conditions are met, including the independent segregation of the alleles at each locus. Indeed, 9:3:3:1 is very close to the ratio Mendel observed in his studies of dihybrid crosses, leading him to state his Second Law, the **Law of Independent Assortment**, which we now express as follows: two loci assort independently of each other during gamete formation.

Table 6.1 Phenotypic classes expected in monohybrid and dihybrid crosses for two seed traits in pea.

Frequency of phenotypic crosses within separate monohybrid crosses:

seed shape:	¾ round	¼ wrinkled
seed color:	¾ green	¼ yellow

Frequency of phenotypic crosses within a dihybrid cross:

¾ round	×	¾ green	=	9/16 round & green
¾ round	×	¼ yellow	=	3/16 round & yellow
¼ wrinkled	×	¾ green	=	3/16 wrinkled & green
¼ wrinkled	×	¼ yellow	=	1/16 wrinkled &yellow

The 9:3:3:1 phenotypic ratio that we calculated using the product rule can also be obtained using Punnett Square (Figure 6.4). First, we list the genotypes of the possible gametes along each axis of the Punnett Square. In a diploid with two heterozygous genes of interest, there are up to four combinations of alleles in the gametes of each parent. The gametes from the respective rows and column are then combined in the each cell of the array. When working with two loci, genotypes are written with the symbols for both alleles of one locus, followed by both alleles of the next locus (e.g. *AaBb*, not *ABab*). Note that the order in which the loci are written does not imply anything about the actual position of the loci on the chromosomes

To calculate the expected phenotypic ratios, we assign a phenotype to each of the 16 genotypes in the Punnet Square, based on our knowledge of the alleles and their dominance relationships. In the case of Mendel's seeds, any genotype with at least one *R* allele and one *Y* allele will be round and green; these genotypes are shown in the nine, green-shaded cells in Figure 6.4. We can represent all of four of the different genotypes shown in these cells with the notation (*R_Y_*), where the blank line (__), means "any allele". The three offspring that have at least one R allele and are homozygous recessive for *y* (i.e. *R_yy*) will have a round, yellow phenotype. Conversely the three progeny that are homozygous recessive *r*, but have at least one *Y* allele (*rrY_*) will have wrinkled, green seeds. Finally, the rarest phenotypic class of wrinkled, yellow seeds is produced by the doubly homozygous recessive genotype, *rryy*, which is expected to occur in only one of the sixteen possible offspring represented in the square.

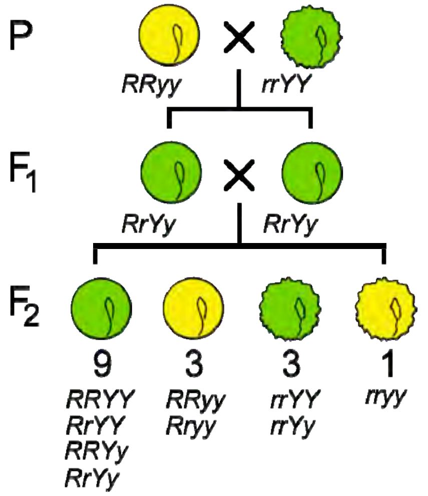
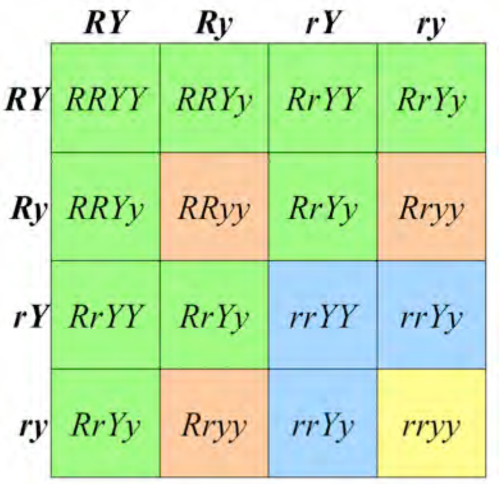


Figure 6.3 Pure-breeding lines are crossed to produce dihybrids in the F_1 generation. The cross of these particular dihybrids produces four phenotypic classes.



	RY	*Ry*	*rY*	*ry*
RY	*RRYY*	*RRYy*	*RrYY*	*RrYy*
Ry	*RRYy*	*RRyy*	*RrYy*	*Rryy*
rY	*RrYY*	*RrYy*	*rrYY*	*rrYy*
ry	*RrYy*	*Rryy*	*rrYy*	*rryy*

Figure 6.4 A Punnett Square showing the results of the dihybrid cross from Figure 6.3. Each of the four phenotypic classes is represented by a different color of shading: round & green (green); round & yellow (red); wrinkled & green (blue); wrinkled & yellow (yellow).

ASSUMPTIONS OF THE 9:3:3:1 RATIO

Both the product rule and the Punnett Square approaches showed that a 9:3:3:1 phenotypic ratio is expected among the progeny of a dihybrid cross such as Mendel's *RrYy* × *RrYy*. In making these calculations, we assumed that: both loci assort independently; one allele at each locus is completely dominant; and each of four possible phenotypes can be distinguished unambiguously, with no interactions between the two genes that would alter the phenotypes. Deviations from the 9:3:3:1 phenotypic ratio may indicate that one of more of these conditions has not been met. Modified ratios in the progeny of a dihybrid cross can therefore reveal useful information about the genes involved. **Linkage** is one of the most important reasons for distortion of the ratios expected from independent assortment. Linked genes are located close together on the same chromosome, which affects which combinations of alleles assort together most frequently. We will return to the concept of linkage in Chapter 7. Deviations from 9:3:3:1 ratios can also be due to interactions between genes; these interactions will be discussed throughout the remainder of this chapter. For simplicity, we will focus on examples that involve easily scored phenotypes, such as pigmentation. Nevertheless, keep in mind that the analysis of segregation ratios can provide insight into a wide range of biological processes.

EPISTASIS AND OTHER GENE INTERACTIONS

Some dihybrid crosses produce a phenotypic ratio that differs from 9:3:3:1 because of epistasis. Epistasis (which means "standing upon") occurs when the phenotype of one locus masks the phenotype of another locus. Thus, following a dihybrid cross with epistasis, fewer than four phenotypic classes will be observed, with ratios such as 9:3:4, 12:3:1, 9:7, or 15:1. Note that each of these modified ratios can be obtained by summing one or more of the 9:3:3:1 classes expected from our original dihybrid cross. In the following paragraphs, we will look at some modified phenotypic ratios obtained from dihybrid crosses and what they might tell us about interactions between genes.

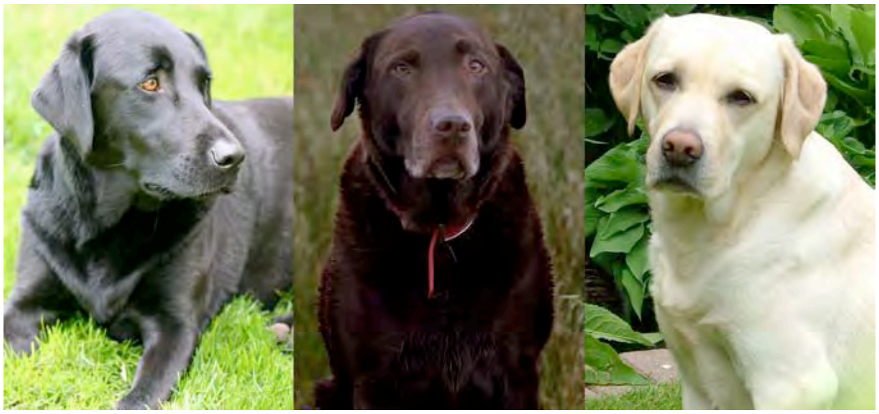
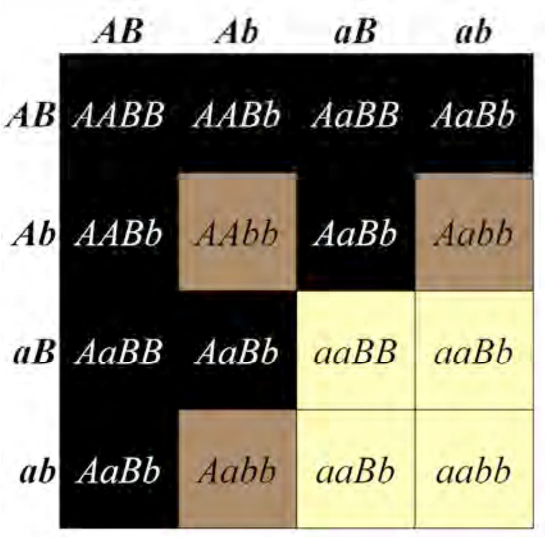


Figure 6.5 Labrador Retrievers with different coat colors: (from left to right) black, chocolate, yellow

RECESSIVE EPISTASIS

As we have already discussed, in the absence of epistasis, there are four phenotypic classes among the progeny of a dihybrid cross. The four phenotypic classes correspond to the genotypes: *A_B_, A_bb, aaB_,* and *aabb*. If either of the singly homozygous recessive genotypes (i.e. *A_bb* or *aaB_*) has the same phenotype as the double homozygous recessive (*aabb*), then a **9:3:4** phenotypic ratio will be obtained. For example, in the Labrador Retriever breed of dogs (Figure 6.5), the B locus encodes a gene for an important step in the production of melanin. The dominant allele, *B* is more efficient at pigment production than the recessive *b* allele, thus *B_* hair appears black, and *bb* hair appears brown. A second locus, which we will call *A*, regulates the production of melanin. At least one functional *A* allele is required to produce any pigment, whether it is black or brown. Thus, all retrievers that are *aa* fail to produce any melanin (and so appear pale yellow), regardless of the genotype at the *B* locus (Figure 6.6). The *aa* genotype is therefore said to be epistatic to both the B and b alleles, since the homozygous *aa* phenotype masks the phenotype of the B locus. Because the masking allele is in this case is recessive, this is called **recessive epistasis**.



	AB	*Ab*	*aB*	*ab*
AB	*AABB*	*AABb*	*AaBB*	*AaBb*
Ab	*AABb*	*AAbb*	*AaBb*	*Aabb*
aB	*AaBB*	*AaBb*	*aaBB*	*aaBb*
ab	*AaBb*	*Aabb*	*aaBb*	*aabb*

Figure 6.6 Genotypes and phenotypes among the progeny of a dihybrid cross of Labrador Retrievers heterozygous for two loci affecting coat color. The phenotypes of the progeny are indicated by the shading of the cells in the table: black coat (black, *A_B_*); chocolate coat (brown, *A_bb*); yellow coat (yellow, *aaB_* or *aabb*).

DOMINANT EPISTASIS

In some cases, a dominant allele at one locus may mask the phenotype of a second locus. This is called **dominant epistasis**, which produces a segregation ratio such as **12:3:1**, which can be viewed as a modification of the 9:3:3:1 ratio in which the *A_B_* class is combined with one of the other genotypic classes that contains a dominant allele. One of the best known examples of a 12:3:1 segregation ratio is fruit color in some types of squash (Figure 6.7). Alleles of a locus that we will call *B* produce either yellow (*B_*) or green (*bb*) fruit. However, in the presence of a dominant allele at a second locus that we call *A*, no pigment is produced at all, and fruit are white. The *A* allele is therefore epistatic to both *B* and *bb* combinations (Figure 6.8). One possible biological interpretation of this segregation pattern is that the function of the *A* allele somehow blocks an early stage of pigment synthesis, before either yellow or green pigments are produced.

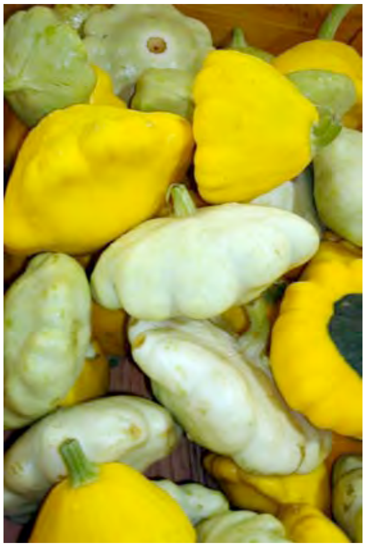
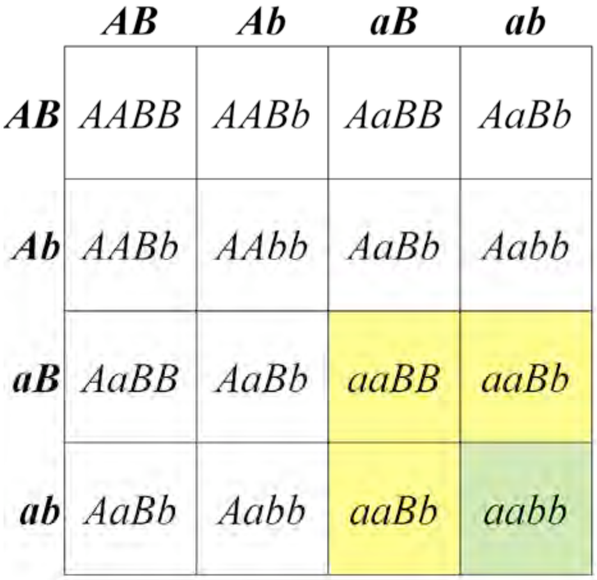


Figure 6.7 Green, yellow, and white fruits of squash.



	AB	***Ab***	***aB***	***ab***
AB	*AABB*	*AABb*	*AaBB*	*AaBb*
Ab	*AABb*	*AAbb*	*AaBb*	*Aabb*
aB	*AaBB*	*AaBb*	*aaBB*	*aaBb*
ab	*AaBb*	*Aabb*	*aaBb*	*aabb*

Figure 6.8 Genotypes and phenotypes among the progeny of a dihybrid cross of squash plants heterozygous for two loci affecting fruit color.

COMPLEMENTARY GENE ACTION

The progeny of a dihybrid cross may produce just two phenotypic classes, in an approximately **9:7** ratio. An interpretation of this ratio is that the loss of function of either *A* or *B* gene function has the same phenotype as the loss of function of both genes, due to **complementary gene action** (meaning that the functions of both genes work together to produce a final product). For example, consider a simple biochemical pathway in which a colorless substrate is converted by the action of gene A to another colorless product, which is then converted by the action of gene B to a visible pigment (Figure 6.9). Loss of function of either A or B, or both, will have the same result: no pigment production. Thus *A_bb*, *aaB_*, and *aabb* will all be colorless, while only *A_B_* genotypes will produce pigmented product (Figure 6.10). The modified 9:7 ratio may therefore be obtained when two genes act together in the same biochemical pathway, and when their loss of function phenotypes are indistinguishable from each other or from the loss of both genes.

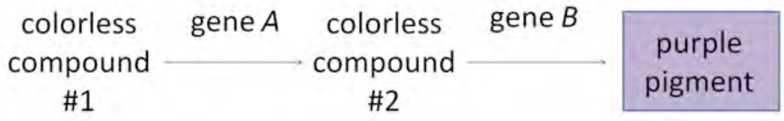
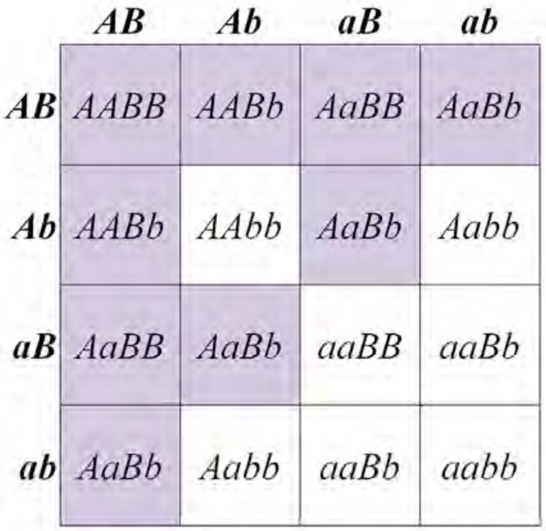


Figure 6.9 A simplified biochemical pathway showing complementary gene action of A and B. Note that in this case, the same phenotypic ratios would be obtained if gene B acted before gene A in the pathway



	AB	***Ab***	***aB***	***ab***
AB	*AABB*	*AABb*	*AaBB*	*AaBb*
Ab	*AABb*	*AAbb*	*AaBb*	*Aabb*
aB	*AaBB*	*AaBb*	*aaBB*	*aaBb*
ab	*AaBb*	*Aabb*	*aaBb*	*aabb*

Figure 6.10 Genotypes and phenotypes among the progeny of a dihybrid cross of a hypothetical plant heterozygous for two loci affecting flower color.

DUPLICATE GENE ACTION

When a dihybrid cross produces progeny in two phenotypic classes in a 15:1 ratio, this can be because the two loci have redundant functions within the same biological pathway. Yet another pigmentation pathway, in this case in wheat, provides an example of this **duplicate gene action**. The biosynthesis of red pigment near the surface of wheat seeds (Figure 6.11) involves many genes, two of which we will label *A* and *B*. Normal, red coloration of the wheat seeds is maintained if function of either of these genes is lost in a homozygous mutant (e.g. in either *aaB_* or *A_bb*). Only the doubly recessive mutant (*aabb*), which lacks function of both genes, shows a phenotype that differs from that produced by any of the other genotypes (Figure 6.12). A reasonable interpretation of this result is that both genes encode the same biological function, and either one alone is sufficient for the normal activity of that pathway.

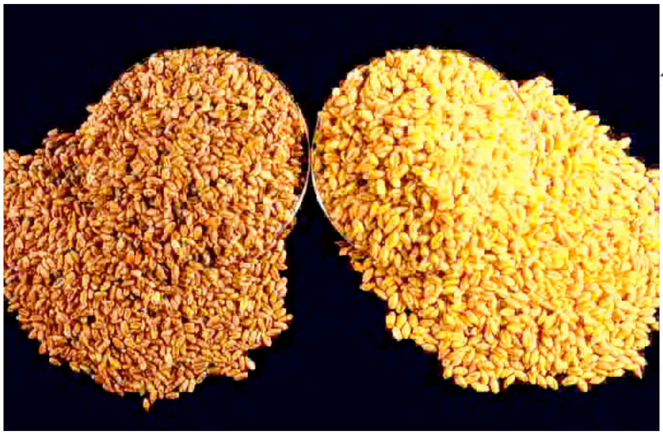
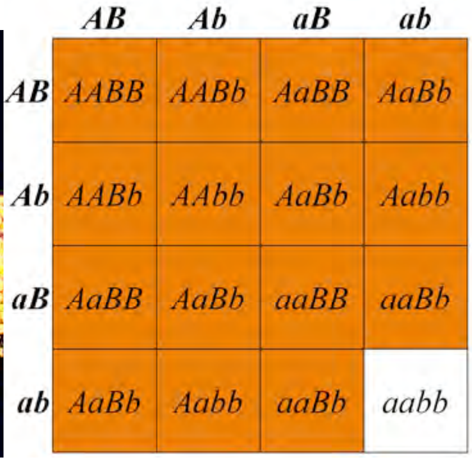


Figure 6.11 Red (left) and white (right) wheat seeds. cropwatch.unl.edu



	AB	*Ab*	*aB*	*ab*
AB	*AABB*	*AABb*	*AaBB*	*AaBb*
Ab	*AABb*	*AAbb*	*AaBb*	*Aabb*
aB	*AaBB*	*AaBb*	*aaBB*	*aaBb*
ab	*AaBb*	*Aabb*	*aaBb*	*aabb*

Figure 6.12 Genotypes and phenotypes among the progeny of a dihybrid cross of a wheat plants heterozygous for two loci affecting seed color.

ENHANCER /SUPPRESSOR SCREENS

From the examples of genetic interactions already discussed in this chapter, it should be apparent that the analysis of phenotypes at two loci simultaneously can provide insight into the functions of genes and biochemical pathways. Accordingly, geneticists sometimes attempt to identify interacting genes through mutant screening, using a technique called enhancer/suppressor screening. Typically, a researcher will mutagenize a population that is already has an interesting mutant phenotype in one gene (*aa*), then screen through the progeny of the re-mutagenized individuals to find additional mutations that either make the original *aa* mutants look more like wild-type (i.e. **suppressors** or **revertants**), or else increase (i.e. **enhance**) the severity of the original *aa* phenotype. Note that this use of the term enhancer is unrelated to its use in the context of transcriptional gene regulation (see Chapter 10).

Reversions, as implied by their name, are mutation that reverses the sequence changes caused by a preceding mutation. Reversions therefore restore the wild-type function of a gene. For example, if the first mutation introduced a stop codon within the coding sequence of a gene, a reversion would result from mutation of that stop codon to again code for an amino acid that restores normal function of the protein. Reversions are genetically less interesting than some other outcomes potential outcomes of an enhancer/suppressor screen.

Suppressors also restore some or all of the wild-type function that is lost in an existing mutant (*aa*), either through mutation of a different site within the same gene (i.e. an **intragenic** suppressor), or by mutation of a different gene (i.e. an **intergenic** suppressor). There are many mechanisms by which intergenic suppressor mutations may restore wild-type function. One way is through the modification of an interacting protein. For example, imagine that proteins produced by genes *A* and *B* must physically connect with each other to perform their normal function (Figure 6.13). If a mutant allele (*a*) was structurally altered in some way that made it unable to interact with the protein produced by *B*, then there would be a blockage of the biochemical pathway normally catalyzed by the two wild-type genes. However, if through additional random mutagenesis, *B* was mutated in a way that changed its structure to allow it to again connect with a, then the wild-type function of the protein complex might be restored. In this case, *b* would be said to suppress *aa*. In this way, a novel gene (*B*) that normally interacts with the original gene of interest (*A*), could be identified through the mutagenesis of a mutant.

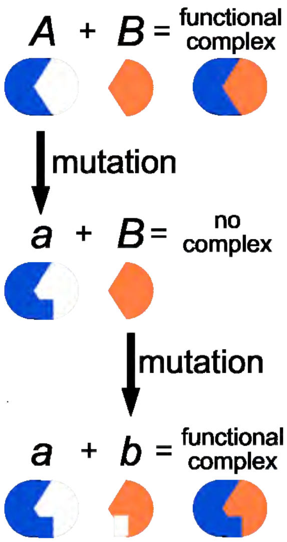


Figure 6.13 Mutation in allele *A* leads to a loss of function because of a failure to form an active complex with *B*. However, a second (suppressor) mutation allows the two mutant alleles to again connect in a functional complex.

Enhancer mutations increase or expand an existing mutant phenotype, often by reducing the function of other genes that normally interact co-operatively or redundantly within a biochemical pathway. Like revertants and suppressors, enhancers can be identified by screening after mutagenesis of a (usually homozygous) mutant (e.g. *aa*).

A particularly interesting example of a gene identified by enhancer screening is provided by Arabidopsis. Researchers had already identified mutants of Arabidopsis called apetala1 (*ap1*) that lacked the outer organs (sepals, petals) of the flower. To find additional genes that interact with *ap1*, the researchers attempted an enhancer/suppressor screen by mutagenizing *ap1*/*ap1* plants, and then looking for novel mutant phenotypes among their selfed (F_2) progeny. One of the novel, enhanced phenotypes caused structures that looked like tiny cauliflowers to be produced in place of the *ap1* flowers. They named the enhancer gene *CAULIFLOWER* (*CAL*), and further analysis showed that the cauliflower phenotype was produced only when both *ap1* and *cal* were homozygous mutant; in fact *cal*/*cal* alone looked just like wild-type. Since Arabidopsis is in the same taxonomic family as cauliflower, cabbage, they hypothesized that something like the Arabidopsis *cal and ap1* mutations might be what makes the vegetable, cauliflower, look the way that it does. Indeed, analysis of cauliflower from a grocery store ultimately showed that its distinct appearance is the result of mutation in genes like *cal*/*cal* and *ap1*/*ap1*. This is just one example of the unexpected information that can come from enhancer/suppressor screening.

SUMMARY

- The alleles of different genes are inherited independently of each other, unless they are genetically linked.
- The expected phenotypic ratio of a dihybrid cross is 9:3:3:1, except in cases of linkage or gene interactions that modify this ratio.
- Epistasis occurs when the phenotype of one gene masks the phenotype of another gene. This usually indicates that the two genes interact within the same biological pathway.
- Enhancer/suppressor screening is a way to identify genetically interacting genes through further mutagenesis of an existing mutant.

KEY TERMS

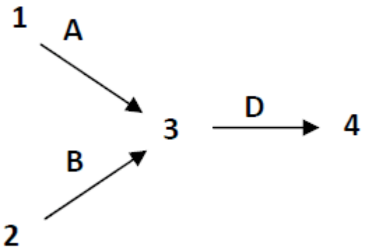
Mendel's Second Law
dihybrid
9:3:3:1
9:3:4
9:7
12:3:1
15:1

independent assortment
linkage
epistasis
recessive epistasis
dominant epistasis
complementary action
redundancy

duplicate gene action
enhancer/suppressor
revertant
intergenic
intragenic
CAULIFLOWER

STUDY QUESTIONS

Answer questions 6.1 -6.3 using the following biochemical pathway for fruit color. Assume all mutations (lower case allele symbols) are recessive, and that *either* precursor 1 or precursor 2 can be used to produce precursor 3. If the alleles for a particular gene are not listed in a genotype, you can assume that they are wild-type.



6.1 If 1 and 2 and 3 are all colorless, and 4 is red, what will be the phenotypes associated with the following genotypes?

a) *aa*
b) *bb*
c) *dd*
d) *aabb*
e) *aadd*
f) *bbdd*
g) *aabbdd*
h) What will be the phenotypic ratios among the offspring of a cross *AaBb* × *AaBb*?
i) What will be the phenotypic ratios among the offspring of a cross *BbDd* × *BbDd*?
j) What will be the phenotypic ratios among the offspring of a cross *AaDd* × *AaDd*?

6.2 If 1 and 2 are both colorless, and 3 is blue and 4 is red, what will be the phenotypes associated with the following genotypes?

a) *aa*
b) *bb*
c) *dd*
d) *aabb*
e) *aadd*
f) *bbdd*
g) *aabbdd*
h) What will be the phenotypic ratios among the offspring of a cross *AaBb* × *AaBb*?
i) What will be the phenotypic ratios among the offspring of a cross *BbDd* × *BbDd*?
j) What will be the phenotypic ratios among the offspring of a cross *AaDd* × *AaDd*?

6.3 If 1 is colorless, 2 is yellow and 3 is blue and 4 is red, what will be the phenotypes associated with the following genotypes?

a) *aa*
b) *bb*
c) *dd*
d) *aabb*
e) *aadd*
f) *bbdd*
g) *aabbdd*
h) What will be the phenotypic ratios among the offspring of a cross *AaBb* × *AaBb*?
i) What will be the phenotypic ratios among the offspring of a cross *BbDd* × *BbDd*?

j) What will be the phenotypic ratios among the offspring of a cross *AaDd* × *AaDd*?

6.4 Which of the situations in questions 6.1 – 6.3 demonstrate epistasis?

6.5 You recover mutants from an enhancer/suppressor screen that look like wild-type. How can you tell whether the restoration of the phenotype is due to a mutation in the same gene as was originally mutated. or it a second gene has been mutated?

6.6 If the genotypes written within the Punnett Square aref rom the F_2 generation, what would be the phenotypes and genotypes of the F_1 and P generations for:

a) Figure 6.6

b) Figure 6.8

c) Figure 6.10

d) Figure 6.12

6.7 To better understand how genes control the development of three-dimensional structures, you conducted a mutant screen in Arabidopsis and identified a recessive point mutation allele of a single gene (*g*) that causes leaves to develop as narrow tubes rather than the broad flat surfacesthat develop in wild-type (*G*). Allele *g* causes a complete loss of function. Now you want to identify more genes involved in the same process. Diagram a process you could use to identify other genes that interact with gene *g*. Show all of the possible genotypes that could arise in the F_1 generation.

6.8 With reference to question 6.7, if the recessive allele, *g* is mutated again to make allele *g**, what are the possible phenotypes of a homozygous *g* g** individual?

6.9 Again, in reference to question 6.7, what are the possible phenotypes of a homozygous *aagg* individual, where *a* is a recessive allele of a second gene? In each case, also specify the phenotypic ratios that would be observed among the F_1 progeny of a cross of *AaGg* x *AaGg*

6.10 Calculate the phenotypic ratios from a dihybrid cross involving the two loci shown in Figure 6.13. There may be more than one possible set of ratios, depending on the assumptions you make about the phenotype of allele *b*.

6.11 Use the product rule to calculate the phenotypic ratios expected from a trihybrid cross. Assume independent assortment and no epistasis/gene interactions.

Chapter 7 LINKAGE & MAPPING

Figure 7.1 Linkage affects the frequency at which some combinations of traits are observed.

As we learned in Chapter 6, Mendel reported that the pairs of genes he observed behaved independently of each other; for example, the segregation of seed color alleles was independent from the segregation of alleles for seed shape. This observation was the basis for his Second Law, and contributed greatly to our understanding of heredity. However, further research showed that Mendel's Second Law did not apply to every pair of genes that could be studied. In fact, we now know that alleles of genes that are located close together on the same chromosome tend to be inherited together. This phenomenon is called **linkage**, and is a major exception to Mendel's Second Law. Understanding linkage is important to natural processes of heredity and evolution, and we will learn how researchers use linkage to determine the location of genes along chromosomes in a process called genetic mapping.

RECOMBINATION

Recombination is term that is used in several different contexts in genetics. In reference to heredity, **recombination** is defined as any process that results in gametes with combinations of alleles that were not present in the gametes of a previous generation (see Figure 7.2). Recombination may occur either through random assortment of alleles that are on different chromosomes, or through **crossovers** between loci on the same chromosomes (as described below). It is important to remember that in both of these cases, recombination is a process that occurs during meiosis (mitotic recombination may also occur in some species, but it is relatively rare). If meiosis results in recombination, the products are said to have a **recombinant** genotype. On the other hand, if no recombination occurs during meiosis, the products are said to have a non-recombinant, or **parental** genotype. Recombination is important because it contributes to the variation that may be observed between individuals within a population.

As an example of recombination, consider loci on two different chromosomes as shown in Figure 7.2. We know that if these loci are on different chromosomes, there are no physical connections between them, so they are **unlinked** and will segregate independently as did Mendel's traits. When loci are unlinked, can we predict which combinations of alleles will segregate together in any given meiosis? No, not with certainty: the segregation depends on the relative orientation of each pair of chromosomes at metaphase. Since the orientation is random and independent of other chromosomes, each of the arrangements (and their meiotic products) is equally possible for two unlinked loci as shown in Figure 7.2. More precisely, there is a 50% probability of recombinant genotypes, and a 50% probability of parental genotypes within the gametes produced by unlinked loci. Indeed, if we examined all of the gametes produced by this individual (which are the products of multiple independent meioses), we would note that approximately 50% of the gametes would be recombinant, and 50% would be parental. **Recombination frequency** is simply the number of recombinant gametes, divided by the total number of gametes. A frequency of approximately 50% recombination is therefore a defining characteristic of unlinked loci.

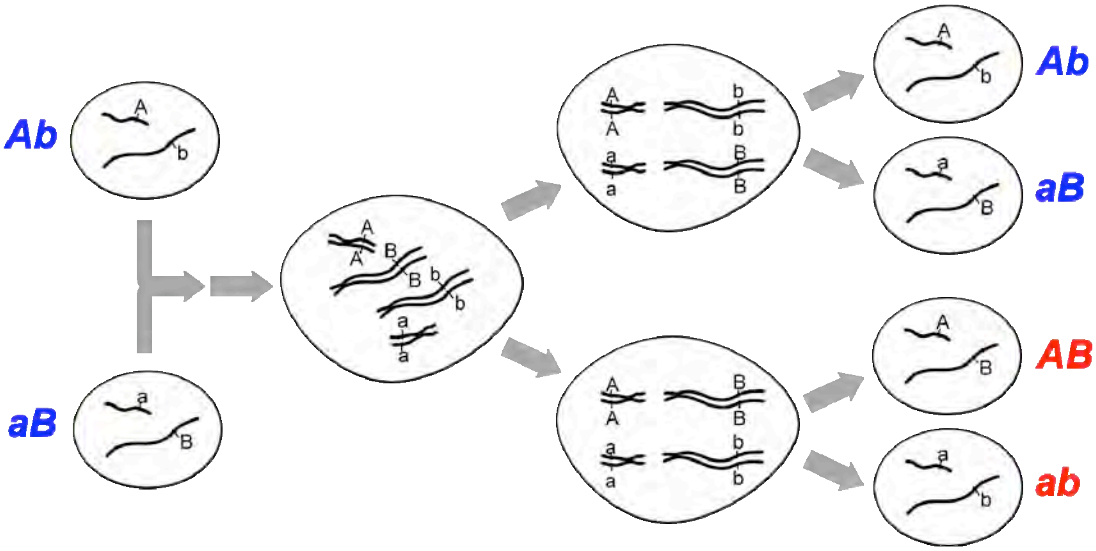


Figure 7.2 When two loci are on non-homologous chromosomes, their alleles will segregate in combinations identical to those present in the parental gametes (*Ab*, *aB*), and in recombinant genotypes (*AB*, *ab*) that are different from the parental gametes. Not all of the identical products of each meiosis are shown here.

LINKAGE SUPPRESSES RECOMBINATION

Having considered unlinked loci, let us turn to the opposite situation, in which two loci are so close together on a chromosome that the parental combinations of alleles always segregate together (Figure 7.3). This is because alleles at the two loci are physically attached and so they always follow each other into the same gamete. In this case, no recombinants will be detected following meiosis, and the recombination frequency will be 0%. **Complete linkage** is rare, as the loci must be so close together that crossovers are never detected between them.

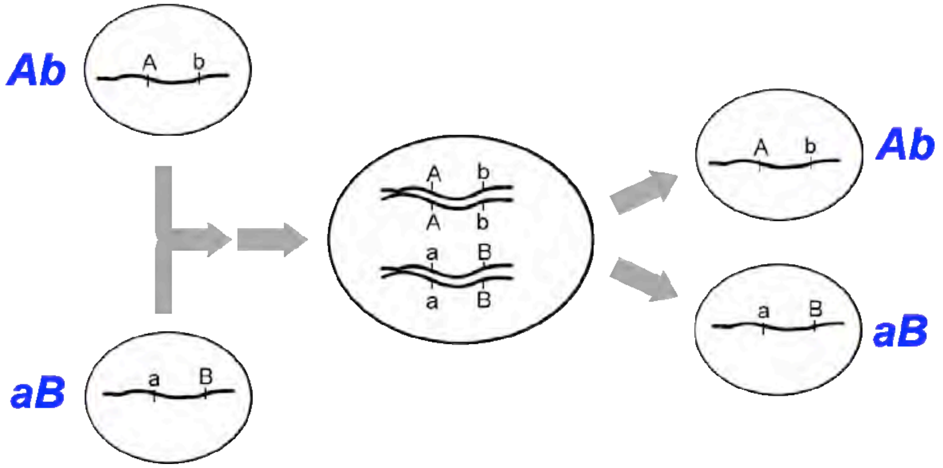


Figure 7.3 If two loci are completely linked, their alleles will segregate in combinations identical to those present in the parental gametes (*Ab*, *aB*). No recombinants will be observed.

CROSSOVERS ALLOW RECOMBINATION BETWEEN LINKED LOCI

Thus far, we have only considered situations with either no linkage (50% recombination) or else complete linkage (0% recombination). It is also possible to obtain recombination frequencies between 0% and 50%, which is a situation we call partial linkage or **incomplete linkage**. Incomplete linkage occurs when two loci are located on the same chromosome (i.e. they are physically linked), but the loci are far enough apart so that crossovers occur between them during some, but not all, meioses.

Crossovers occur during prophase I of meiosis, when pairs of homologous chromosomes have aligned with each other in a process called **synapsis**. Crossing over begins with the double-strand breakage of a pair of non-sister chromatids. The breaks usually occur at corresponding positions on two chromatids, and then the broken ends of non-sister chromatids are connected to each other resulting in a reciprocal exchange of double-stranded DNA (Figure 7.4). Generally every pair of chromosomes has at least one (and often more) crossovers during meioses (Figure 7.5)

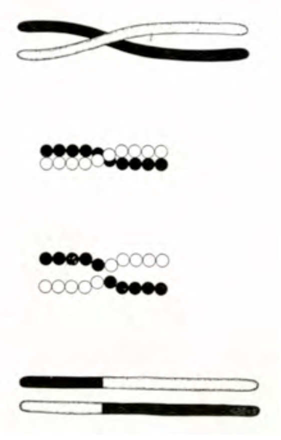


Figure 7.4 A depiction of a crossover from Morgan's 1916 manuscript. Only one pair of non-sister chromatids is shown.

Because the location of crossovers is essentially random, the greater the distance between two loci, the more likely it is that a crossover will occur between them. Furthermore, loci that are on the same chromosome, but are sufficiently far apart from each other, will have crossovers so often that they will behave as though they are completely unlinked. A recombination frequency of 50% is therefore the maximum recombination frequency that can be observed, and is indicative of loci that are either on separate chromosomes, or are located very far apart on the same chromosome.

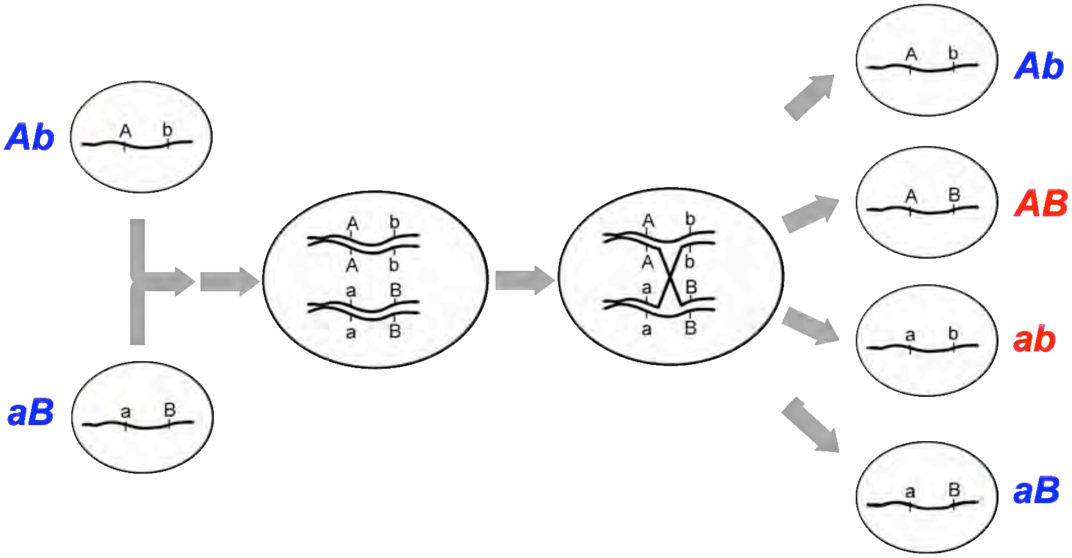


Figure 7.5 A crossover between two linked loci can generate recombinant genotypes (*AB*, *ab*), from the chromatids involved in the crossover. Remember that multiple, independent meioses occur in each organism, so this particular pattern of recombination may not be observed among all the gametes from this individual.

INFERRING RECOMBINATION FROM GENETIC DATA

In the preceding examples, we had the advantage of knowing the approximate chromosomal positions of each allele involved, before we calculated the recombination frequencies. Knowing this information beforehand made it relatively easy to define the parental and recombinant genotypes, and to calculate recombination frequencies. However, in most experiments, we cannot directly examine the chromosomes, or even the gametes, so we must infer the arrangement of alleles from the phenotypes over two or more generations. Importantly, it is generally not sufficient to know the genotype of individuals in just one generation; for example, given an individual with the genotype *AaBb*, we do not know from the genotype alone whether the loci are located on the same chromosome, and if so, whether the arrangement of alleles on each chromosome is or *AB* and *ab* (also called the **cis** , or **coupling** arrangement; Figure 7.6) or *Ab* and *aB* (**trans**, or **repulsion**).

Fortunately for geneticists, the arrangement of alleles can sometimes be inferred if the genotypes of a previous generation are known. For example, if the parents of *AaBb* had genotypes *AAB*B and *aabb* respectively, then the parental gametes that fused to produce *AaBb* would have been genotype *AB* and genotype *ab*. Therefore, prior to meiosis in the dihybrid, the arrangement of alleles would likewise be *AB* and *ab* (Figure 7.7). Conversely, if the parents of *AaBb* had genotypes *aaBB* and *AAbb*, then the arrangement of alleles on the chromosomes of the dihybrid would be *aB* and *Ab*. Thus, the genotype of the previous generation can determine which of an individual's gametes are considered recombinant, and which are considered parental.

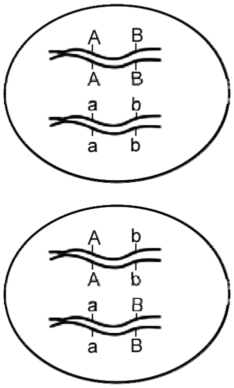


Figure 7.6 Alleles in cis configuration (top) or trans configuration (bottom).

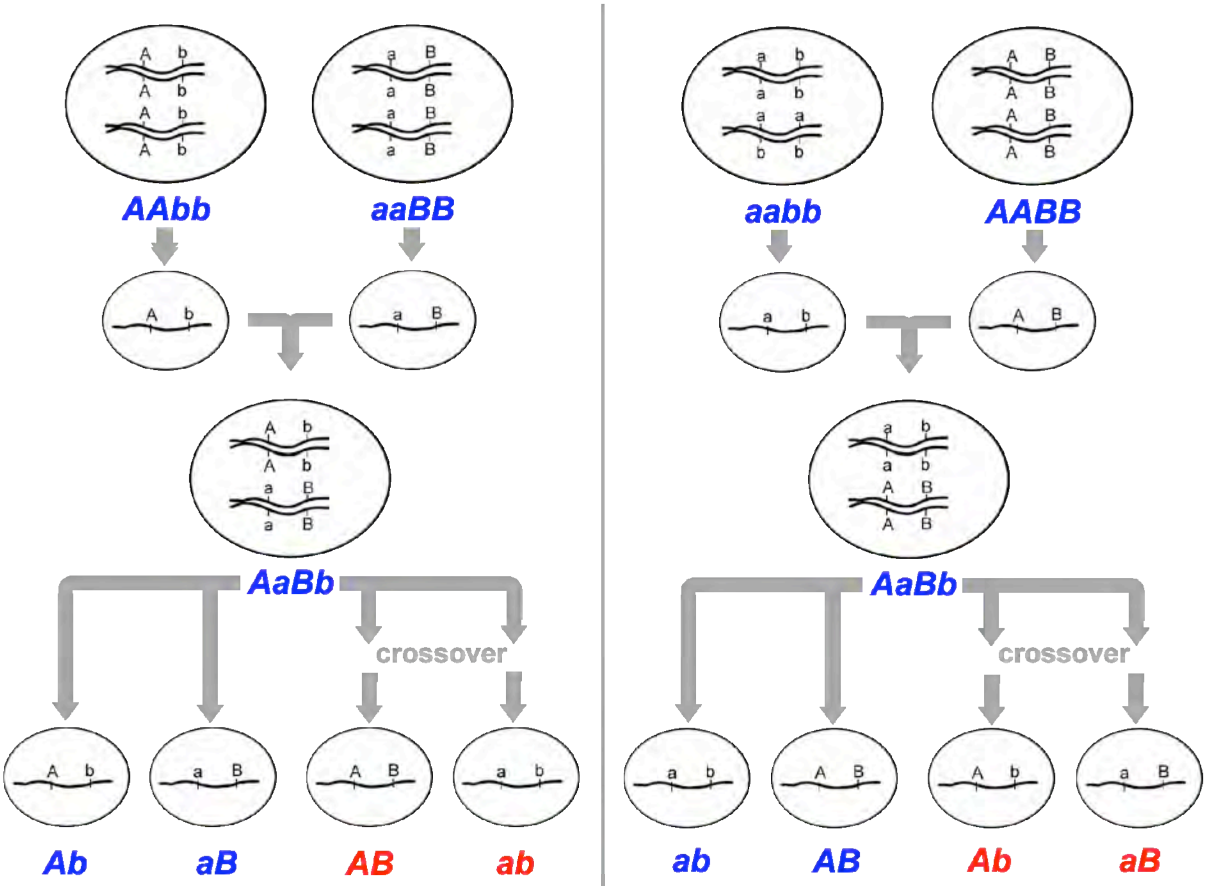


Figure 7.7 The genotype of gametes can be inferred unambiguously if the gametes are produced by homozygotes. However, recombination frequencies can only be measured among the progeny of heterozygotes (i.e. dihybrids). Note that the dihybrid on the left contains a different configuration of alleles than the dihybrid on the rightdue to differences in the genotypes of their respective parents. Therefore, different gametes are defined as recombinant and parental among the progeny of the two dihybrids. In the cross at left, the recombinant gametes will be genotype *AB* and *ab*, and in the cross on the right, the recombinant gametes will be *Ab* and *aB*

Let us now consider a complete experiment in which our objective is to measure recombination frequency (Figure 7.8). We need at least two alleles for each of two genes, and we must know which combinations of alleles were present in the parental gametes. The simplest way to do this is to start with pure-breeding lines that have contrasting alleles at two loci. For example, we could cross short-tailed mice, brown mice (*aaBB*) with long-tailed, white mice (*AAbb*). Based on the genotypes of the parents, we know that the parental gametes will be *aB* or *Ab* (but not *ab* or *AB*), and all of the progeny will be dihybrids, *AaBb*. We do not know at this point whether the two loci are on different pairs of homologous chromosomes, or whether they are on the same chromosome, and if so, how close together they are.

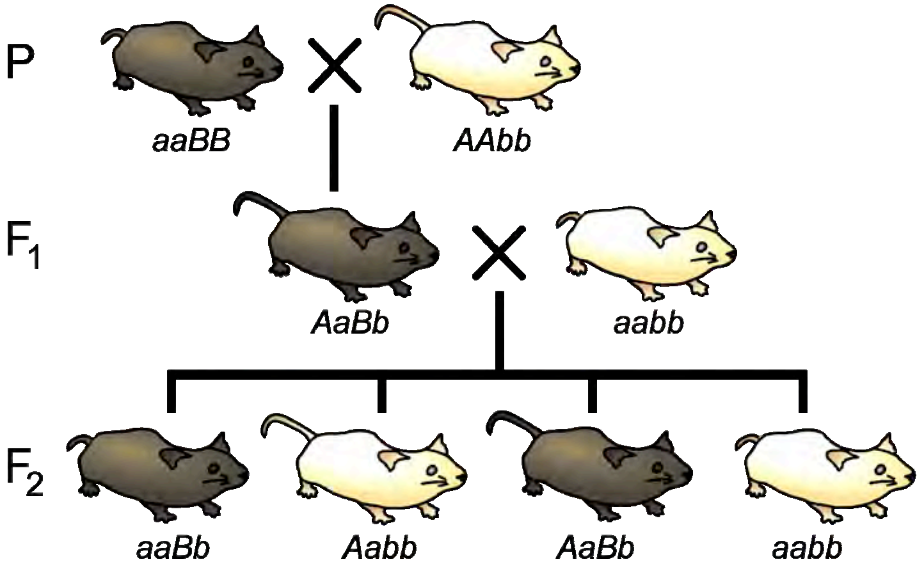


Figure 7.8 An experiment to measure recombination frequency between two loci. The loci affect coat color and tail length.

The recombination events that may be detected will occur during meiosis in the dihybrid individual. If the loci are completely or partially linked, then prior to meiosis, alleles *aB* will be located on one chromosome, and alleles *Ab* will be on the other chromosome (based on our knowledge of the genotypes of the gametes that produced the dihybrid). Thus, recombinant gametes produced by the dihybrid will have the genotypes *ab* or *AB*, and non-recombinant (i.e. parental) gametes will have the genotypes *aB* or *Ab*.

How do we determine the genotype of the gametes produced by the dihybrid individual? The most practical method is to use a testcross (Figure 7.8), in other words to mate *AaBb* to an individual that has only recessive alleles at both loci (*aabb*). This will give a different phenotype in the F_2 generation for each of the four possible combinations of alleles in the gametes of the dihybrid. We can then infer unambiguously the genotype of the gametes produced by the dihybrid individual, and therefore calculate the recombination frequency between these two loci. For example, if only two phenotypic classes were observed in the F_2 (i.e. short tails and brown fur (*aaBb*), and white fur with long tails (*Aabb*) we would know that the only gametes produced following meiosis of the dihybrid individual were of the parental type: *aB* and *Ab*, and the recombination frequency would therefore be 0%. Alternatively, we may observe multiple classes of phenotypes in the F_2 in ratios such as shown in Table 7.1:

tail phenotype	fur phenotype	number of progeny	gamete from dihybrid	genotype of F_2 from test cross	(P)arental or (R)ecombinant
short	brown	48	*aB*	*aaBb*	P
long	white	42	*Ab*	*Aabb*	P
short	white	13	*ab*	*aabb*	R
long	brown	17	*AB*	*AaBb*	R

Table 7.1 An example of quantitative data that may be observed in a genetic mapping experiment involving two loci. The data correspond to the F_2 generation in the cross shown in Figure 7.8.

Given the data in Table 7.1, the calculation of recombination frequency is straightforward:

$$\text{recombination frequency} = \frac{\text{number of recombinant gametes}}{\text{total number of gametes scored}}$$

$$\text{R.F.} = \frac{13+17}{48+42+13+17}$$

$$= 25\ \%$$

GENETIC MAPPING

Because the frequency of recombination between two loci (up to 50%) is proportional to the chromosomal distance between them, we can use recombination frequencies to produce genetic maps. The units of genetic distance are called centiMorgans (cM, also called map units), in honor of Thomas Hunt Morgan. We can easily convert recombination frequencies to cM: the recombination frequency in percent is the same as the map distance in cM. For example, if two loci have a recombination frequency of 25% they are said to be 25cM apart on a chromosome (Figure 7.9). Note that the map distance of two loci alone does not tell us anything about the orientation of these loci relative to other features on the chromosome.

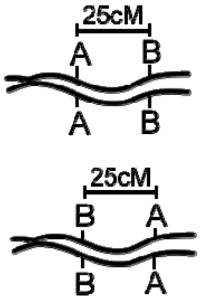


Figure 7.9 Two genetic maps consistent with a recombination frequency of 25% between A and B.

Map distances are always calculated for one pair of loci at a time. However, by combining the results of multiple calculations, a genetic map of many loci on a chromosome can be produced (Figure 7.10). A genetic map shows the map distance, in cM, that separates any two loci, and the position of these loci relative to all other mapped loci. The map distance corresponds roughly to the physical distance, i.e. the amount of DNA between two loci. For example, in Arabidopsis, 1.0 cM corresponds roughly to 150,000bp and contains approximately 5,000 genes. The exact number of DNA bases in a cM depends on the organism, and on the particular position in the chromosome; some parts of chromosomes ("hot spots") have slightly higher rates of recombination than others.

When a novel gene is identified by mutation or polymorphism, its approximate position on a chromosome can be determined by crossing it with previously mapped genes, and then calculating the recombination frequency. If the novel gene and the previously mapped genes show complete or partial linkage, the recombination frequency will indicate the approximate position of the novel gene within the genetic map. This information is useful in isolating (i.e. cloning) the specific fragment of DNA that encodes the novel gene, through a process called map-based cloning. Genetic maps are also useful in breeding crops and animals, in studying evolutionary relationships between species, and in determining the causes and individual susceptibility of some human diseases.

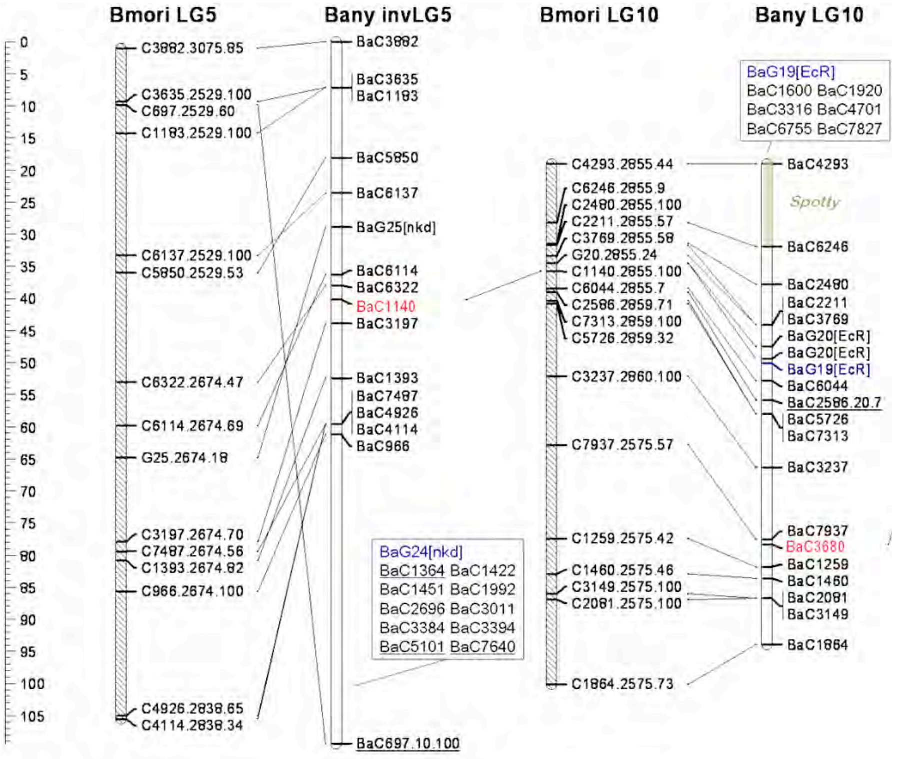


Figure 7.10 Genetic maps for regions of two chromosomes from two species of the moth, *Bombyx*. The scale at left shows distance in cM, and the position of various loci is indicated on each chromosome. Diagonal lines connecting loci on different chromosomes show the position of corresponding loci in different species

Genetic maps are useful tools, but genetic map distances are only an approximation of the actual physical distance between loci. The correlation between recombination frequency and chromosomal distance is more accurate for short distances than long distances. Observed recombination frequencies between two relatively distant markers tend to underestimate the actual number of crossovers that occurred. This is because as the distance between loci increases, so also increases the possibility of having two crossovers occur between the loci. This is a problem for geneticists, because with respect to the loci being studied, these double-crossovers produce gametes with the same genotypes as if no recombination events had occurred (Figure 7.11). Researchers will sometimes use specific mathematical functions to adjust large recombination frequencies to account for the possibility of multiple crossovers.

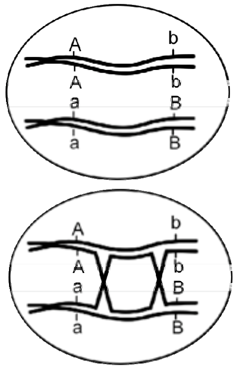


Figure 7.11 A double crossover between two loci (bottom) will produce gametes with parental genotypes (with respect to these loci).

MAPPING WITH THREE-POINT CROSSES

A particularly efficient method of mapping genes is the **three-point cross**, which allows the order and distance between three potentially linked genes to be determined in a single experiment (Figure 7.12). The basic strategy is the same as for the dihybrid mapping experiment described above; pure breeding lines with contrasting genotypes are crossed to produce an individual heterozygous at three loci (a trihybrid), which is then testcrossed to determine the recombination frequency between each pair of genes.

One useful feature of the three-point cross is that the order of the loci relative to each other can usually be determined by a simple visual inspection of the F_2 segregation data. If the genes are linked, there will often be two phenotypic classes that are much more infrequent than any of the others. In these cases, the rare phenotypic classes are usually those that arose from two crossover events, in which the locus in the middle is flanked by a crossover on either side of it. Thus, among the two rarest recombinant phenotypic classes, the one allele that differs from the other two alleles relative to the parental genotypes likely represents the locus that is in the middle of the other two loci. For example, based on the phenotypes of the pure-breeding parents in Figure 7.12, the parental genotypes are *aBC* and *AbC* (remember the order of the loci is unknown, and it is not necessarily the alphabetical order in which we wrote the genotypes). Because we can deduce from the outcome of the testcross (Table 7.2) that the rarest genotypes were *abC* and *ABc*, we can conclude that locus *A* that is most likely located between the other two loci, since it would require a recombination event between both *A* and *B* and between *A* and *C* in order to generate these gametes. Thus, the order of loci is *BAC* (which is equivalent to *CAB*).

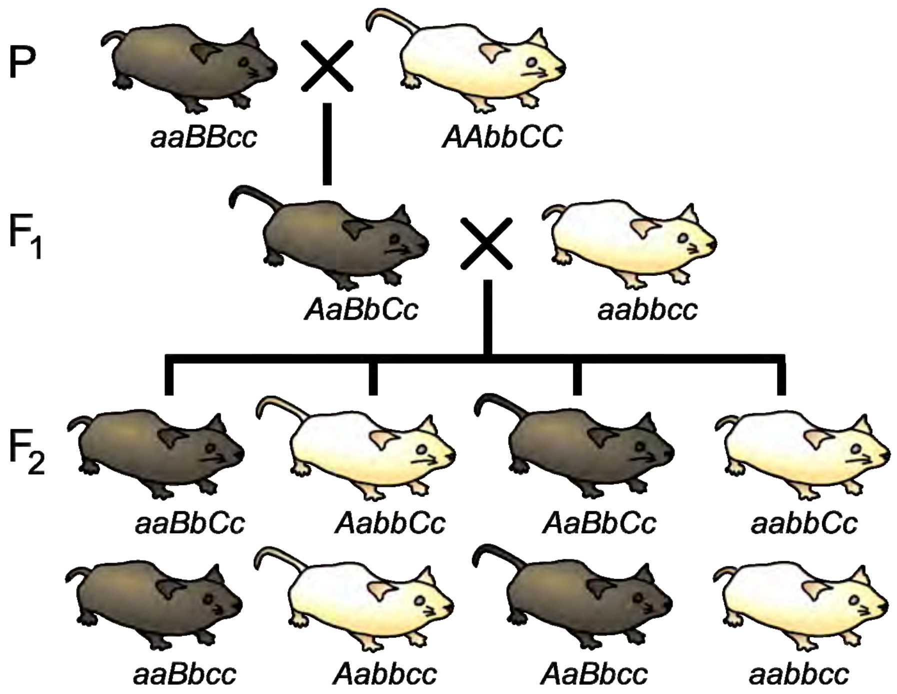


Figure 7.12 A three point cross for loci affecting tail length, fur color, and whisker length.

tail phenotype	fur phenotype	whisker phenotype	number of progeny	gamete from trihybrid	genotype of F_2 from test cross	loci A, B	loci A, C	loci B, C
short	brown	long	5	*aBC*	*aaBbCc*	P	R	R
long	white	long	38	*AbC*	*AabbCc*	P	P	P
short	white	long	1	*abC*	*aabbCc*	R	R	P
long	brown	long	16	*ABC*	*AaBbCc*	R	P	R
short	brown	short	42	*aBc*	*aaBbcc*	P	P	P
long	white	short	5	*Abc*	*Aabbcc*	P	R	R
short	white	short	12	*abc*	*aabbcc*	R	P	R
long	brown	short	1	*ABc*	*AaBbcc*	R	R	P

Table 7.2 An example of data that might be obtained from the F_2 generation of the three-point cross shown in Figure 7.12. The rarest phenotypic classes correspond to double recombinant gametes *ABc* and *abC*. Each phenotypic class and the gamete from the trihybrid that produced it can also be classified as parental (P) or recombinant (R) with respect to each pair of loci (A,B), (A,C), (B,C) analyzed in the experiment.

Recombination frequencies may be calculated for each pair of loci in the three-point cross as we did before for one pair of loci in our dihybrid (Figure 7. 8)

$$\text{loci A,B R.F.} = \frac{1+16+12+1}{120} = 25\%$$

$$\text{loci A,C R.F.} = \frac{1+5+1+5}{120} = 10\%$$

$$\text{loci B,C R.F.} = \frac{5+16+12+5}{120} = 32\%$$

(not corrected for double crossovers)

However, note that in the three point cross, the sum of the distances between A-B and A-C (35%) is less than the distance calculated for B-C (32%)(Figure 7.13). this is because of double crossovers between B and C, which were undetected when we considered only pairwise data for B and C. We can easily account for some of these double crossovers, and include them in calculating the map distance between B and C, as follows. We already deduced that the map order must be *BAC* (or *CAB*), based on the genotypes of the two rarest phenotypic classes in Table 7.2. However, these double recombinants, *ABc* and *abC,* were not included in our calculations of recombination frequency between loci *B* and *C*. If we included these double recombinant classes (multiplied by 2, since they each represent two recombination events), the calculation of recombination frequency between B and C is as follows, and the result is now more consistent with the sum of map distances between A-B and A-C.

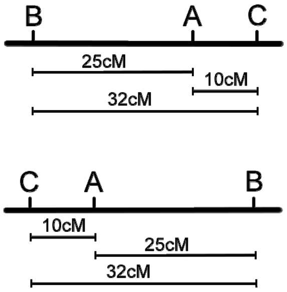


Figure 7.13 Equivalent maps based on the data in Table 7.2 (without correction for double crossovers).

$$\text{loci B,C R.F.} = \frac{5+16+12+5+2(1)+2(1)}{120} = 35\%$$

(corrected for double recombinants)

Thus, the three point cross was useful for determining the order of three loci relative to each other, calculating map distances between the loci, and detecting some of the double crossover events that would otherwise lead to an underestimation of map distance. However, it is possible that other, double crossovers events remain undetected, for example double crossovers between loci A,B or between loci A,C. Geneticists have developed a variety of mathematical procedures to try to correct for things like double crossovers during large-scale mapping experiments.

SUMMARY

- Recombination is defined as any process that results in gametes with combinations of alleles that were not present in the gametes of a previous generation.

- The recombination frequency between any two loci depends on their relative chromosomal locations.

- Unlinked loci show a maximum 50% recombination frequency.

- Loci that are close together on a chromosome are linked and tend to segregate with the same combinations of alleles that were present in either parent.

- Crossovers are a normal part of most meioses, and allow for recombination between linked loci.

- Measuring recombination frequency is easiest when starting with pure-breeding lines with two alleles for each locus, and with suitable lines for test crossing.

- Because recombination frequency is proportional to the distance between loci, recombination frequencies can be used to create genetic maps.

- Recombination frequencies tend to underestimate map distances, especially over long distances, since double crossovers may be genetically indistinguishable from non-recombinants.

- Three-point crosses can be used to determine the order and map distance between of three loci, and can correct for some of the double crossovers between the two outer loci.

KEY TERMS

linkage
recombination
crossover
recombinant
parental
linked
unlinked
complete linkage
recombination frequency
synapsis
cis
coupling
trans
repulsion
cM
linkage map
genetic map

STUDY QUESTIONS

7.1 Compare recombination and crossover. How are these similar? How are they different?

7.2 Explain why it usually necessary to start with pure-breeding lines when measuring genetic linkage by the methods presented in this chapter.

7.3 If you knew that a locus that affected earlobe shape was tightly linked to a locus that affected susceptibility to cardiovascular disease human, under what circumstances would this information be clinically useful?

7.4 In a previous chapter, we said a 9:3:3:1 phenotypic ratio was expected among the progeny of a dihybrid cross, in absence of gene interaction.

a) What does this ratio assume about the linkage between the two loci in the dihybrid cross?

b) What ratio would be expected if the loci were completely linked? Be sure to consider every possible configuration of alleles in the dihybrids.

7.5 Given a dihybrid with the genotype *CcEe*:

a) If the alleles are in cis-configuration, what will be the genotypes of the parental and recombinant progeny from a test-cross?

b) If the alleles are in trans-configuration, what will be the genotypes of the parental and recombinant progeny from a test-cross?

7.6 Imagine the white flowers are recessive to purple flowers, and yellow seeds are recessive to green seeds. If a green-seeded, purple-flowered dihybrid is testcrossed, and half of the progeny have yellow seeds, what can you conclude about linkage between these loci? What do you need to know about the parents of the dihybrid in this case?

7.7 In corn (i.e. maize, a diploid species), imagine that alleles for resistance to a particular pathogen are recessive and are linked to a locus that affects tassel length (short tassels are recessive to long tassels). Design a series of crosses to determine the map distance between these two loci. You can start with any genotypes you want, but be sure to specify the phenotypes of individuals at each stage of the process. Outline the crosses similar to what is shown in Figure 7.8, and specify which progeny will be considered recombinant. You do not need to calculate recombination frequency.

7.8 In a mutant screen in Drosophila, you identified a gene related to memory, as evidenced by the inability of recessive homozygotes to learn to associate a particular scent with the availability of food. Given another line of flies with an autosomal mutation that produces orange eyes, design a series of crosses to determine the map distance between these two loci. Outline the crosses similar to what is shown in Figure 7.8, and specify which progeny will be considered recombinant. You do not need to calculate recombination frequency.

7.9 Image that methionine heterotrophy, chlorosis (loss of chlorophyll), and absence of leaf hairs (trichomes) are each caused by recessive mutations at three different loci in Arabidopsis. Given a triple mutant, and assuming the loci are on the same chromosome, explain how you would determine the order of the loci relative to each other.

7.10 If the progeny of the cross *aaBB* x *AAbb* is testcrossed, and the following genotypes are observed among the progeny of the testcross, what is the frequency of recombination between these loci?

AaBb	135
Aabb	430
aaBb	390
aabb	120

7.11 Three loci are linked in the order B-C-A. If the A-B map distance is 1cM, and the B-C map distance is 0.6cM, given the lines *AaBbCc* and *aabbcc*, what will be the frequency of *Aabb* genotypes among their progeny if one of the parents of the dihybrid had the genotypes *AABBCC*?

7.12 Genes for body color (B black dominant to b yellow) and wing shape (C straight dominant to c curved) are located on the same chromosome in flies. If single mutants for each of these traits are crossed (i.e. a yellow fly crossed to a curved-wing fly), and their progeny is testcrossed, the following phenotypic ratios are observed among their progeny.

black, straight	17
yellow, curved	12
black, curved	337
yellow, straight	364

a) Calculate the map distance between B and C.
b) Why are the frequencies of the two smallest classes not exactly the same?

7.13 Given the map distance you calculated between B-C in question 12, if you crossed a double mutant (i.e. yellow body and curved wing) with a wild-type fly, and testcrossed the progeny, what phenotypes in what proportions would you expect to observe among the F_2 generation?

7.14 In a three-point cross, individuals *AAbbcc* and *aaBBCC* are crossed, and their F_1 progeny is testcrossed. Answer the following questions based on these F_2 frequency data.

aaBbCc	480
AaBbcc	15
AaBbCc	10
aaBbcc	1
aabbCc	13
Aabbcc	472
AabbCc	1
aabbcc	8

a) Without calculating recombination frequencies, determine the relative order of these genes.
b) Calculate pair-wise recombination frequencies (without considering double cross overs) and produce a genetic map.
c) Recalculate recombination frequencies accounting for double recombinants.

7.15 Wild-type mice have brown fur and short tails. Loss of function of a particular gene produces white fur, while loss of function of another gene produces long tails, and loss of function at a third locus produces agitated behaviour. Each of these loss of function alleles is recessive. If a wild-type mouse is crossed with a triple mutant, and their F_1 progeny is test-crossed, the following recombination frequencies are observed among their progeny. Produce a genetic map for these loci

fur	tail	behaviour	
white	short	normal	16
brown	short	agitated	0
brown	short	normal	955
white	short	agitated	36
white	long	normal	0
brown	long	agitated	14
brown	long	normal	46
white	long	agitated	933

Chapter 8 TECHNIQUES OF MOLECULAR BIOLOGY

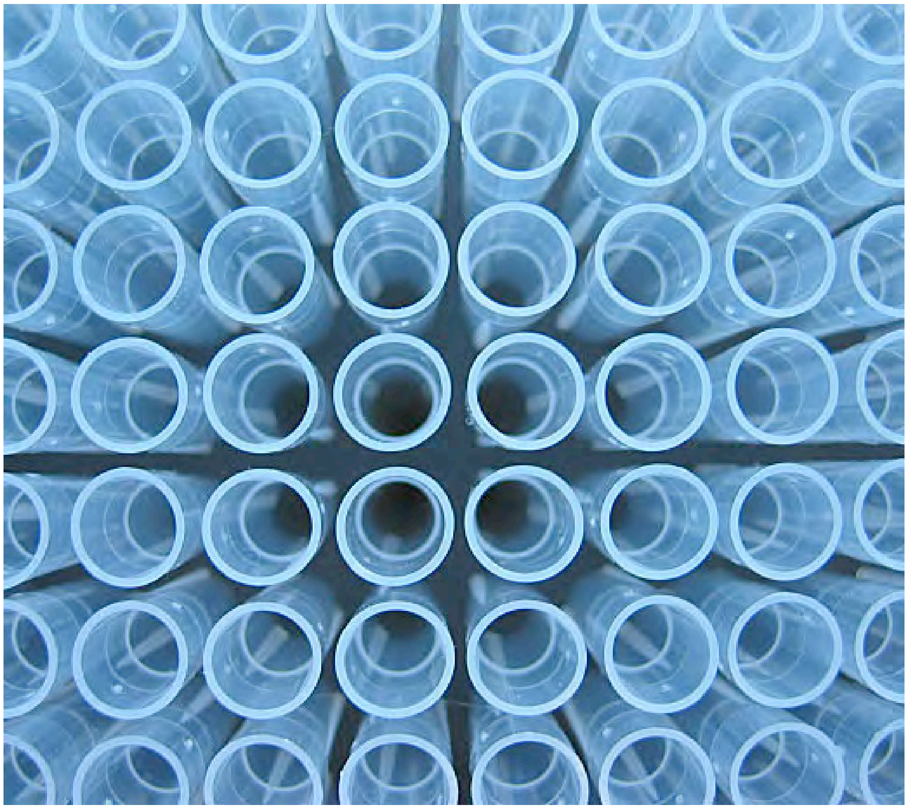


Figure 8.1 Disposable tips for a pipette are used to distribute microliter volumes of liquid in molecular biology.

Genetics is the study of the inheritance and variation of biological traits. We have previously noted that it is possible to conduct genetic research without directly studying DNA. Indeed some of the greatest geneticists had no special knowledge of DNA at all, but relied instead on analysis of phenotypes and their ratios in carefully designed crosses. Today, classical genetics is often complemented by **molecular biology**, which is the study of DNA and other **macromolecules** that have been isolated from an organism. Usually, molecular biology experiments involve some combination of techniques to isolate, then analyze the DNA or RNA of a particular gene. In some cases, the DNA may be subsequently manipulated by mutation or by recombination with other DNA fragments. Techniques of molecular biology have wide application in many fields of biology, as well as forensics, biotechnology, and medicine.

ISOLATING GENOMIC DNA

DNA purification strategies rely on the chemical properties of DNA that distinguish it from other molecules in the cell, namely that it is a very long, negatively charged molecule. To extract purified DNA from a tissue sample, cells are broken open by grinding or **lysing** in a solution that contains chemicals that protect the DNA while disrupting other components of the cell (Figure 8.2). These chemicals may include **detergents**, which dissolve membranes and denature proteins. A cation such as Na^+ helps to stabilize the negatively charged DNA and separate it from proteins including histones. A **chelating agent**, such as **EDTA**, is added to protect DNA by sequestering Mg^{2+}, which can otherwise serve as a necessary co-factor for **nucleases** (enzymes that digest DNA). As a result, free, double-stranded DNA molecules are released from the chromatin into the extraction buffer, which also contains proteins and all other cellular components.

The free DNA molecules are subsequently isolated by one of several methods, such as adjusting the salt concentration so that proteins precipitate. The supernatant, which contains DNA and other, smaller metabolites, is then mixed with ethanol, which causes the DNA to precipitate. A small **pellet** of DNA can be collected by centrifugation, and after removal of the ethanol, the DNA pellet can be dissolved in water (usually with a small amount of EDTA and a pH buffer) for the use in other reactions. Note that this process has purified all of the DNA from a tissue sample; if we want to further isolate a specific gene or DNA fragment, we must use additional techniques, such as PCR.

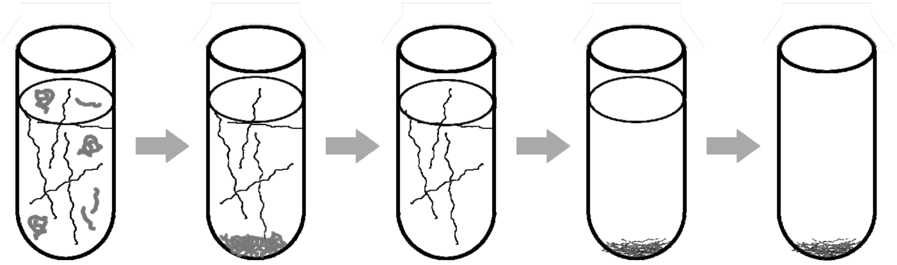


Figure 8.2 Extraction of DNA from a mixture of solubilized cellular components by successive precipitations. Proteins are precipitated, then DNA (in the supernatant) is precipitated in ethanol, leaving a pellet of DNA.

ISOLATING OR DETECTING A SPECIFIC SEQUENCE BY PCR

The **Polymerase Chain Reaction (PCR)** is a method of DNA replication that is performed in a test tube (i.e. in vitro). Here “polymerase” refers to a DNA polymerase enzyme extracted from bacteria, and “chain reaction” refers to the ability of this technique produce millions of copies of a DNA molecule, by using each newly replicated double helix as a template to synthesize two new DNA double helices. PCR is therefore a very efficient method of amplifying DNA.

Besides its ability to make lots of DNA, there is a second characteristic of PCR that makes it extremely useful. Recall that most DNA polymerases can only add nucleotides to the end of an existing strand of DNA, and therefore require a **primer** to begin the process of replication. For PCR, chemically synthesized primers of about 16 nucleotides are used. In an ideal PCR, primers only hybridize to their exact complementary sequence on the template strand (Figure 8.3).

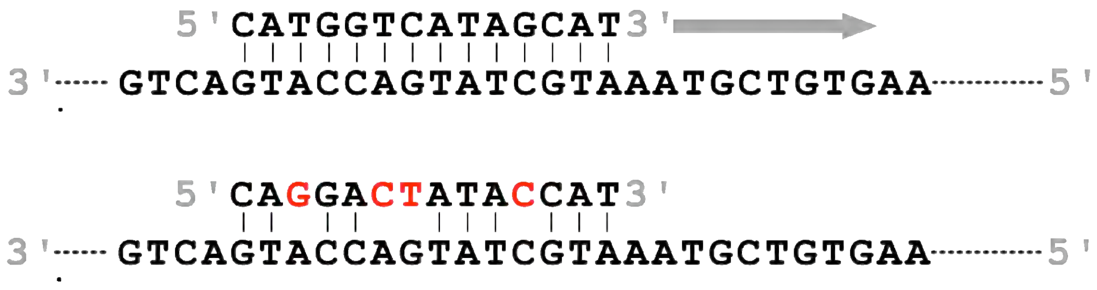

5 ' CATGGTCATAGCAT 3 '
3 '------ GTCAGTACCAGTATCGTAAATGCTGTGAA------------ 5 '
5 ' CAGGACTATACCAT 3 '
3 '------ GTCAGTACCAGTATCGTAAATGCTGTGAA------------ 5 '

Figure 8.3 The primer-template duplex at the top part of the figure is perfectly matched, and will be stable at a higher temperature than the duplex in the bottom part of the figure, which contains many mismatches and therefore fewer hydrogen bonds. If the annealing temperature is sufficiently high, only the perfectly matched primer will be able to initiate extension (grey arrow) from this site on the template.

The experimenter can therefore control exactly what region of a DNA template is amplified by controlling the sequence of the primers used in the reaction.

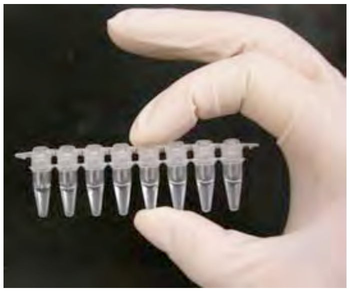


Figure 8.4 A strip of PCR tubes

To conduct a PCR, an experimenter combines in a small, thin-walled tube (Figure 8.4), all of the necessary components for DNA replication, including DNA polymerase and solutions containing nucleotides (dATP, dCTP, dGTP, dTTP), a DNA template, DNA primers, a pH buffer, and ions (e.g. Mg^{2+}) required by the polymerase. Successful PCR reactions have been conducted using a single DNA molecule as a template, but in practice, most PCR reactions contain many thousands of template molecules. The template DNA (e.g. total genomic DNA) has usually already been purified from cells or tissues using the techniques described above. However, in some situations it is possible to put whole cells directly in a PCR reaction for use as a template.

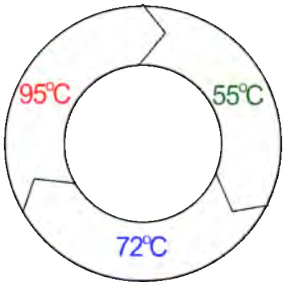
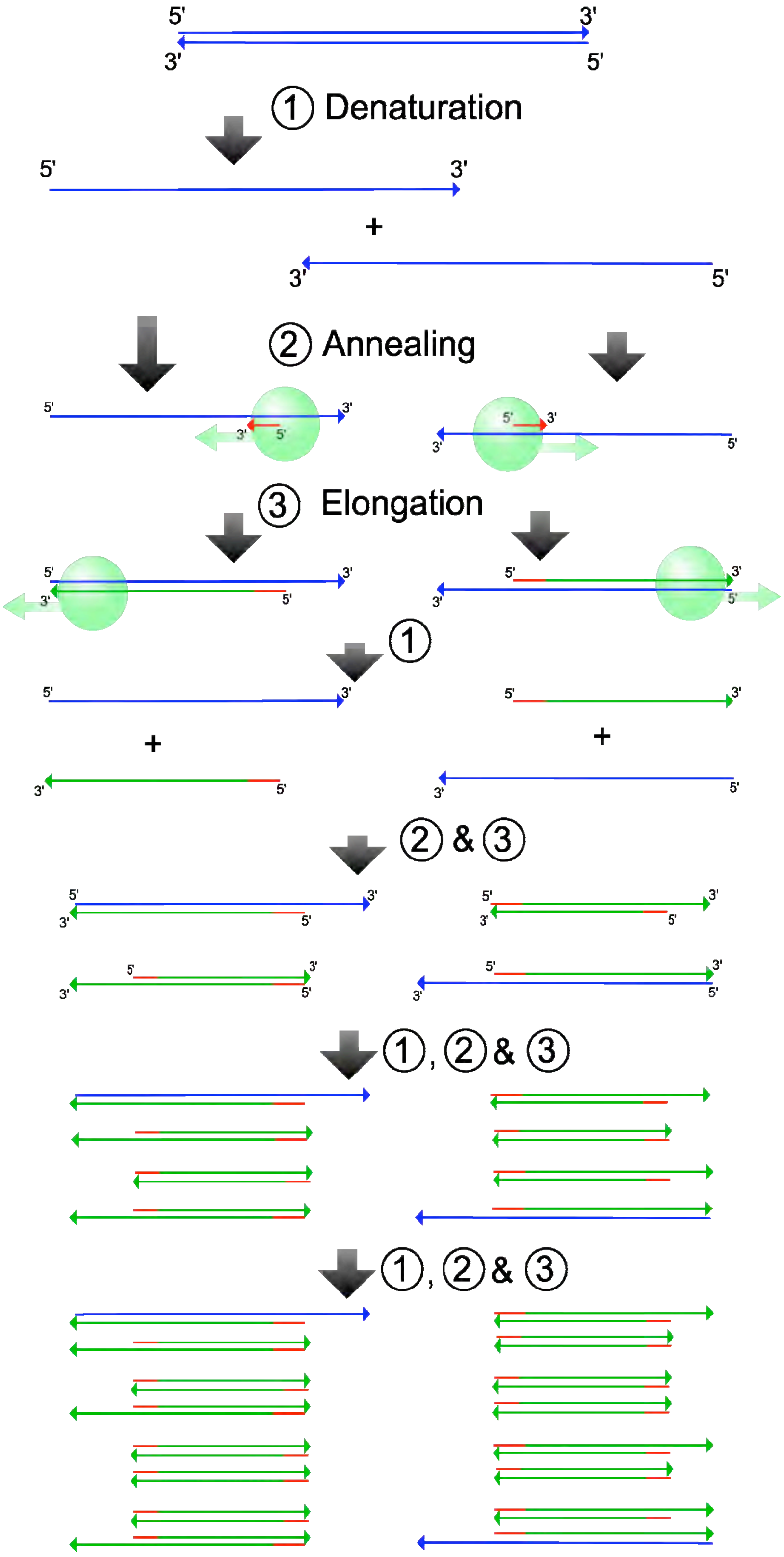


Figure 8.5 Example of a thermalcycle, in which the annealing temperature is 55°C.

An essential aspect of PCR is **thermalcycling**, meaning the exposure of the reaction to a series of precisely defined temperatures (Figure 8.5). The reaction mixture is first heated to 95°C. This causes the hydrogen bonds between the strands of the template DNA molecules to melt, or **denature**. This produces two single-stranded DNA molecules from each double helix (Figure 8.6). In the next step (**annealing**), the mixture is cooled to 45-65°C (the exact temperature depends on the primer sequence used and the objectives of the experiment). This allows the formation of double stranded helices between complementary DNA molecules, including the annealing of primers to the template. In the final step (**extension**) the mixture is heated to 72°C. This is the temperature at which the particular DNA polymerase used in PCR is most active. During extension, the new DNA strand is synthesized, starting from the 3' end of the primer, along the length of the template strand. The entire PCR process is very quick, with each temperature phase usually lasting 30 seconds or less. Each cycle of three temperatures (denaturation, annealing, extension) is usually repeated about 30 times, amplifying the target region approximately

Figure 8.6 PCR with the three phases of the thermalcycle numbered. The template strand (blue) is replicated from primers (red), with newly synthesized strands in green. The green strands flanked by two primer binding sites will increase in abundance exponentially through successive PCR cycles.



2^{30}-fold. Notice from the figure that most of the newly synthesized strands in PCR begin and end with sequences either identical to or complementary to the primer sequences; although a few strands are longer than this, they are in such a small minority that they can almost always be ignored. The earliest PCR reactions used a polymerase from *E. coli*. Because the high temperature of the denaturation step destroyed the enzyme, new polymerase had to be added after each cycle. To overcome this, researchers identified thermostable DNA polymerases such as **Taq DNA pol**, from *Thermus acquaticus*, a thermophilic bacterium that lives in hot springs. Taq cannot usually amplify fragments longer than about 3kbp, but under some specialized conditions, PCR can amplify fragments up to approximately 10kbp.

After completion of the thermalcycling, part of the PCR reaction is usually loaded onto an **electrophoretic gel** (described below) to determine whether a DNA fragment of the expected length was amplified or not. Usually, the template DNA will be so dilute that it will not show up on the gel, so the presence of a sharp band of the expected length indicates that PCR was able to amplify its target. If the purpose of the PCR was to test for the presence of a particular template sequence, this is the end of the experiment. Otherwise, the remaining PCR product can be used as starting material for a variety of other techniques such as sequencing or cloning.

AN APPLICATION OF PCR: THE STARLINK AFFAIR

PCR is sensitive (meaning it can amplify very small starting amounts of DNA), and specific (meaning it can amplify only the target sequence from a mixture of many DNA sequences). This made PCR the perfect tool to test whether genetically modified corn was present in consumer products on supermarket shelves. Although the majority of corn in the United States is genetically modified, and contains genes that government regulators have approved for human consumption, an environmental group became suspicious that a different strain of genetically modified corn, which had been approved for use only as animal feed, had been mixed in with corn used in production of things like taco shells. To prove this, the group purchased taco shells from stores in the Washington DC area, extracted DNA from the taco shells and used it as a template in a PCR reaction with primers specific to the unauthorized gene. Their suspicions were confirmed when they ran this PCR product on an agarose gel and saw a band of expected size. The company that sold both the approved and non-approved transgenic seed to farmers had to pay for the destruction of large amounts of corn, and was the target of a class action law-suit by angry consumers who claimed they had been made sick by the taco shells. No legitimate cases of harm were ever proven, and the plaintiffs were awarded $9 million, of which $3 million went to the legal fees, and the remainder of the judgment went to the consumers in the form of coupons for taco shells. The affair destroyed the company, and exposed a weakness in the way the genetically modified crops were handled in the United States at the time.

CUTTING AND PASTING DNA: RESTRICTION AND LIGATION

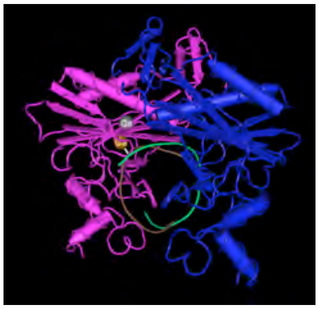


Figure 8.7 An *EcoRI* dimer (blue, purple) sits like a saddle on a double helix of DNA (one strand is green, one is brown). This image is looking down the center of the helix.

Many bacteria have enzymes that recognize specific DNA sequences (usually 4 or 6 nucleotides) and then cut the double stranded DNA helix at this sequence (Figure 8.7). These enzymes are called site-specific **restriction endonucleases**, or more simply “restriction enzymes”, and they naturally function as part of bacterial defenses against viruses and other sources of foreign DNA. To cut DNA at known locations, researchers use restriction enzymes that have been purified from various bacterial species, and which can be purchased from commercial sources. These enzymes are usually named after the bacterium from which they were first isolated. For example, ***EcoRI*** and *EcoRV* are both enzymes from *E. coli*. *EcoRI* cuts double stranded DNA at the sequence GAATTC, but note that this enzyme, like many others, does not cut in exactly the middle of the restriction sequence (Figure 8.8). The ends of a molecule cut by *EcoRI* have an overhanging region of single stranded DNA, and so are sometimes called **sticky-ends**. On the other hand, *EcoRV* is an example of an enzyme that cuts both strands in exactly the middle of its recognition sequence, producing what are called **blunt-ends**, which lack overhangs.

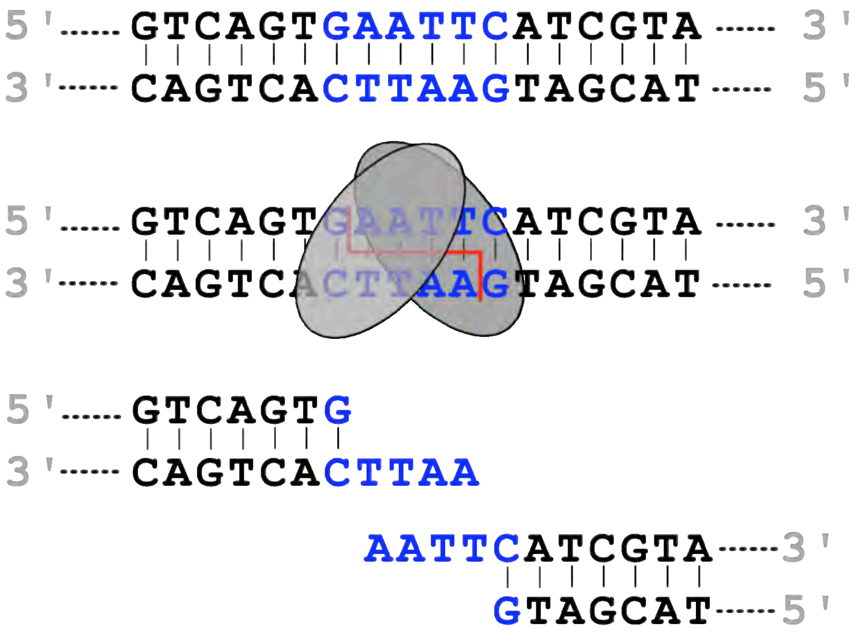

5 '------ GTCAGTGAATTCATCGTA ------ 3 '
3 '------ CAGTCACTTAAGTAGCAT ------ 5 '
5 '------ GTCAGTGAATTCATCGTA ------ 3 '
3 '------ CAGTCACTTAAGTAGCAT ------ 5 '
5 '------ GTCAGTG
3 '------ CAGTCACTTAA
AATTCATCGTA------3 '
GTAGCAT------5 '

Figure 8.8 The recognition sequence for *EcoRI* (blue) is cleaved by the enzyme (grey). This particular enzyme cuts DNA at a position offset from the center of the restriction site. This creates an overhanging, sticky-end

The process of **ligation** occurs when double-stranded DNA molecules are covalently joined, end-to-end, through the action of an enzyme called **ligase**. Ligation is therefore central to the production of recombinant DNA, including the insertion of a cloned DNA fragment into a plasmid. Sticky-ended molecules with complementary overhanging sequences are said to have **compatible ends.** Likewise, two blunt-ended sequences are also considered compatible, although they may not ligate together as efficiently as sticky-ends. On the other hand, sticky-ended molecules with non-complementary sequences cannot be ligated together.

CLONING DNA: PLASMID VECTORS

Many bacteria contain extra-chromosomal DNA elements called **plasmids**. These are usually small, circular, double stranded molecules that replicate independently of the chromosome and can be present in high copy numbers within a cell. Plasmids can be transferred between individuals during bacterial mating and are sometimes even transferred between different species. Plasmids are particularly important in medicine because they often carry some genes for pathogenicity and drug-resistance.

To insert a DNA fragment into a plasmid, both the fragment and the plasmid are cut using a restriction enzyme that produces compatible ends (Figure 8.9). Given the large number of restriction enzymes that are currently available, it is usually not too difficult to find an enzyme for which corresponding recognition sequences are present in both the plasmid and the DNA fragment, particularly because most plasmid vectors used in molecular biology have been engineered to contain recognition sites for a large number of endonucleases.

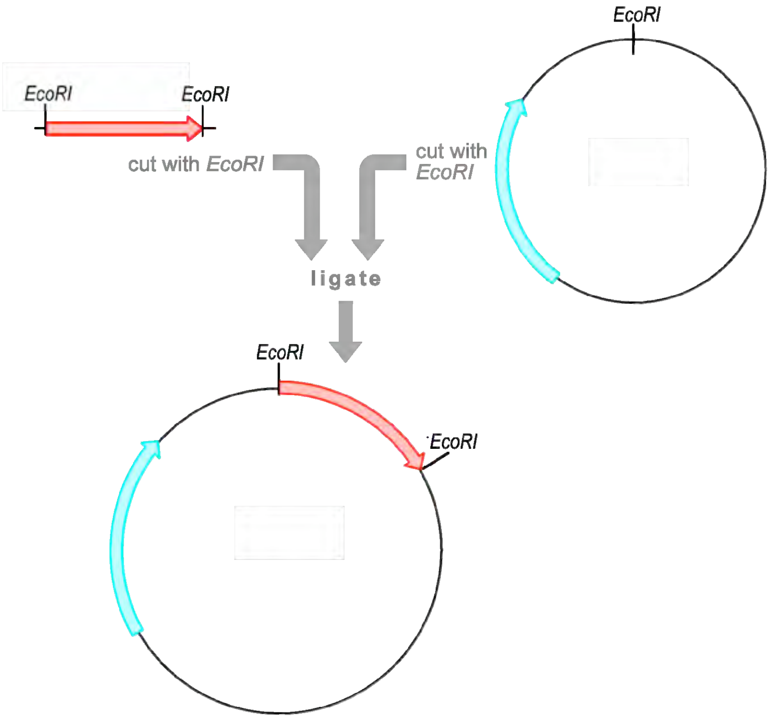


Figure 8.9 Cloning of a DNA fragment (red) into a plasmid vector. The vector already contains a selectable marker gene (blue)

After restriction digestion, the desired fragments may be further purified before they are mixed together with ligase. Following a short incubation, the newly ligated plasmids, containing the gene of interest are **transformed** into *E. coli*. Transformation is accomplished by mixing the ligated DNA with *E. coli* that has been specially prepared (i.e. made **competent**) to uptake DNA when exposed to compounds such as $CaCl_2$ or to electrical fields (**electroporation**). Because only a small fraction of cells that are mixed with DNA will actually be transformed, a **selectable marker**, such as a gene for antibiotic resistance, is usually also present on the plasmid. After combining DNA with competent cells, bacteria are therefore spread on a plate containing an appropriate antibiotic so that only cells that have actually incorporated the plasmid will grow and form colonies for further study.

Molecular biologists use plasmids as **vectors** to contain, amplify, transfer, and sometimes express genes of interest. Often, the first step in a molecular biology experiment is to **clone** (i.e. copy) a gene into a plasmid, then introduce transform this recombinant plasmid back into bacteria so that essentially unlimited copies of the gene (and the plasmid that carries it) can be made as the bacteria reproduce. This is a practical necessity for further manipulations of the DNA, since most techniques of molecular biology are not sensitive enough to work with just a single molecule at a time. Many molecular cloning and recombination experiments are therefore iterative processes in which:

- a DNA fragment (usually isolated by PCR and/or restriction digestion) is cloned into a plasmid cut with a compatible restriction enzyme
- the recombinant plasmid is transformed into bacteria
- the bacteria are allowed to divide, usually in liquid culture
- a large quantity of the recombinant plasmid DNA is isolated from the bacterial culture
- further manipulations (such as site directed mutagenesis or the introduction of another piece of DNA) are conducted on the recombinant plasmid
- the modified plasmid is again transformed into bacteria, prior to further manipulations, or for expression

AN APPLICATION OF MOLECULAR CLONING: PRODUCTION OF RECOMBINANT INSULIN

Purified insulin is critical to the treatment of diabetes. Historically, insulin for clinical use was isolated from human cadavers or from slaughtered animals such as pigs. Human-derived insulin generally had better pharmacological properties, but was in limited supply and carried risks of disease transmission. By cloning the human insulin gene and expressing it in *E. coli*, large quantities of insulin identical to the human hormone could be produced in fermentors, safely and efficiently. Production of recombinant insulin also allows specialized variants of the protein to be produced: for example, by changing a few amino acids, longer-acting forms of the hormone can be made. The active insulin hormone contains two peptide fragments of 21 and 30 amino acids, respectively. Today, essentially all insulin is produced from recombinant sources (Figure 8.10), i.e. human genes and their derivatives expressed in bacteria or yeast.

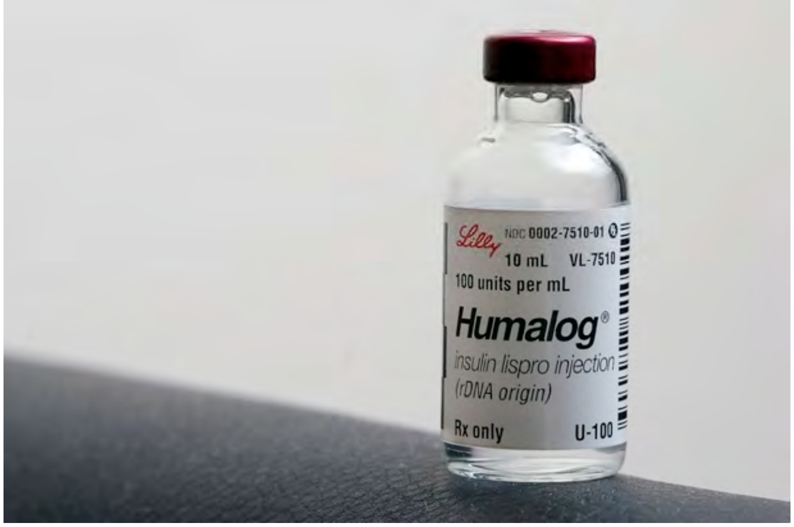


Figure 8.10 A vial of insulin. Note that the label lists the origin as “rDNA”, which stands for recombinant DNA.

DNA ANALYSIS: GEL ELECTROPHORESIS

A solution of DNA is colorless, and except for being viscous at high concentrations, is visually indistinguishable from water. Therefore, techniques such as **gel electrophoresis** have been developed to detect and analyze DNA (Figure 8.11). To start, a solution of DNA is deposited at one end of a gel slab. This gel is made from polymers such as **agarose**, which is a polysaccharide isolated from certain seaweed. The DNA is then pulled through the gel by an electrical current, with DNA molecules moving toward the positive electrode (Figure 8.12).

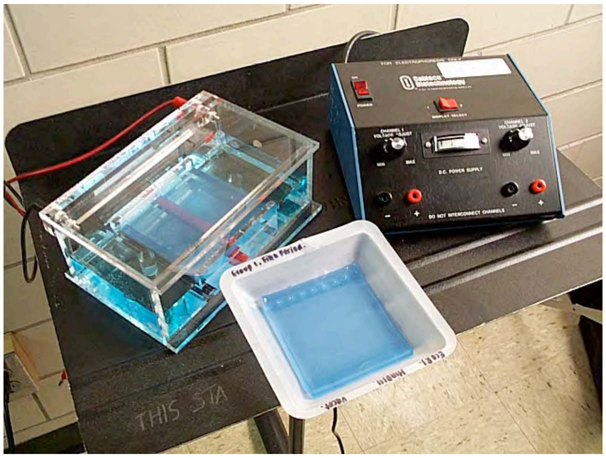


Figure 8.11 Apparatus for agarose gel electrophoresis. A waterproof tank is used to pass current through a slab gel, which is submerged in a buffer in the tank. The current is supplied by an adjustable power supply. A gel (stained blue by a dye sometimes used when loading DNA on the gel) sits in a tray, awaiting further analysis.

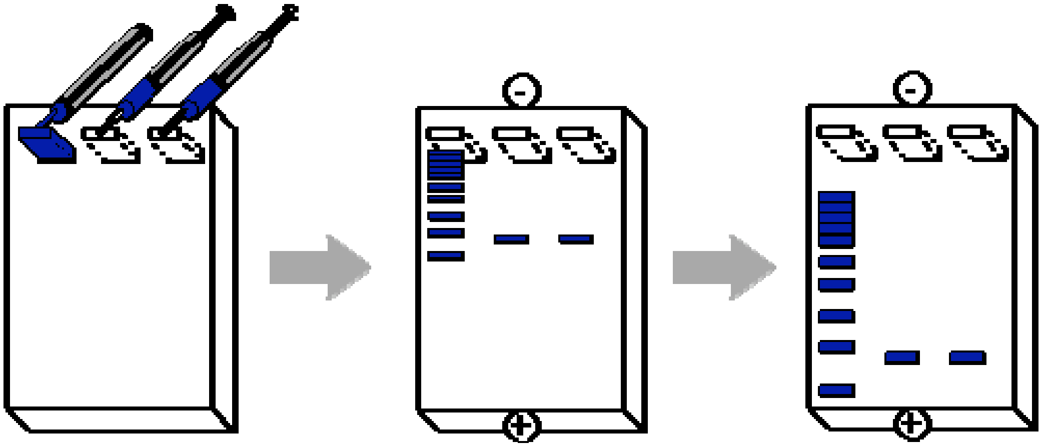


Figure 8.12 Agarose gel electrophoresis. DNA is loaded into wells at the top of a gel. A current is passed through the gel, pulling DNA towards the positively charged electrode. The DNA fragments are separated by size, with smaller fragments moving fastest towards the electrode.

As it migrates, each strand of DNA threads its way through the pores that form between the polymers of the gel. Because shorter fragments move through these pores faster than longer fragments, gel electrophoresis separates molecules based on their size, with smaller DNA fragments moving faster than long ones. This process also concentrates DNA fragments of a similar size in one location in each gel, called a **band**. Concentrating DNA fragments of similar size in this way make it easy to visualize DNA after staining the DNA with a fluorescent dye such as **ethidium bromide** (Figure 8.13). By electrophoresing a mixture of DNA fragments of known size in adjacent lanes on the same gel, the length of an uncharacterized DNA fragment can be estimated. Bands can also be cut out of the gel, and the DNA in the band can be extracted and used in other types of reactions.

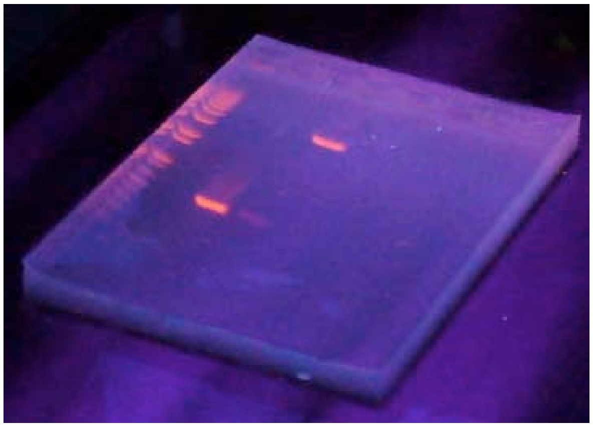


Figure 8.13 An agarose gel stained with ethidium bromide and illuminated by UV light. The stain associated with DNA is fluorescent.

DNA ANALYSIS: BLOTTING AND HYBRIDIZATION

DNA only forms distinct bands in gel electrophoresis if most of the DNA molecules are of the same size, such as following a PCR reaction, or restriction digestion of a plasmid. In other situations, such as after restriction digestion of chromosomal DNA, there will be so many molecules of so many different sizes, that the mixture will appear as a smear, rather than distinct bands. In this case, it is necessary to use additional techniques to detect the presence of a specific DNA sequence within the electrophoretic gel.

In a **Southern blot** (the technique is named after Ed Southern, its inventor), DNA that has been digested with restriction enzymes is separated by gel electrophoresis, and then a sheet of nylon or similar material is laid under the gel and the DNA is transferred to the nylon by drawing the liquid out of the gel, in a process called **blotting** (Figure 8.14). The blotted DNA is usually covalently attached to the nylon **membrane** by briefly exposing the blot to UV light. Transferring the DNA to the nylon is necessary be because the fragile gel would fall apart during the next steps in the process: **hybridization** and **washing**. For hybridization, a piece of DNA that is complementary in sequence to the target molecule is labeled using fluorescent or radioactive molecules, and then this labeled **probe** is flooded over the surface of the nylon membrane prior to denaturing the DNA on the membrane.

If the hybridization is performed properly, the probe DNA will form a stable helix only with those DNA molecules on the membrane that exactly match it, and the radioactive or fluorescent signal will appear in a distinct band after washing off the unbound probe. In this way, the presence of a particular DNA sequence within a mixture of DNA can be detected. The temperature of the hybridization and washing solutions is important to the **stringency** of the hybridization. At maximum stringency (higher temperature), probes will only hybridize with target sequences that are perfectly complementary and therefore form the maximum number of hydrogen bonds. At lower temperatures, probes will be able to hybridize to targets to which they do not match exactly.

Southern blotting is useful when trying to detect the presence of a DNA sequence within a mixture of DNA molecules. Southern blotting was invented before PCR, and although PCR has replaced blotting in some applications, Southern blots are still useful when detecting fragments larger than those normally amplified by PCR, and when trying to detect fragments that may be only distantly related to a known sequence. Applications of Southern blotting will be discussed further in the context of molecular markers in a subsequent chapter.

Following the development of the Southern blot, other types of blotting techniques were invented, including the northern blot (in which RNA is separated on a gel before probing), and the western blot (protein is separated on a gel before probing with an antibody).

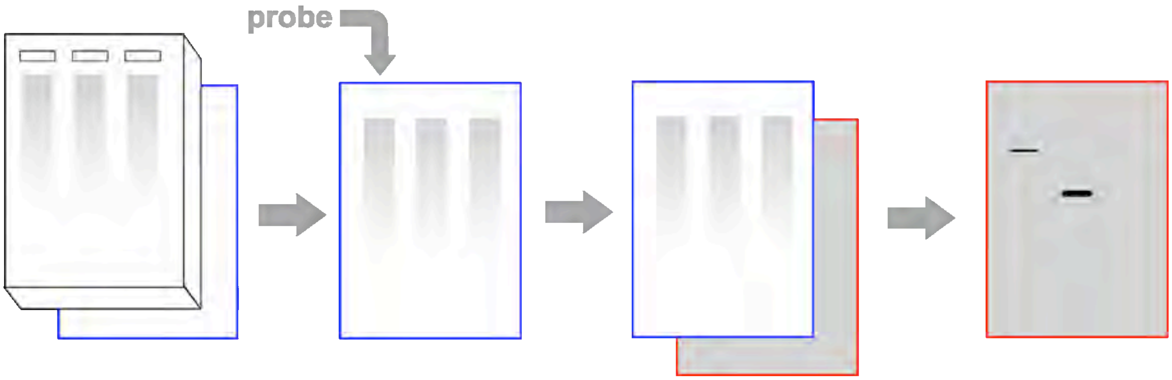


Figure 8.14 An example of Southern blotting. Genomic DNA that has been digested with a restriction enzyme is separated on an agarose gel, then the DNA is transferred from the gel to a nylon membrane (blue outline) by blotting. The DNA is immobilized on the membrane, then probed with a radioactively labeled DNA fragment that is complementary to a target sequence. After stringent washing, the blot is exposed to X-ray film to detect where the probe is bound. In this case, the probe bound to different-sized fragments in lanes 1 and 2, and no fragments in lane 3.

SUMMARY:

- Molecular biology involves the isolation and analysis of DNA and other macromolecules
- Isolation of total genomic DNA involves separating DNA from protein and other cellular components, for example by ethanol precipitation of DNA.
- PCR can be used as part of a sensitive method to detect the presence of a particular DNA sequence
- PCR can also be used as part of a method to isolate and prepare large quantities of a particular DNA sequence
- Restriction enzymes are natural endonucleases used in molecular biology to cut DNA sequences at specific sites.
- DNA fragments with compatible ends can be joined together through ligation. If the ligation produces a sequence not found in nature, the molecule is said to be recombinant.
- Transformation is the introduction of DNA (usually recombinant plasmids) into bacteria.
- Cloning of genes in E. coli is a common technique in molecular biology, since it allows large quantities of a DNA for gene to made, which allows further analysis or manipulation

- Cloning can also be used to produce useful proteins, such as insulin, in microbes.
- Southern blotting can be used to detect the presence of any sequence that matches a probe, within a mixture of DNA (such as total genomic DNA).
- The stringency of hybridization in blotting and in PCR is dependent on physical parameters such as temperature.

KEY TERMS:

macromolecules
lysis
detergent
chelating agent
EDTA
nuclease
pellet
PCR
primer
thermalcycle
denature
anneal
extend
Taq DNApol
electrophoretic gel
restriction
endonuclease
EcoRI
sticky end
blunt end
compatible end
ligation
ligase
plasmid
transformation
competent
electroporation
selectable marker
agarose
vector
band
ethidium bromide
Southern blot
membrane
hybridization
washing
probe
stringency
northern blot
western blot

STUDY QUESTIONS:

8.1 What information, and what reagents would you need to use PCR to detect HIV in a blood sample?

8.2 A 6000bp PCR fragment flanked by recognition sites for the *HindIII* restriction enzyme is cut with *HindIII* then ligated with a 3kb plasmid vector that has also been cut with *HindIII*. This recombinant plasmid is transformed into *E. coli* and a large preparation of plasmid is digested with HindIII.
a) When the product of the *HindIII* digestion is analyzed by gel electrophoresis, what will be the size of the bands observed?
b) What bands would be observed if the recombinant plasmid was cut with *EcoRI*, which has only one site, directly in the middle of the PCR fragment?
c) What bands would be observed if the recombinant plasmid was cut with both *EcoRI* and *HindIII* at the same time?

8.3 If you started with 10 molecules of double stranded DNA template, what is the maximum number of molecules you would you have after 10 PCR cycles?

8.4 What is present in a PCR tube at the end of a reaction? With this in mind, why do you usually only see a single, sharp band on a gel when a successful PCR reaction is analyzed by electrophoresis?

8.5 A coat protein from a particular virus can be used to immunize children against further infection. However, inoculation of children with proteins extracted from natural viruses sometimes causes fatal disease, due to contamination with live viruses. How could you use molecular biology to produce an optimal vaccine?

8.6 How would cloning be different if there were no selectable markers?

8.7 You believe that a particular form of cancer is caused by a 200bp deletion in a particular human gene that is normally 2kb long.
a) Explain how you would use Southern blotting to diagnose the disease.
b) How would the blot appear in a heterozygote?
c) How would the blot appear in a homozygote with the deletion?
d) How would the blot appear in a homozygote without the deletion?
e) How would any of the blots appear if you hybridized and washed at very low temperature?

8.8 Refer to question 8.7.
a) Explain how you would detect the presence of the same deletion using PCR, rather than Southern blot.
b) How would the results appear in a heterozygote?
c) How would the results appear in a homozygote with the deletion?
d) How would the results appear in a homozygote without the deletion?
e) How would any of the blots appear if you annealed at very low temperature?

8.9 You have a PCR fragment for a human olfactory receptor gene (perception of smells). You want to know what genes a dog might have that are related to this human gene. How can you use your PCR fragment and genomic DNA from a dog to find this out?

8.10 You mix ligase with a sticky-ended plasmid and stick-ended insert fragment, which both have compatible ends. Unbeknownst to you, someone in the lab left the stock of ligase enzyme out of the freezer and it no longer works. Explain in detail what will happen in your ligation experiment in this situation.

Chapter 9 MOLECULAR MARKERS & QUANTITATIVE TRAITS

Figure 9.1 Many interesting traits, such as body mass, show continuous variation. Although environment obviously also affects this trait, some of the variation observed between individuals is heritable, and is dependent on interactions involving multiple genes. The study of quantitative traits is one of many applications of molecular markers.

Imagine that you could compare the complete DNA of any two people you meet today. Although their genomic DNA sequences would be very similar, they would not be identical at each of the 3 billion base pair positions you examined (unless, perhaps, your subjects were identical twins). In fact, the genomic sequences of almost any two unrelated people differ at millions of nucleotide positions. Some of these polymorphisms are found in the regions of genes that code for proteins. Other polymorphisms might affect the amount of transcript that is made for a particular gene. A person's health, appearance, behavior, and other characteristics depend in part on these polymorphisms. However, the vast majority of DNA polymorphisms have no effect on protein sequences, because they occur within regions of DNA that neither encode proteins, nor regulate the expression of genes. These polymorphisms are nevertheless very useful because they can be used as **molecular markers** in medicine, forensics, ecology, agriculture, and many other fields. In most situations, molecular markers obey the same rules of inheritance that we have already described for other types of loci, and so can be used to create genetic maps and to identify linked genes.

ORIGINS OF MOLECULAR POLYMORPHISMS

Mutations of DNA sequences can arise in many ways (Chapter 4). Some of these changes occur during DNA replication processes, resulting in an insertion, deletion, or substitution of one or a few nucleotides. Larger mutations can be caused by mobile genetic elements such as transposons, which are inserted more or less randomly into chromosomal DNA, sometimes occurring in clusters. In these and other types of **repetitive DNA** sequences, the number of repeated units is prone to change through unequal crossovers and other rare events.

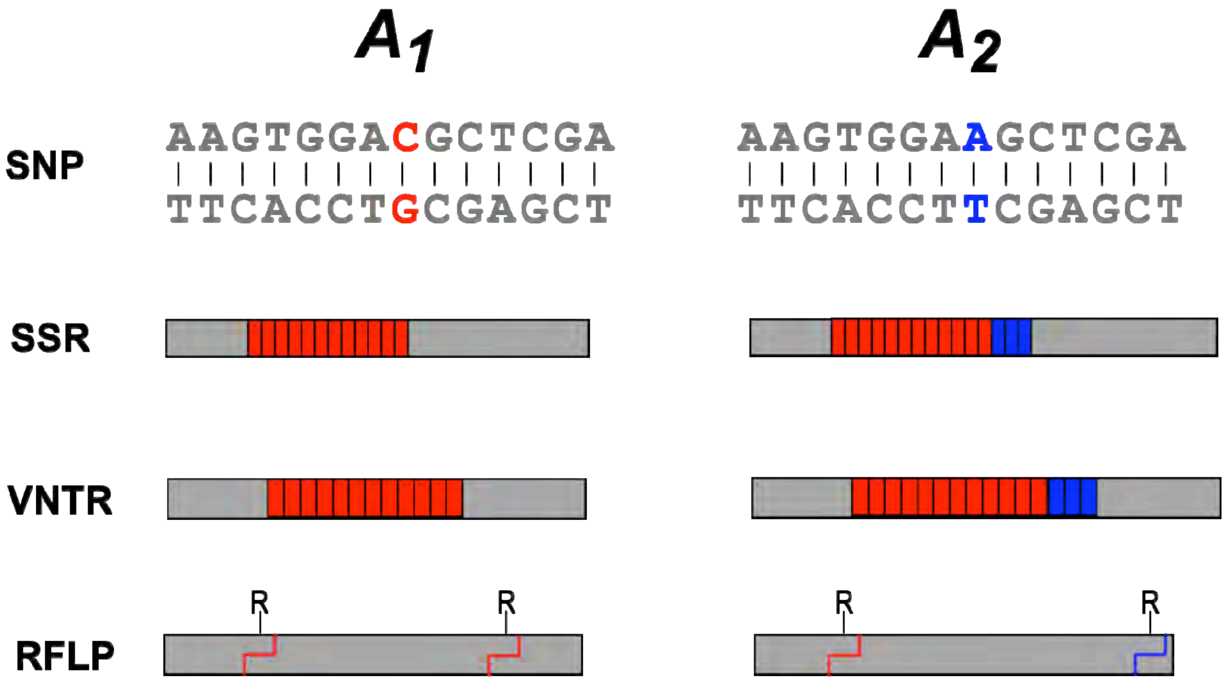


Figure 9.2 Some examples of DNA polymorphisms. The variant region is marked in blue, and each variant sequence is arbitrarily assigned one of two allele labels. Abbreviations: SNP (Single Nucleotide Polymorphism); SSR (Simple Sequence Repeat) = SSLP (Simple Sequence Length Polymorphism); **VNTR** (Variable Number of Tandem Repeats); **RFLP** (Restriction Fragment Length Polymorphisms. VNTRs and SSRs differ in the size of the repeat unit; VNTRs are larger than SSRs.

Classification and Detection of Molecular Markers

Regardless of their origins, molecular markers can be classified as polymorphisms that either vary in the length of a DNA sequence, or vary only in the identity of nucleotides at a particular position on a chromosome (Figure 9.2). In both cases, because two or more alternative versions of the DNA sequence exist, we can treat each variant as a different allele of a single locus. Each allele gives a different molecular phenotype. For example, polymorphisms of **SSRs** (short sequence repeats) can be distinguished based on the length of PCR products: one allele of a particular SSR locus might produce a 100bp band, while the same primers used with a different allele as a template might produce a 120bp band (Figure 9.3). A different type of marker, called a **SNP** (single nucleotide polymorphism), is an example of polymorphism that varies in nucleotide identity, but not length. SNPs occur at the highest density of any molecular markers, and the genotypes of thousands of SNP loci can be determined in parallel, using new, hybridization based instruments are one the of the most common types of Note that the alleles of most molecular markers are co-dominant, since it is possible to distinguish the phenotype of a heterozygote from either homozygote.

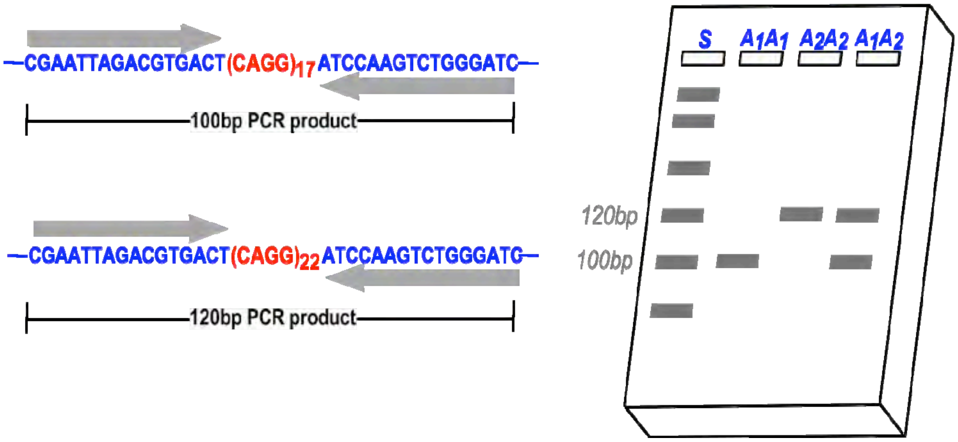

—CGAATTAGACGTGACT(CAGG)17ATCCAAGTCTGGGATC—
100bp PCR product
—CGAATTAGACGTGACT(CAGG)22ATCCAAGTCTGGGATC—
120bp PCR product
S
A1A1
A2A2
A1A2
120bp
100bp

Figure 9.3
Determining the genotype of an individual at a single SSR locus using a specific pair of PCR primers and agarose gel electrophoresis. S= size standard

Mutations that do not affect the function of protein sequences or gene expression are likely to be retained in a population as polymorphisms, since there will be no selection either in favor or against them (i.e. they are **neutral**). Note that the although the rate of spontaneous mutation in natural populations is sufficient to generate millions of polymorphisms that accumulate over thousands of generations, the rate of mutation is on the other hand sufficiently low that existing polymorphisms are stable throughout the generations we study in a typical genetic experiment.

Applications of Molecular Markers

Several characteristics of molecular markers make them useful to geneticists. Because of the way DNA polymorphisms arise and are retained, they are present at high density throughout the genome. Because they are phenotypically neutral, they can also be highly polymorphic, meaning that it is relatively easy to find markers that differ between two individuals. The neutrality of molecular markers makes it possible to study hundreds of loci without worrying about gene interactions or other influences that make it difficult to infer genotype from phenotype. Moreover, unlike visible traits such as eye color or petal color, the phenotype of a molecular marker can be detected in any tissue or developmental stage, and the same type of assay can be used to score molecular phenotypes at millions of different loci. Thus, the neutrality, high density, high degree of polymorphism, co-dominance, and ease of detection of molecular markers have lead to their wide adoption in many areas of research.

It is worth emphasizing again that DNA polymorphisms are a natural part of most genomes. Geneticists discover these polymorphisms in various ways, including comparison of random DNA sequence fragments from several individuals in a population. Once molecular markers have been identified, they can be used in many ways, including:

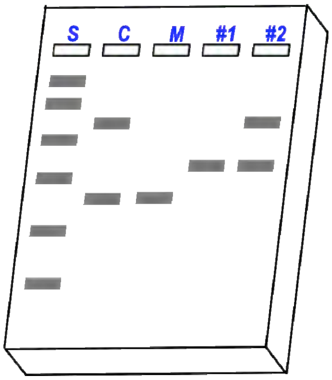


Figure 9.4 Paternity testing. Given the molecular phenotype of the child (C) and mother (M), only one of the possible fathers (#2) has alleles that are consistent with the child's phenotype.

DNA FINGERPRINTING

By comparing the genotypes of several molecular markers, it is possible to determine the relationship between two DNA samples. For example, a forensic scientist can demonstrate that some of the blood on a weapon came from a particular suspect, or even that leaves in the back of a suspect's pick-up truck came from a particular tree at a crime scene (Figure 9.4). DNA fingerprinting is also useful in paternity testing and in commercial applications such as verification of species of origin of certain foods and herbal medicines.

CONSTRUCTION OF GENETIC LINKAGE MAPS

By calculating the recombination frequency between pairs of molecular markers, a map of each chromosome can be generated for almost any organism (Figure 9.5). These maps are calculated using the same mapping techniques described for genes in Chapter 7, however, the high density and ease with which molecular markers can be genotyped makes them more useful than other phenotypes for constructing genetic maps. These maps are useful in further studies, including map-based cloning of protein coding genes that were identified by mutation.

POPULATION STUDIES

As described in Chapter 5, the observed frequency of alleles, including alleles of molecular markers, can be compared to frequencies expected for populations in Hardy-Weinberg equilibrium to determine whether the population is in equilibrium. By monitoring molecular markers, ecologists and wildlife biologists can make inferences about migration, selection, diversity, and other population-level parameters. Molecular markers can also be used by anthropologists to study migration events in human ancestry.

IDENTIFICATION OF LINKED TRAITS

It is often possible to identify an allele of a molecular marker whose presence is correlated with a particular disease or other trait of interest. One way to make this correlation is to obtain genomic DNA samples from hundreds of individuals with a particular disease, as well as samples from a control population of healthy individuals. The genotype of each individual is scored at hundreds or thousands of molecular marker loci (e.g. SNPs), to find alleles that are usually

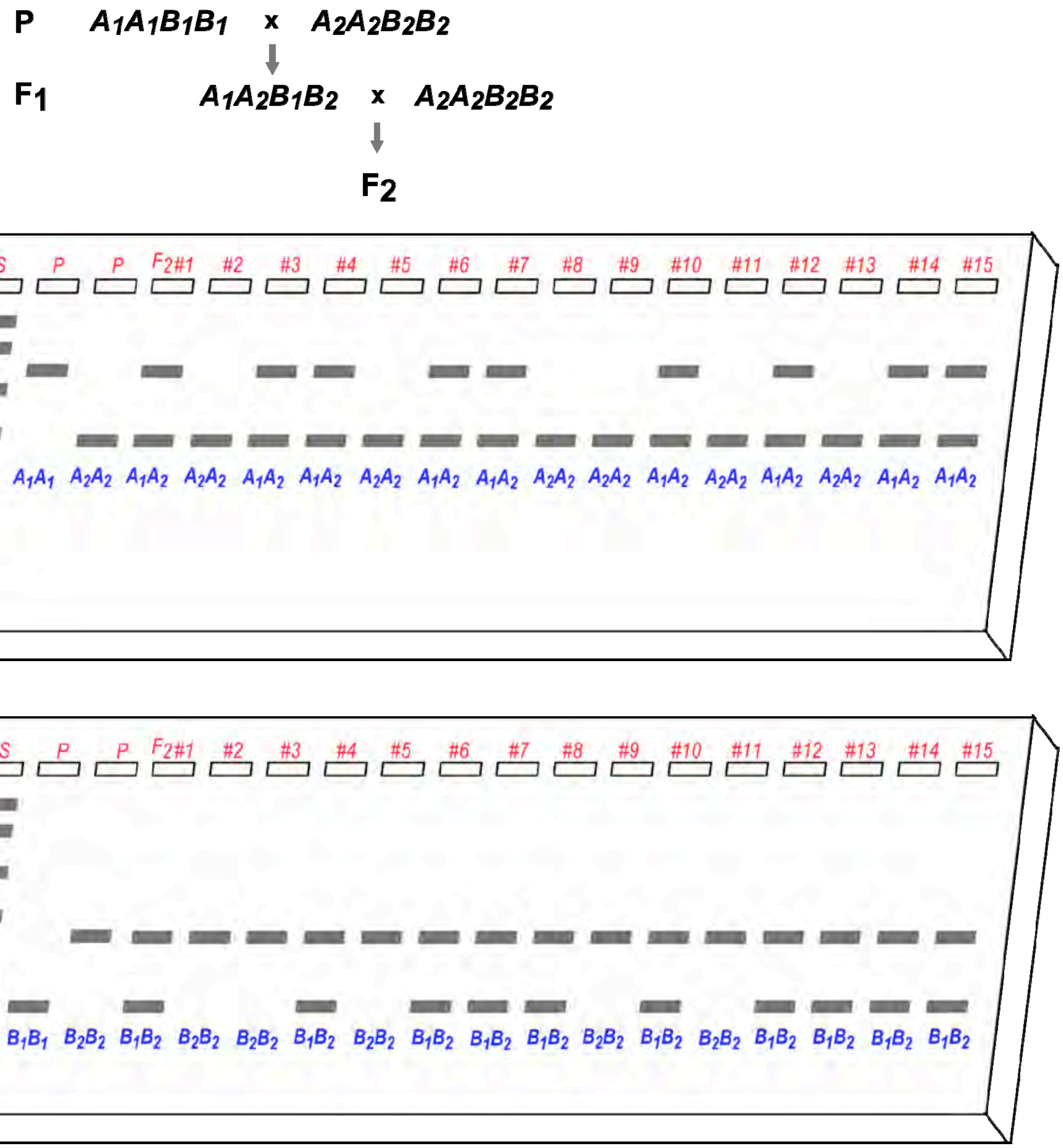

P $A_1A_1B_1B_1$ x $A_2A_2B_2B_2$
F1 $A_1A_2B_1B_2$ x $A_2A_2B_2B_2$
F2
S P P F2#1 #2 #3 #4 #5 #6 #7 #8 #9 #10 #11 #12 #13 #14 #15
A_1A_1 A_2A_2 A_1A_2 A_2A_2 A_1A_2 A_1A_2 A_2A_2 A_1A_2 A_1A_2 A_2A_2 A_2A_2 A_1A_2 A_2A_2 A_1A_2 A_2A_2 A_1A_2 A_1A_2
S P P F2#1 #2 #3 #4 #5 #6 #7 #8 #9 #10 #11 #12 #13 #14 #15
B_1B_1 B_2B_2 B_1B_2 B_2B_2 B_2B_2 B_1B_2 B_2B_2 B_1B_2 B_1B_2 B_1B_2 B_2B_2 B_1B_2 B_2B_2 B_1B_2 B_1B_2 B_1B_2 B_1B_2

Figure 9.5 Measuring recombination frequency between two molecular marker loci, *A* and *B*. A different pair of primers is used to amplify DNA from either parent (P) and 15 of the F_2 offspring from the cross shown. Recombinant progeny will have the genotype $A_1A_2B_2B_2$ or $A_2A_2B_1B_2$. Individuals #3, #8, #13 are recombinant, so the recombination frequency is 3/15=20%.

present in persons with the disease, but not in healthy subjects. The molecular marker is presumed to be tightly linked to the gene that causes the disease, although this protein-coding gene may itself be unknown. The presence of a particular molecular polymorphism may therefore be used to diagnose a disease, or to advise an individual of susceptibility to a disease.

Molecular markers may also be used in a similar way in agriculture. For example, markers associated with desirable traits can be identified by screening both the traits and molecular marker genotypes of hundreds of individuals. Markers that are linked to desirable traits can then be used during breeding to select varieties with economically useful combinations of traits, even when the genes underlying the traits are not known.

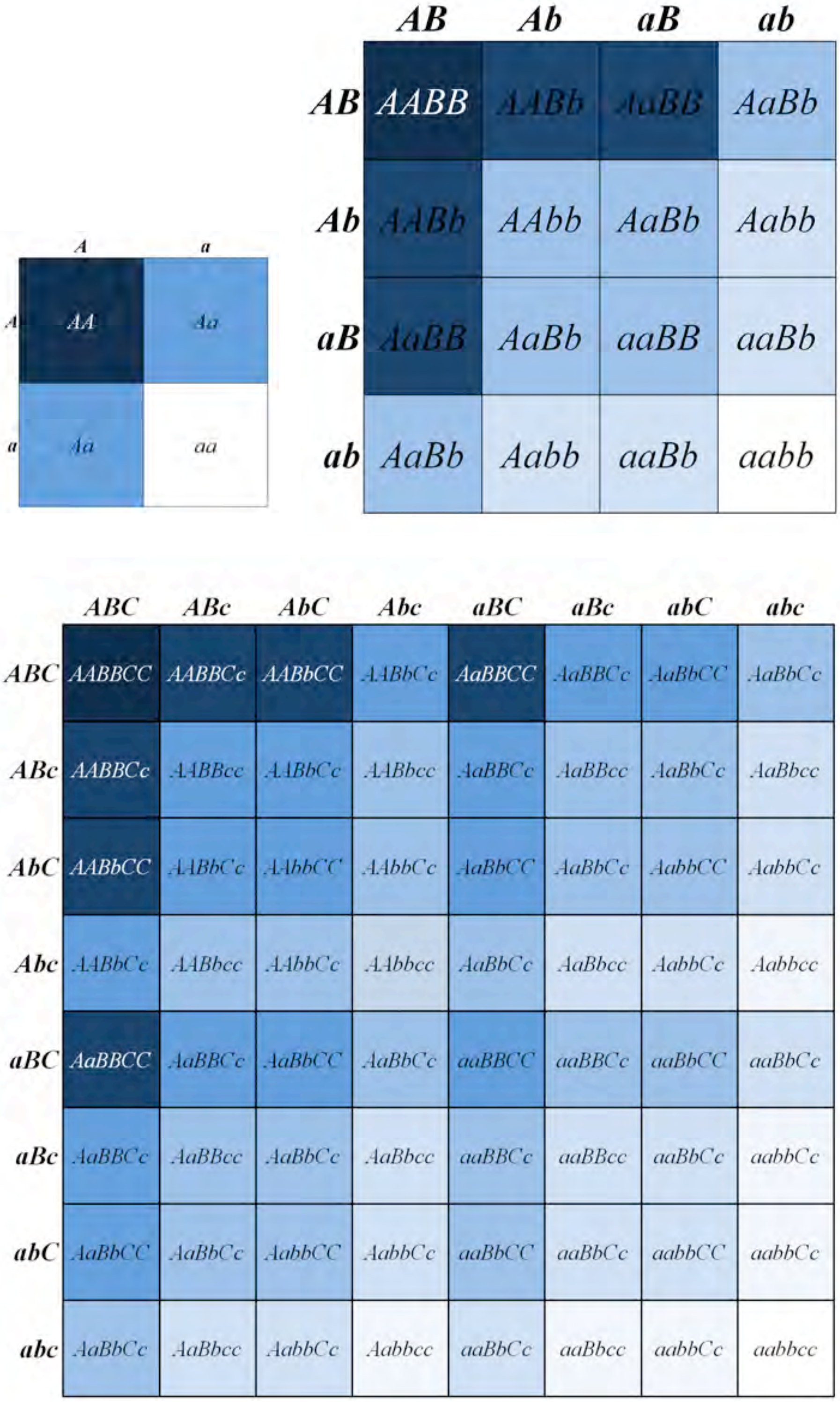
QUANTITATIVE TRAIT LOCUS (QTL) MAPPING

Molecular markers can be used to identify regions of chromosomes that contain genes that act together to produce complex traits. This process involves finding combinations of alleles of molecular markers that are correlated with a quantitative phenotype such as body mass, or intelligence. QTL mapping is described in more detail in the following section.

QUANTITATIVE TRAIT LOCUS ANALYSIS

Most of the phenotypic traits commonly used in introductory genetics are qualitative, meaning that the phenotype exists in only two (or possibly a few more) alternative forms, such as either purple or white flowers, or red or white eyes. These qualitative traits are therefore said to exhibit **discrete variation**. On the other hand, many interesting and important traits exhibit **continuous variation**; these exhibit a continuous range of phenotypes that are usually measured quantitatively, such as intelligence, body mass, blood pressure in animals including humans, and yield, water use, or vitamin content in crops. Traits with continuous variation are often complex, and do not show the simple Mendelian segregation ratios (e.g. 3:1) observed with some qualitative traits. Many complex traits are also influenced heavily by the environment. Nevertheless, complex traits can often be shown to have a component that is heritable, and which must therefore involve one or more genes.

How can genes, which are inherited (in the case of a diploid) as at most two variants each, explain the wide range of continuous variation observed for many traits? The lack of an immediately obvious explanation to this question was one of the early objections to Mendel's explanation of the mechanisms of heredity. However, upon further consideration, it becomes clear that the more loci that contribute to trait, the more phenotypic classes may be observed for that trait (Figure 9.6). If the number of phenotypic classes is sufficiently large,



	A	a
A	AA	Aa
a	Aa	aa

	AB	Ab	aB	ab
AB	AABB	AABb	AaBB	AaBb
Ab	AABb	AAbb	AaBb	Aabb
aB	AaBB	AaBb	aaBB	aaBb
ab	AaBb	Aabb	aaBb	aabb

	ABC	ABc	AbC	Abc	aBC	aBc	abC	abc
ABC	AABBCC	AABBCc	AABbCC	AABbCc	AaBBCC	AaBBCc	AaBbCC	AaBbCc
ABc	AABBCc	AABBcc	AABbCc	AABbcc	AaBBCc	AaBBcc	AaBbCc	AaBbcc
AbC	AABbCC	AABbCc	AAbbCC	AAbbCc	AaBbCC	AaBbCc	AabbCC	AabbCc
Abc	AABbCc	AABbcc	AAbbCc	AAbbcc	AaBbCc	AaBbcc	AabbCc	Aabbcc
aBC	AaBBCC	AaBBCc	AaBbCC	AaBbCc	aaBBCC	aaBBCc	aaBbCC	aaBbCc
aBc	AaBBCc	AaBBcc	AaBbCc	AaBbcc	aaBBCc	aaBBcc	aaBbCc	aabbCc
abC	AaBbCC	AaBbCc	AabbCC	AabbCc	aaBbCC	aaBbCc	aabbCC	aabbCc
abc	AaBbCc	AaBbcc	AabbCc	Aabbcc	aaBbCc	aaBbcc	aabbCc	aabbcc

Figure 9.6 Punnett Squares for one, two, or three loci. We are using a simplified example of up to three semi-dominant genes, and in each case the effect on the phenotype is additive, meaning the more "upper case" alleles present, the stronger the phenotype. Comparison of the Punnett Squares and the associated phenotypes shows that under these conditions, the larger the number of genes that affect a trait, the more intermediate phenotypic classes that will be expected.

individual classes may become indistinguishable from each other (particularly when environmental effects are included), and the result is continuous variation (Figure 9.7). Thus, quantitative traits are sometimes called **polygenic traits**, because it is assumed that their phenotypes are controlled by the combined activity of many genes. Note that this does not imply that each of the individual genes has an equal influence on a polygenic trait. Furthermore, any give gene may influence more than one trait, whether these traits are quantitative or qualitative traits.

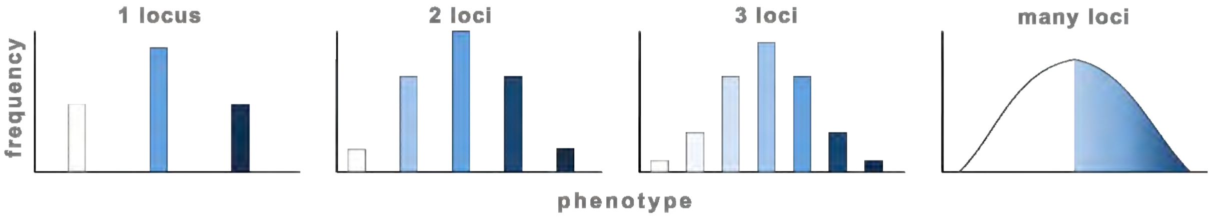


Figure 9.7 The more loci that affect a trait, the larger the number of phenotypic classes that can be expected. For some traits, the number of contributing loci is so large that the phenotypic classes blend together in apparently continuous variation.

We can use molecular markers to identify at least some of the genes that affect a given quantitative trait. This is essentially an extension of the mapping techniques we have already considered for discrete traits. A QTL mapping experiment will ideally start with two pure-breeding lines that differ greatly from each other in respect to one or more quantitative traits (Figure 9.8). The parents and all of their progeny should be raised in under similar environmental conditions, to ensure that observed variation is due to genetic rather than external factors. These parental lines must also be polymorphic for a large number of molecular loci, meaning that they must have different alleles from each other at hundreds of loci. The parental lines are crossed, and then this F_1 individual, in which recombination between parental chromosomes has occurred is self-fertilized (or back-crossed). Because of recombination, each of the F_2 individuals will contain a different combination of molecular markers, and also a different combination of alleles for the genes that control the quantitative trait of interest (Table 9.1). By comparing the molecular marker genotypes of several hundred F_2 individuals with their quantitative phenotypes, a researcher can identify molecular markers for which the presence of particular alleles is always associated with extreme values of the trait (Figure 9.9). In this way, regions of chromosomes that contain genes that contribute to quantitative traits can be identified. It then takes much more work (further mapping and other experimentation) to identify the individual genes in each of the regions that control the quantitative trait.

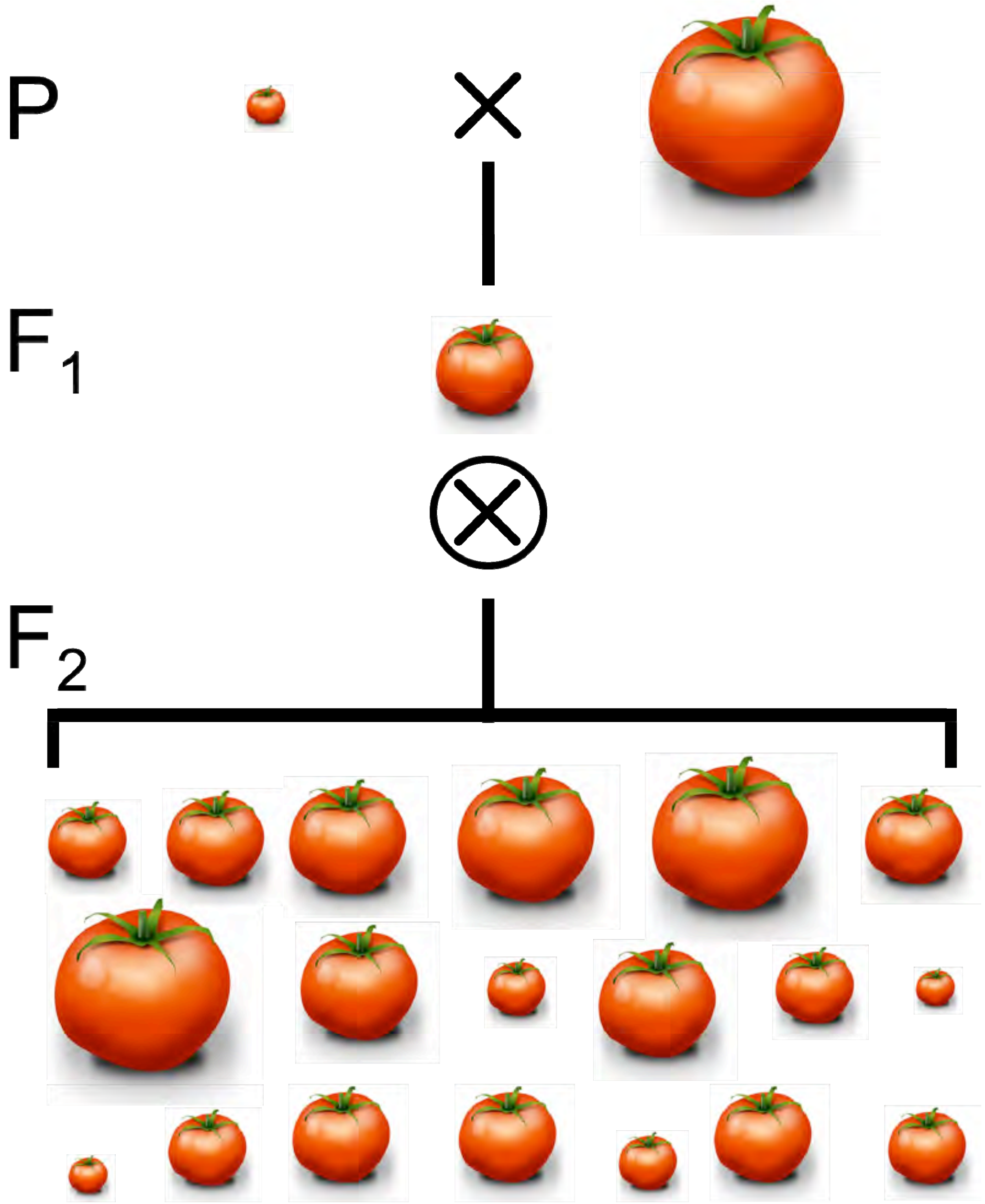


Figure 9.9 Strategy for a typical QTL mapping experiment. Two parents that differ in a quantitative trait (e.g. fruit mass) are crossed, and the F_1 is self-fertilized (as shown by the cross-in-circle symbol). The F_2 progeny will show a range of quantitative values for the trait. The task is then to identify alleles of markers from one parent that are strongly correlated with the quantitative trait. For example, markers from the large-fruit parent that are always present in large-fruit F_2 individuals (but never in small-fruit individuals) are likely linked to loci that control fruit mass.

Table 9.1 Genotypes and quantitative data for some individuals from the crosses shown in Figure 9.9

	genotype	fruit mass
P	$A_1A_1B_1B_1C_1C_1D_1D_1E_1E_1F_1F_1G_1G_1H_1H_1J_1J_1K_1K_1$	10g
P	$A_2A_2B_2B_2C_2C_2D_2D_2E_2E_2F_2F_2G_2G_2H_2H_2J_2J_2K_2K_2$	90g
F_1	$A_1A_2B_1B_2C_1C_2D_1D_2E_1E_2F_1F_2G_1G_2H_1H_2J_1J_2K_1K_2$	50g
F_2 #001	$A_1A_1B_1B_2C_1C_1D_2D_2E_1E_2F_1F_2G_1G_2H_1H_1J_1J_2K_1K_2$	80g
F_2 #002	$A_1A_2B_1B_2C_1C_2D_1D_1E_1E_2F_1F_2G_2G_2H_1H_2J_2J_2K_1K_1$	10g
F_2 #003	$A_2A_2B_1B_2C_2C_2D_1D_2E_1E_2F_1F_2G_1G_1H_1H_2J_1J_2K_1K_2$	50g
F_2 #004	$A_1A_2B_1B_2C_1C_2D_1D_2E_1E_2F_1F_2G_1G_2H_1H_2J_1J_2K_2K_2$	60g
F_2 #005	$A_1A_2B_1B_1C_1C_2D_2D_2E_1E_2F_1F_2G_1G_2H_1H_2J_1J_1K_2K_2$	90g
F_2 #006	$A_1A_2B_2B_2C_1C_2D_1D_2E_1E_2F_1F_2G_2G_2H_1H_2J_1J_2K_1K_2$	60g
F_2 #007	$A_2A_2B_1B_1C_1C_2D_2D_2E_1E_2F_1F_2G_1G_1H_1H_2J_1J_1K_1K_2$	80g
F_2 #008	$A_1A_1B_1B_2C_1C_2D_1D_2E_1E_2F_1F_2G_1G_2H_1H_2J_1J_2K_1K_2$	50g
F_2 #009	$A_1A_2B_1B_2C_2C_2D_1D_2E_1E_2F_1F_2G_1G_2H_1H_2J_1J_2K_1K_2$	50g
F_2 #010	$A_1A_2B_1B_2C_1C_2D_1D_2E_1E_2F_1F_2G_1G_2H_1H_2J_1J_2K_2K_2$	30g
F_2 #011	$A_1A_2B_1B_2C_1C_2D_2D_2E_1E_1F_1F_2G_1G_2H_1H_2J_1J_2K_1K_2$	80g
F_2 #012	$A_1A_1B_1B_2C_1C_2D_1D_1E_1E_2F_2F_2G_1G_2H_1H_2J_1J_2K_2K_2$	30g
F_2 #013	$A_2A_2B_1B_1C_1C_2D_1D_1E_1E_2F_1F_1G_1G_2H_2H_2J_1J_1K_1K_1$	10g
F_2 #014	$A_2A_2B_1B_1C_1C_1D_2D_2E_1E_2F_1F_2G_1G_2H_1H_2J_1J_1K_1K_1$	70g

F_2 #015	$A_2A_2B_2B_2C_1C_2D_1D_2E_1E_2F_2F_2G_1G_2H_1H_1J_2J_2K_1K_2$	40g
F_2 #016	$A_1A_2B_2B_2C_1C_2D_1D_1E_1E_2F_1F_1G_2G_2H_1H_1J_1J_2K_1K_1$	10g
F_2 #017	$A_1A_2B_2B_2C_1C_2D_2D_2E_2E_2F_1F_1G_2G_2H_1H_2J_1J_2K_2K_2$	90g
F_2 #018	$A_1A_2B_2B_2C_1C_2D_1D_2E_1E_2F_1F_1G_2G_2H_1H_2J_1J_2K_1K_1$	40g
F_2 #019	$A_1A_1B_1B_2C_1C_2D_1D_1E_1E_2F_2F_2G_1G_1H_1H_1J_1J_2K_1K_2$	20g
F_2 #100	$A_1A_1B_1B_2C_1C_2D_2D_2E_1E_2F_1F_2G_2G_2H_1H_2J_2J_2K_1K_2$	80g

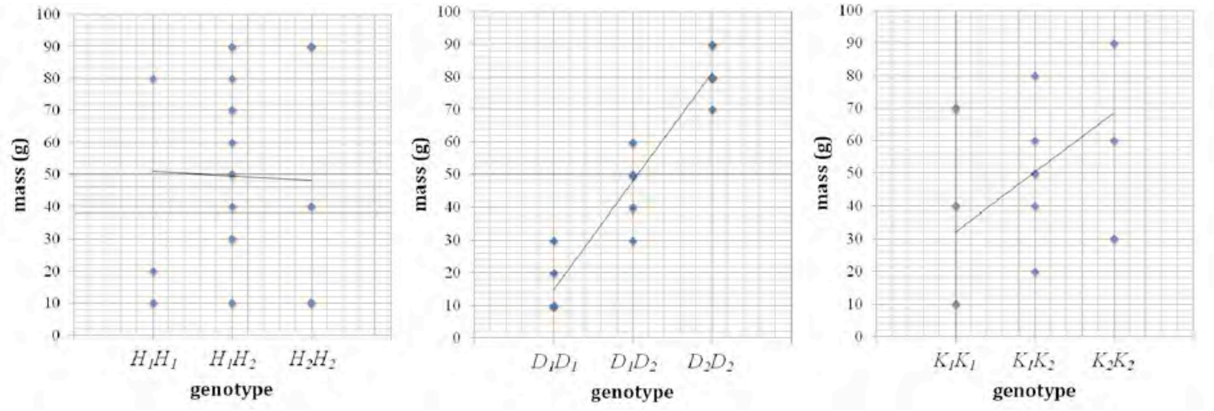


Figure 9.10 Plots of fruit mass and genotype for selected loci from Table 9.1. For most loci (e.g. *H*), the genotype shows no significant correlation with fruit weight. However, for some molecular markers, the genotype will be highly correlated with fruit weight. Both *D* and *K* influence fruit weight, but the effect of genotype at locus *D* is larger than at locus *K*.

SUMMARY:

- Natural variations in the length or identity of DNA sequences occur at millions of locations throughout most genomes.
- DNA polymorphisms are often neutral, but because of linkage may be used as molecular markers to identify regions of genomes that contain genes of interest.
- Molecular markers are useful because of their neutrality, co-dominance, density, allele frequencies, ease of detection, and expression in all tissues.
- Molecular markers can be used for any application in which the identity of two DNA samples is to be compared, or when a particular region of a chromosome is to be correlated with inheritance of a trait.
- Many important traits show continuous, rather than discrete variation. These are also called quantitative traits.
- Many quantitative traits are influenced by a combination of environment and genetics.
- The heritable component of quantitative traits can best be studied under controlled conditions, with pure-breeding parents that are polymorphic for both a quantitative trait and a large number of molecular markers.
- Molecular markers can be identified for which specific alleles are tightly correlated with the quantitative value of a particular phenotype. The genes that are linked to these markers can be identified through subsequent research.

KEY TERMS:

molecular marker
repetitive DNA
SSR
SSLP
VNTR
SNP
RFLP
neutral mutation
QTL
discrete variation
continuous variation
polygenic

STUDY QUESTIONS:

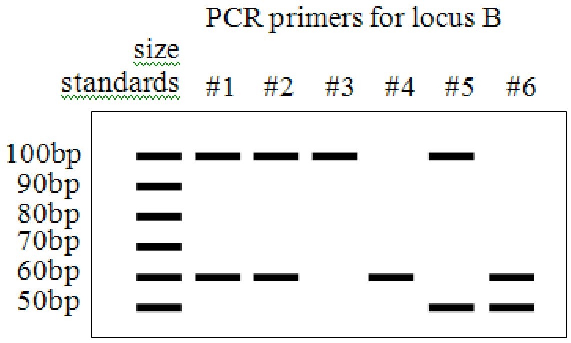
9.1 Three different polymorphisms have been identified at a particular molecular marker locus. A single pair of PCR primers will amplify either a 50bp fragment (*B2*), a 60bp fragment (*B3*), or a 100bp fragment (*B4*).

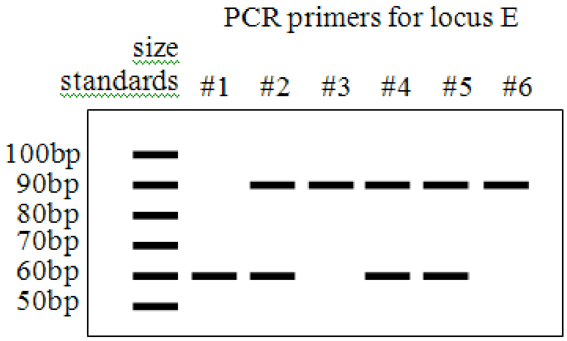
Draw the PCR bands that would be expected if these primers were used to amplify DNA from individuals with each of the following genotypes:

a) B_2B_2
b) B_4B_4
c) B_2B_3
d) B_2B_4

9.2 In addition to the primers used to genotype locus *B* (described above), a separate pair of primers can amplify another polymorphic SSR locus *E,* with either a 60bp product (E_1) or a 90bp (E_2) product. DNA was extracted from six individuals (#1- #6), and DNA from each individual was used as a template in separate PCR reactions with primers for either locus *B* or primers for locus *E*, and the PCR products were visualized on electrophoretic gels as shown below.

Based on the following PCR banding patterns, what is the full genotype of each of the six individuals?





9.3 Based on the genotypes you recorded in Question 9.2, can you determine which of the individuals could be a parent of individual #1?

9.4 Here is part of the DNA sequence of a chromosome:

```
TAAAGGAATCAATTACTTCTGTGTGTGTGTGTGTGTGTGTGTGTGTTCTTAGTTGTTTAAGTTTTAAGTTGTGA
||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
ATTTCCTTAGTTAATGAAGACACACACACACACACACACACACACAAGAATCAACAAATTCAAAATTCAACACT
```

Identify the following features on the sequence:

a) the region of the fragment that is most likely to be polymorphic
b) any simple sequence repeats
c) the best target sites for PCR primers that could be used to detect polymorphisms in the length of the simple sequence repeat region in different individuals

9.5 In a particular diploid plant, seed color is a polygenic trait. If true-breeding plants that produce red seeds are crossed with true breeding plants that produce white seeds, the F_1 produces seeds that are intermediate in color (i.e. pink). When an F_1 plant self-fertilizes, white seeds are observed in the next generation. How many genes are involved in seed color for each of the following frequencies of white seeds in the F_2 generation?

a) 1/4 white seeds
b) 1/16 white seeds
c) 1/64 white seeds
d) 1/256 white seeds

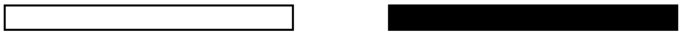
9.6 If height in humans is a polygenic trait, explain why it occasionally happens that two tall parents have a child who grows up to be much shorter than either of them.

9.7 In quantitative trait (QTL) mapping, researchers cross two parents that differ in expression of some quantitative trait, then allow chromosomes from these parents to recombine randomly, and after several generations of inbreeding, produce a large number of offspring ("recombinant inbred lines"). Because the position of crossovers is random, each of the offspring contain a different combination of chromosomal regions from each of the two parents. The researchers then use molecular markers to determine which chromosomal regions have the greatest influence on the quantitative trait, e.g. in tall offspring, which chromosomal regions always come from the tall parent?

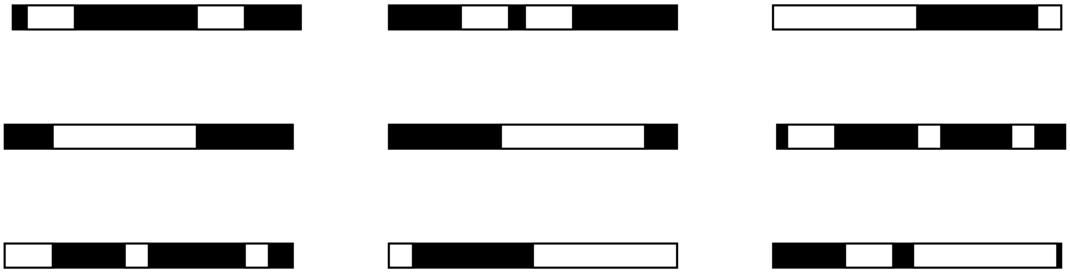
Imagine that two mice strains have been identified that differ in the time required to complete a maze, which may be an indication of intelligence. The time for maze completion is heritable and these parental strains “breed true” for the same completion time in each generation. Imagine also that their chromosomes are different colors and we can track the inheritance of chromosomal regions from each parent based on this color.

Based on the following diagrams of one chromosome from each individual in a pedigree, identify a chromosomal region that may contain a gene that affects time to complete a maze. The time for each individual is shown below each chromosome. Assume that all individuals are homozygous for all loci.

Parents:



Selected individuals from among F8 progeny of the above parents:



9.8 In a more realistic situation (as compared to question 7), where you could not distinguish the parental origin of different chromosomal regions just by appearance of chromosomes, explain how you could identify which parent was the source of a particular region of a chromosome in recombinant offspring.

Chapter 10 Genomics and Systems Biology

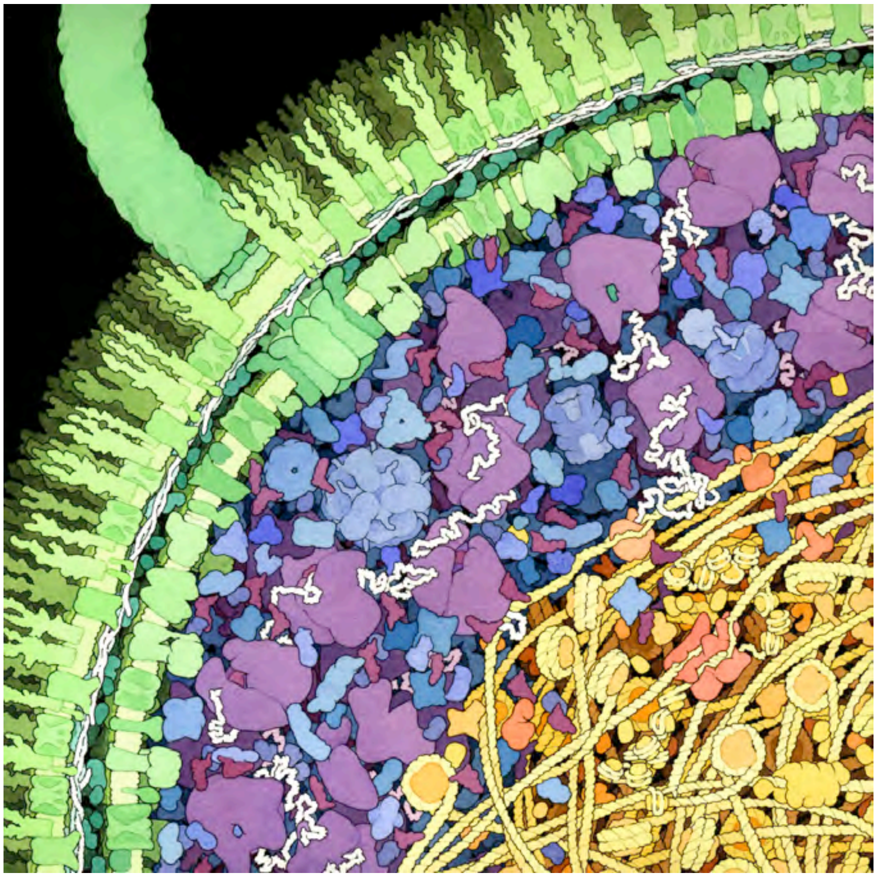


Figure 10.1 An artist's depiction of part of an *E.coli* cell, showing many different types of molecules in their typical abundance. mRNA appears as white lines associated with purple ribosomes, while DNA and proteins such as histones are yellow.

Imagine that you could identify and quantify every molecule within a cell (Figure 10.1) in a single assay. You could use this ability to better understand almost any aspect of biology. For example, by comparing the molecular profiles of plants that differed in their resistance to drought, you might discover which combination of genes or proteins makes a crop drought tolerant. The complete set of DNA within an organism is called its **genome**. Although it is not currently possible to study literally every molecule in a cell in a single experiment, recent advances in molecular biology have made it possible to study many genes (or their products) in parallel. **Genomics** is therefore the large-scale application of techniques of molecular biology; the study of many genes or even many thousands of genes at once. This type of research is facilitated by technologies that increase throughput (i.e. rate of analysis), and decrease cost. The –**omics** suffix has been used to indicate high-throughput analysis of many types of molecules, including transcripts (**transcriptomics**), proteins (**proteomics**), and the products of enzymatic reactions, or metabolites (**metabolomics**; Figure 10.2). Interpretation of the large data sets generated by –omics research depends on a combination of computational, biological, and statistical knowledge provided by experts in **bioinformatics**. Attempts to combine information from different types of –omics studies is sometimes called **systems biology**.

Figure 10.2 A word cloud listing some of the different –omics technologies. Terms that are more widely used are written in the largest characters. There is no significance to the color of each word.

DNA Sequencing

The purpose of DNA sequencing is to determine the order of nucleotide bases within a given fragment of DNA. This information can be used to infer the protein sequence encoded by the gene, from which further inferences may be made about the gene's function and its relationship to other genes. DNA sequence information is also useful in studying the regulation of gene expression. If DNA sequencing is applied to the study of many genes, or even a whole genome, it is considered an example of genomics.

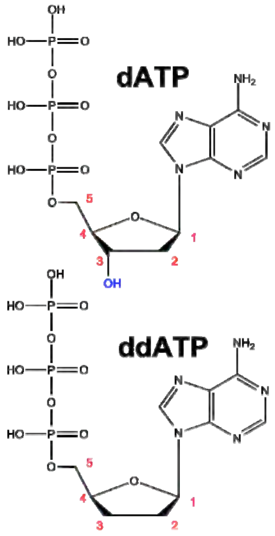


Figure 10.3 Structure of a terminator (ddATP) compared to a regular nucleoside. Note the absence of the hydroxyl group in the 3' position of ddATP

Recall that DNA polymerases incorporate nucleotides (dNTPs) into a growing strand of DNA. DNA polymerases can generally add a new base only to the 3'-OH group of an existing strand of DNA; this is why primers are required in natural DNA synthesis and in techniques such as PCR. Most of the currently used DNA sequencing techniques rely on modified nucleotides called terminators. Examples of terminators are the dideoxy nucleotides (**ddNTPs**), which lack a 3'-OH group and therefore cannot serve as an attachment site for the addition of new bases to a growing strand of DNA (Figure 10.3). Thus, after a ddNTP is incorporated into a strand of DNA, no further elongation can occur. Terminators labeled with fluorescent dyes specific to each of the four nucleotide bases are particularly useful in DNA sequencing.

To sequence a DNA fragment, you need many copies of that fragment (Figure 10.4). Unlike PCR, DNA sequencing does not amplify the target sequence. In another difference from PCR, only one primer is used in DNA sequencing. This primer is hybridized to the denatured template DNA, and determines where on the template strand the sequencing reaction will begin. A mixture of dNTPs, fluorescently labeled terminators, and DNA polymerase is added to a tube containing the primer-template hybrid. The DNA polymerase will then synthesize a new strand of DNA until a fluorescently labeled nucleotide is incorporated. Because the reaction contains millions of template molecules, a corresponding number of shorter molecules is synthesized, each ending in a fluorescent label that corresponds to the last base incorporated. The newly synthesized strands can be denatured from the template, and then separated electrophoretically based on their length (Figure 10.5). Since each band differs in length by one nucleotide, and the identity of that nucleotide is known from its fluorescence, the DNA sequence can be read simply from the order of the colors in successive bands. In practice, the maximum length of sequence that can be read from a single sequencing reaction is about 700bp.

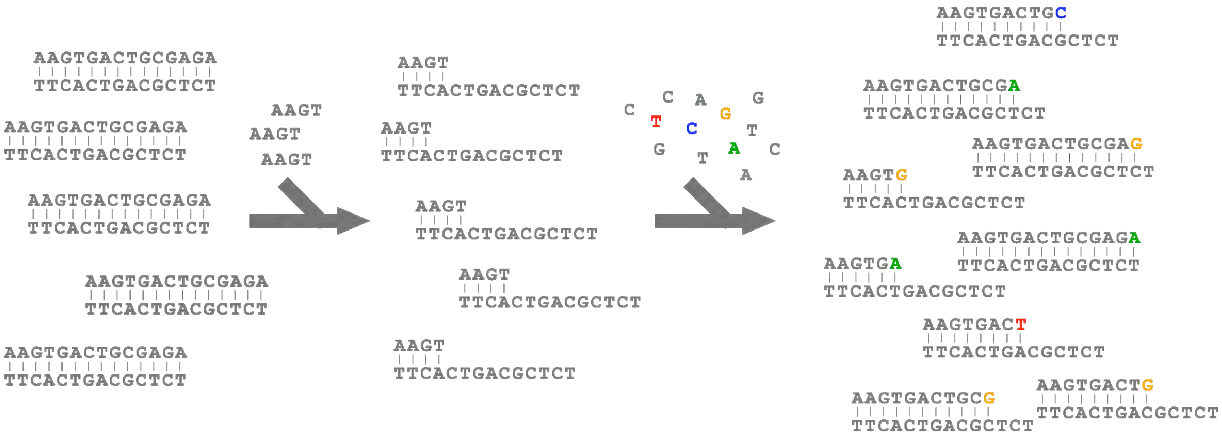


Figure 10.4 A sequencing reaction begins with many identical copies of a template DNA fragment. The template is denatured, then primers are annealed to the template. Following the addition of polymerase, regular dNTPs , and fluorescently labeled terminators, extension begins at the primer site. Elongation proceeds until a fluorescently labeled terminator (shown here in color) is incorporated.

A particularly sensitive electrophoresis method used in the analysis of DNA sequencing reactions is called **capillary electrophoresis** (Figure 10.6). In this method, a current pulls the sequencing products through a gel-like matrix that is encased in a fine tube of clear plastic. As in conventional electrophoresis, the smallest fragments move through the capillary the fastest. As they pass through a point near the end of the capillary, the fluorescent intensity of each dye is read. This produces a graph called a chromatogram (Figure 10.6). The sequence is determine by identifying the highest peak (i.e. the dye with the most intense fluorescent signal) at each position.

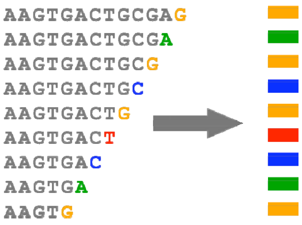


Figure 10.5 Fluorescently labeled products can be separated electrophoretically based on their length.

Advances in technology over the past two decades have increased the speed and quality of sequencing, while decreasing the cost. This has become especially true with the most recently developed methods called **next-generation** sequencing. Not all of these new methods rely on terminators, but one that does is a method used in instruments sold by a company called **Illumina**. Illumina sequencers use a special variant of PCR called bridge PCR to make many thousands of copies of a short (45bp) template fragment. Each of these short template fragments is attached in a cluster in a small spot on a reaction surface. Millions of other clusters, each made by different template fragment, are located at other positions on the reaction surface. DNA synthesis at each template strand then proceeds using dye-labeled terminators that are used are reversible. Synthesis is therefore terminated (temporarily) after the incorporation of each nucleotide. Thus, after the first nucleotide is incorporated in each strand, a camera records the color of fluorescence emitted from each cluster. The terminators are then modified, and a second nucleotide is incorporated in each strand, and again the reaction surface is photographed. This cycle is repeated a total of 45 times. Because millions of 45bp templates are sequenced in parallel in a single process, Illumina sequencing is very efficient compared to other sequencing techniques. However, the short length of the templates currently limits the application of the technology.

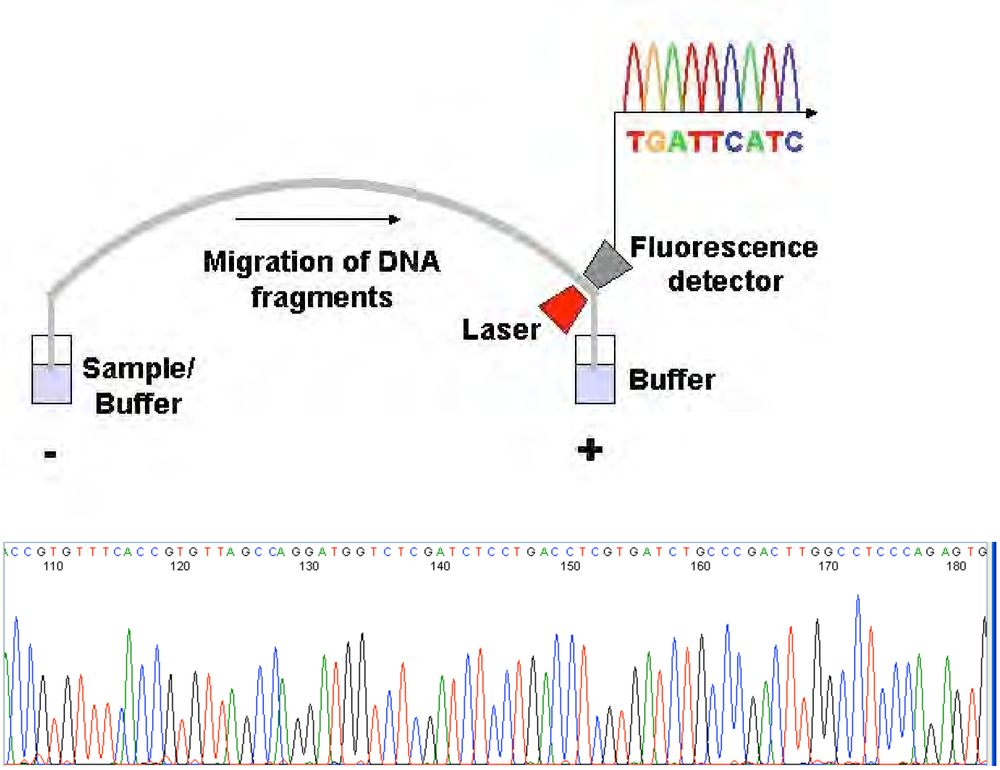

TGATTCATC
Migration of DNA fragments
Fluorescence detector
Laser
Sample/ Buffer
Buffer
-
+
ACCGTGTTTCACCGTGTTAGCCAGGATGGTCTCGATCTCCTGACCTCGTGATCTGCCCGACTTGGCCTCCCAGAGTG
110
120
130
140
150
160
170
180

Figure 10.6 Fluorescently labeled products can be separated by capillary electrophoresis, generating a chromatogram from which thesequence can be read.

Whole Genome Sequencing

Given that the length of a single sequencing read is somewhere between 45bp and 700bp, we are faced with a problem when we want to determine the sequence of longer fragments, including each of the chromosomes in an entire genome such as that of humans (3 x10^9 bp). Obviously, we need to break the genome into smaller fragments. There are two different strategies for doing this: clone-by-clone sequencing, which relies on the creation of a physical map, and whole genome shotgun sequencing, which does not require a physical map.

PHYSICAL MAPPING

A **physical map** is a representation of a genome that is comprised of cloned fragments of DNA (Figure 10.7). The map is therefore made of physical entities (pieces of DNA) rather than abstract concepts such as the linkage frequencies and genes that make up a genetic map. It is usually possible to correlate genetic and physical maps, for example by identifying the clone that contains a particular molecular marker. The connection between physical and genetic maps allows the genes underlying particular mutations to be identified through a process called **map-based cloning**.

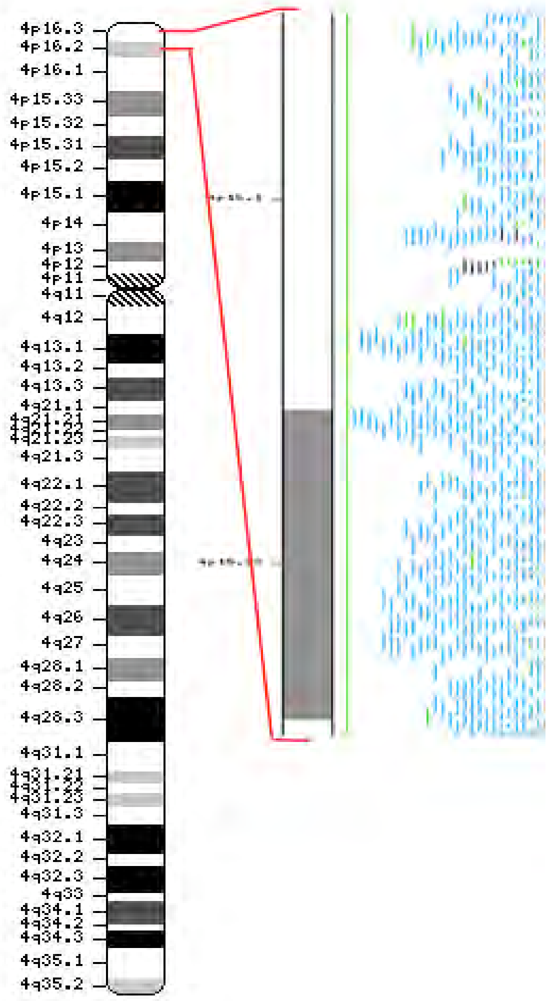


Figure 10.7 A portion of a physical map for human chromosome 4. The entire chromosome is shown at left. The physical map is made up of the small blue lines, each of which represents a cloned piece of DNA approximately100kb in length.

.

To create a physical map, large fragments of the genome are cloned into plasmid vectors, or into larger vectors called bacterial artificial chromosomes (**BACs**). BACs can contain approximately 100kb fragments. The set of BACs produced in a cloning reaction will be redundant, meaning that different clones will contain DNA from the same part of the genome. Because of this redundancy, it is useful to select the minimum set of clones that represent the entire genome, and to order these clones respective to the sequence of the original chromosome. Note that this is all to be done without knowing the sequence of each BAC. Making a physical map may therefore rely on techniques related to Southern blotting: DNA from the ends of one BAC is used as a probe to find another clone that contains at least part of the same sequence. By repeating this process, a series of clones can be identified that together contain the sequence of a larger part of a chromosome.

CLONE-BY-CLONE GENOME SEQUENCING

Physical mapping was once considered a pre-requisite for whole genome sequencing. The **clone-by-clone** sequencing process would begin by breaking the genome into BAC-sized pieces, arranging these BACs into a map, then breaking each BAC up into successively smaller clones, which were usually then also mapped. Eventually, a minimum set of clones would be identified, each of which was small enough to be sequenced. Because the order of clones relative to the complete chromosome was known prior to sequencing, the resulting sequence information could be easily assembled into a complete chromosome at the end of the project. Clone-by-clone sequencing therefore minimizes the number of sequencing reactions that must be performed, and makes sequence assembly straightforward. However, a drawback of this sequencing strategy is the tedious process of building a physical map.

WHOLE GENOME SHOTGUN SEQUENCING

Rather than making a physical map, why not simply break the genome into fragments that are small enough to be sequenced, then reassemble complete chromosome sequences simply by finding overlaps in the small fragments? This **whole genome shotgun** avoids the process of making a physical map. However, this requires many more sequencing reactions than the clone-by-clone method, because in the shotgun approach, there is no way to avoid sequencing redundant fragments. There is also a question of the feasibility of assembling complete chromosomes based simply on the sequence of many small fragments. This is particularly a problem when the size of the fragments is smaller than the length of a repetitive region of DNA. Nevertheless, it has now been demonstrated in the sequencing of rice, human, and many other large genomes, that whole genome shotgun sequencing is an efficient way to obtain nearly complete genome sequence. However, shotgun assemblies often contain some gaps particularly in repeat-rich regions. The human genome project, for example, relied on a combination of shotgun sequence and physical mapping to produce contiguous sequence for the length of each arm of each chromosome. Note that because of the highly repetitive nature of centromeric and telomeric DNA, sequencing projects rarely include these heterochromatic regions.

No matter which strategy is used to obtain genomic sequence, the result is a string of millions of A's,C's,G's, andT's. Which of these nucleotides encode proteins, and which of them represent other features of genes and their regulatory elements, or non-coding DNA with no evident function? The process of **genome annotation** relies on computers to define features such a start and stop codons, introns, exons, and splice sites. However, none of the predictions made by these programs are entirely accurate, and must be verified experimentally for any gene of particular importance.

FUNCTIONAL GENOMICS

Having identified putative genes within a genome sequence, how do we determine their function? Techniques of functional genomics are an experimental approach to address this question. One widely used technique in functional genomics is called microarray analysis. Microarrays measure the abundance of mRNA for hundreds or thousands of genes at once. The abundance of mRNA of a particular gene is sometimes correlated with its activity. For example, genes that are involved in neuronal development generally produce more mRNA in brain tissue than in heart tissue. We can therefore use microarrays and related technologies to identify genes whose transcripts are abundant in one tissue sample, but not in another.

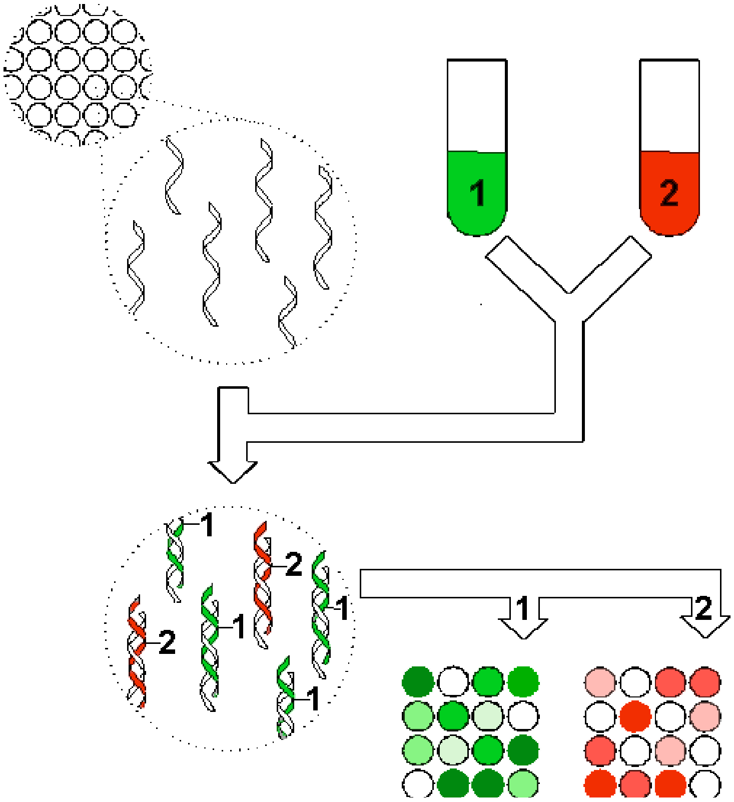


Figure 10.8 An example of a type of DNA microarray. Fluorescently labeled molecules derived from the transcripts of two tissue samples are hybridized immobilized DNA molecules on the surface of an array. The labeled molecules will bind in proportion to their abundance in the original tissue samples. Thus spots on the microarray that are more green than red represent genes that are more abundant in the tissue sample from which green-labeled molecules were derived.

One type of microarray analysis begins with the immobilization of DNA in tens of thousands of small (100µm) spots on the surface of a specially treated microscope slide. Within each spot, there are hundreds of thousands of copies of DNA for the same gene, but each of the spots represents a different gene. The DNA on the surface of this microarray is denatured, making it single-stranded. Then RNA is extracted from two tissue samples that a researcher wants to compare, for example biopsies of cancerous and healthy tissues. The RNA from each sample is labeled with a contrasting fluorescent dye and is usually also reverse transcribed into cDNA, for stability. In our example, all of the transcripts from the tumor sample might be labeled red, and all of the transcripts from the normal sample might be labeled green. These labeled molecules are then hybridized to the surface of the microarray at high stringency, so that they will bind only at spots that contain complementary sequence. Importantly, the labeled cDNA molecules will bind at each spot in proportion to their relative abundance in the original mixture of transcripts from the two samples. Thus, if a gene produced more transcripts in a tumor sample than in a normal sample, the corresponding spot on the microarray would be more red than green. Conversely, a gene that was silenced in a tumor, but was expressed at high levels in a normal tissue would have a spot on the microarray that was more green than red. By analyzing images of fluorescent signals from microarrays, researchers can then identify genes whose expression is highly correlated with one physiological state but not another. Accordingly, microarrays have been used to study everything from psychiatric disorders to wood formation in trees.

SUMMARY:

- Genomics and related technologies differ from other techniques of molecular biology largely because of their scale; they allow many molecules to by studied in parallel.

- DNA sequencing can be applied to either a single gene, or in the case of genomics, to a large number of genes

- Most DNA sequencing relies on the incorporation of dye-labeled terminator molecules which create products that differ in length and end in a known nucleotide. The products can then be separate based on length, and the identity of the last base in each fragment determined from its fluorescence.

- Next-generation sequencing technologies have reduced costs of sequencing further through miniaturization and parallelization.

- Physical maps are ordered sets of clones containing overlapping pieces of DNA, which together represent large pieces of chromosomes.

- Whole genomes may be sequenced using either a clone-by-clone approach, which requires a physical map, or by a shot gun approach, in which small fragments are randomly sequenced.

- Genome analysis does not end after sequence acquisition; various features of the genome including genes (and their introns, exons, etc.) must be identified through a process called annotation.

- Functional genomics techniques including microarray analysis correlate transcript abundance with particular tissue samples. Genes who transcripts are highly abundant under certain biological conditions may cause or respond to that condition.

KEY TERMS:

genome
genomics
-omics
proteomics
transcriptomics
metabolomics
ddNTP
terminator nucleotide
capillary electrophoresis
chromatogram
next-generation sequencing
Illumina
physical map
map-based cloning
BAC
clone-by-clone sequencing
whole genome shotgun
genome annotation
functional genomics
microarray

STUDY QUESTIONS:

10.1 What are the advantages of high-throughput –omics techniques compared to studying a single gene or protein at a time? What are the disadvantages?

10.2 What would the chromatogram from a capillary sequencer look like if you accidentally added only template, primers, polymerase, and fluorescent terminators to your sequencing reaction?

10.3 What are the advantages and disadvantages of clone-by clone vs. whole genome shotgun sequencing?

10.4 How could you use DNA sequencing to identify new species of marine microorganism ?

10.5 Explain how you could use a microarray to identify genes that are affected during drought in wheat.

10.6 A microarray identified 100 genes whose transcripts are highly abundant in tumors, but not in normal tissues. Do any or all of these transcripts cause cancer? Explain your answer.

10.7 How do you ensure that each spot printed on a microarray contains DNA for only one gene?

10.8 What would the spots look like on a microarray after hybridization, if each spot contained a mixture of genes?

10.9 What would the spots look like if the hybridization of green and red labeled DNA was done at low stringency?

Chapter 11 REGULATION OF GENE EXPRESSION

Figure 11.1 The stickleback is an example of an organism for which changes in the regulation of gene expression have been shown to confer a selective advantage in some environments.

Within an individual organism, every cell contains essentially identical genomic sequence. How then do cells develop and function differently from each other? The answer lies in the regulation of **gene expression**. Only certain genes are expressed (i.e. are functionally active) under any particular biological circumstances. Gene expression is regulated at many different stages of the process that converts DNA information into active protein. In the first stage, transcript abundance can be controlled by regulating the rate of initiation of transcription, and the processing and degradation of transcripts. In many cases, higher abundance of a gene's transcripts is correlated with its increased expression. In this chapter, we will focus on **transcriptional regulation**, but be aware that cells also regulate the activity of genes in other ways. For example, by controlling the rate of translation, processing, degradation, and post-translational modification of proteins and protein complexes.

THE *lac* OPERON

Early insights into mechanisms of transcriptional regulation came from studies of *E. coli* by researchers Francois Jacob & Jacques Monod. In *E. coli* and many other bacteria, genes for several different proteins may be encoded on a single transcript in a unit called an **operon**. The genes of an operon share the same transcriptional regulation, but are translated individually. With the exception of *C. elegans* and a few other species, eukaryotes generally do not group genes together in operons.

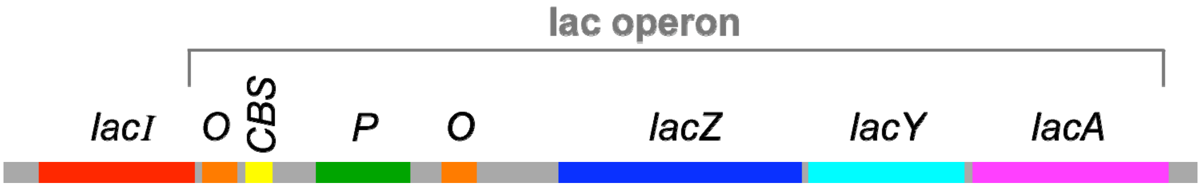


Figure 11.2 Diagram of a segment of an *E. coli* chromosome containing the *lac* operon, as well as the *lacI* coding region. The various genes and *cis*-elements are not drawn to scale.

E. coli encounters many different sugars in its environment. These sugars, such as **lactose** and **glucose**, require different enzymes for their metabolism. Three of the enzymes for lactose metabolism are grouped in the ***lac*** **operon**: ***lacZ***, ***lacY***, and ***lacA*** (Figure 11.2). *LacZ* encodes an enzyme called **β-galactosidase**, which digests lactose into its two constituent sugars: glucose and galactose. *lacY* is a **permease** that helps to transfer lactose into the cell. Finally, *lacA* is a **trans-acetylase**; the relevance of which in lactose metabolism is not entirely clear. *E. coli* activates transcription of the *lac* operon only when lactose is available for it to digest. Presumably, this avoids wasting energy in the synthesis of enzymes for which no substrate is present.

In addition to the three protein-coding genes, the *lac* operon contains short DNA sequences that do not encode proteins, but are instead binding sites for proteins involved in transcriptional regulation of the operon. In the *lac* operon, these sequences are called **P (promoter)**, **O (operator)**, and **CBS (CAP-binding site)**. Collectively, sequence elements such as these are called ***cis*-elements** because they must be located on the same piece of DNA as the genes they regulate. On the other hand, the proteins that bind to these *cis*-elements are called ***trans*-regulators** because (as diffusible molecules) they do not necessarily need to be encoded on the same piece of DNA as the genes they regulate.

lacI IS AN ALLOSTERICALLY REGULATED REPRESSOR

One of the major *trans*-regulators of the *lac* operon is encoded by ***lacI.*** Four identical molecules of *lacI* proteins assemble together to form a **homotetramer** called a **repressor** (Figure 11.3). This repressor binds to two operator sequences adjacent to the promoter of the *lac* operon. Binding of the repressor prevents RNA polymerase from binding to the promoter (Figure 11.4). Therefore, the operon cannot be transcribed when the operator is occupied by a repressor.

Figure 11.3 Structure of *lacI* homotetramer bound to DNA

Besides its ability to bind to specific DNA sequences at the operator, another important property of the *lacI* protein is its ability to bind to lactose. When lactose is bound to *lacI*, the shape of the protein changes in a way that prevents it from binding to the operator. Therefore, in the presence of lactose, RNA polymerase is able to bind to the promoter and transcribe the *lac* operon, leading to a moderate level of expression of the *lacZ*, *lacY*, and *lacA* genes. Proteins such as *lacI* that change their shape and functional properties after binding to a ligand are said to be regulated through an **allosteric** mechanism. The role of *lacI* in regulating the *lac* operon is summarized in Figure 11.4.

CAP IS AN ALLOSTERIC ACTIVATOR OF THE *LAC* OPERON

A second aspect of *lac* operon regulation is conferred by a *trans*-factor called **cAMP binding protein** (**CAP**, Figure 11.5). CAP is another example of an allosterically regulated *trans*-factor. Only when the CAP protein is bound to cAMP can another part of the protein bind to a specific *cis*-element within the *lac* promoter called the **CAP binding sequence (CBS)**. CBS is located very close to the promoter (P). When CAP is bound to at CBS, RNApol is better able to bind to the promoter and initiate transcription. Thus, the presence of cAMP ultimately leads to a further increase in *lac* operon transcription.

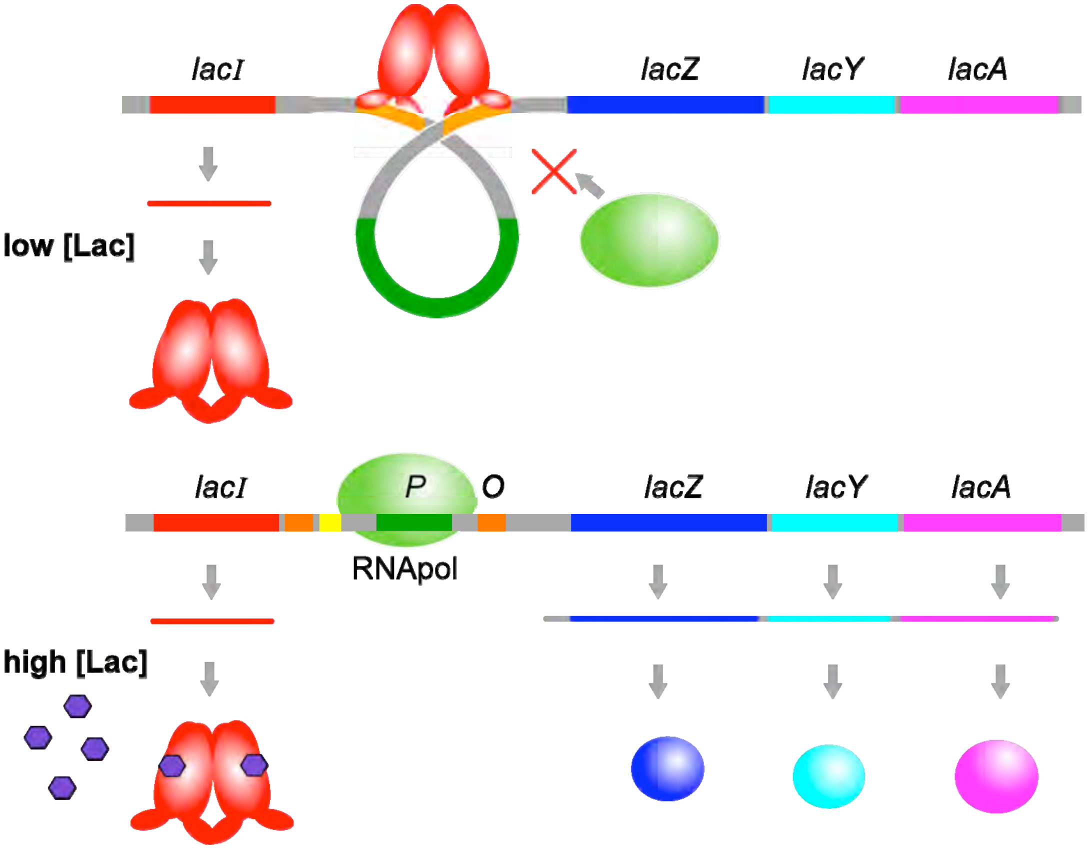


Figure 11.4 When the concentration of lactose [Lac] is low, *lacI* tetramers bind to operator sequences (O), thereby blocking binding of RNApol (green) to the promoter (P). Alternatively, when [Lac] is high, lactose binds to *lacI*, preventing the repressor from attaching to O, and allowing transcription by RNApol.

The physiological significance of regulation by cAMP becomes more obvious in the context of the following information. The concentration of cAMP is inversely proportional to the abundance of glucose: when glucose concentrations are low, an enzyme called **adenylate cyclase** is able to produce cAMP from ATP. Evidently, *E. coli* prefers glucose over lactose, and so expresses the *lac* operon at high levels only when glucose is absent and lactose is present. This provides another layer of logical control of *lac* operon expression: only in the presence of lactose, and in the absence of glucose is the operon expressed at its highest levels.

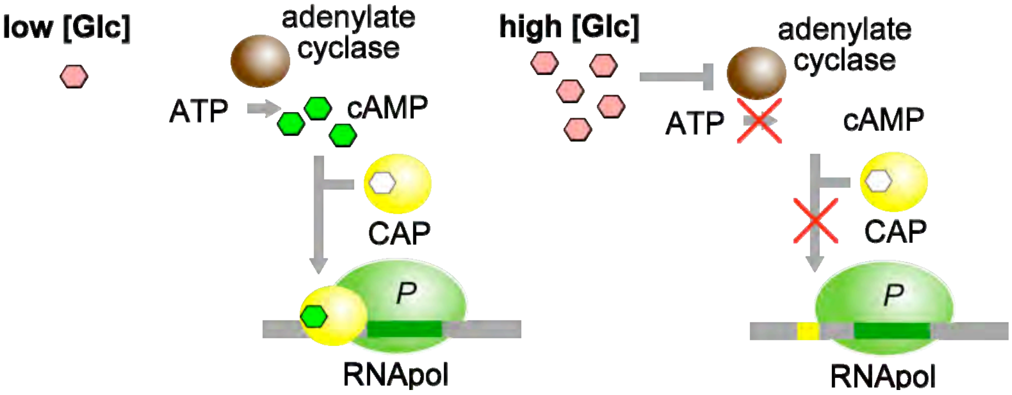


Figure 11.5 CAP, when bound to cAMP, helps RNApol to bind to the *lac* operon. cAMP is produced only when glucose [Glc] is low.

MUTANTS OF THE *lac* OPERON

The *lac* operon and its regulators were first characterized by studying mutants of *E. coli* that exhibited various abnormalities in lactose metabolism. Some mutants expressed the *lac* operon genes constitutively, meaning the operon was expressed whether lactose was present in the medium or not. One example of a **constitutive** mutant is O^c, in which a mutation in an operator sequence prevents *lacI* from recognizing and binding to the operator. Thus, in O^c mutants, *lacZ*, *lacY*, and *lacA* are expressed whether lactose is present or not. A mutant allele of *lacI* (e.g. I^-) that prevents it from binding to DNA is also a constitutive expresser of the *lac* operon. On the other hand, a mutant of *lacI* called I^s prevents it from binding lactose, so that the mutant represses the *lac* operon whether lactose is present or not.

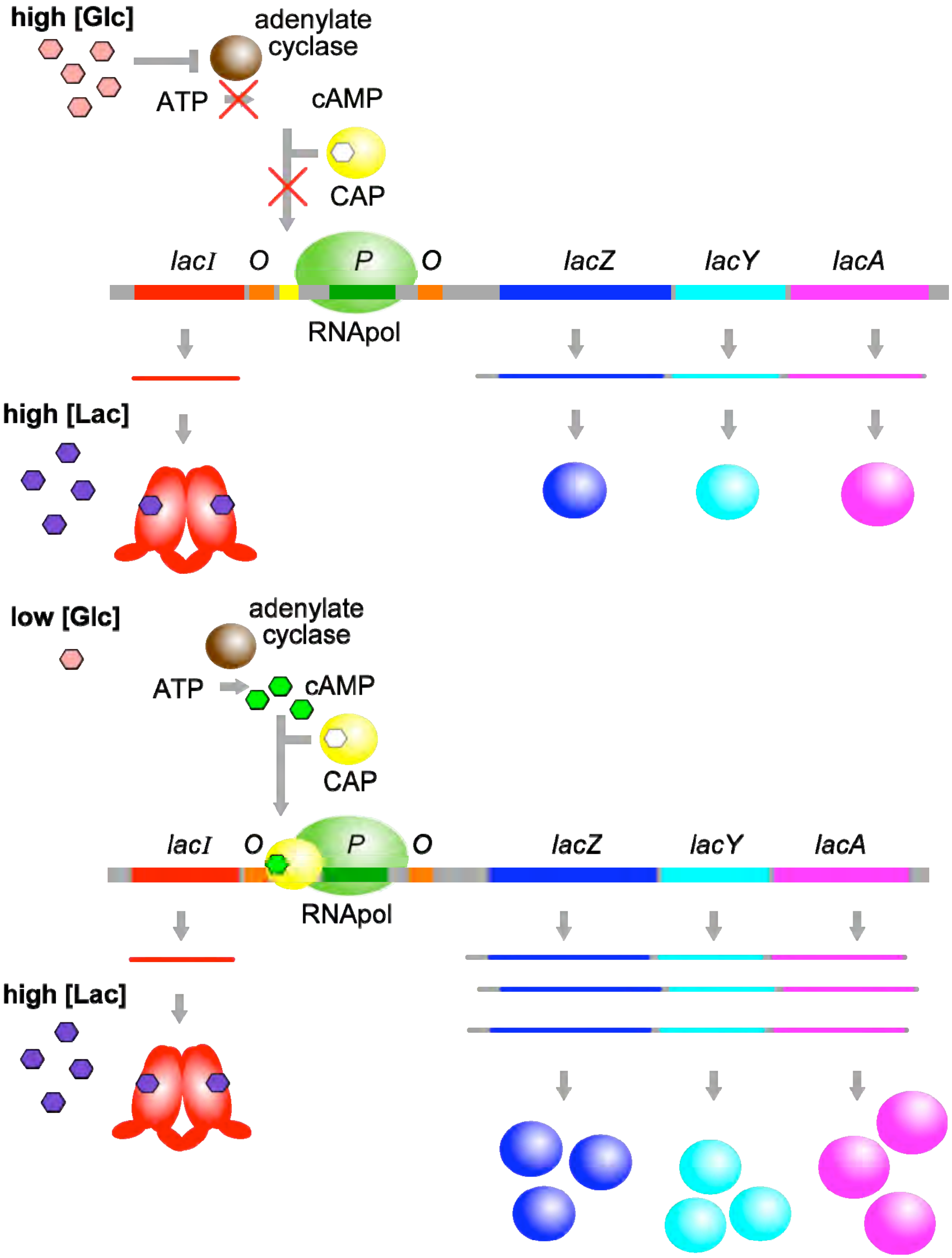


Figure 11.6 When glucose [Glc] and lactose [Lac] are both high, the *lac* operon is transcribed at a moderate level, because CAP (in the absence of cAMP) is unable to bind to its corresponding *cis*-element (yellow) and therefore cannot help to stabilize binding of RNApol at the promoter. Alternatively, when [Glc] is low, and [Lac] is high, CAP and cAMP can bind near the promoter and increase further the transcription of the *lac* operon.

EUKARYOTIC GENE REGULATION

As we have seen in prokaryotes, transcriptional regulation in eukaryotes involves both *cis*-elements and *trans*-factors. A typical eukaryotic gene, including several types of *cis*-elements, is shown in Figure 11.7. RNA polymerase binds to the gene at its promoter. In eukaryotes, RNApol is part of a large protein complex that includes additional proteins that bind to one or more specific *cis*-elements near the promoter, including **GC boxes**, **CAAT boxes**, **and TATA boxes**. High levels of transcription require both the presence of this protein complex at the promoter, as well as their interaction with other *trans*-factors described below. The approximate position of these elements relative to the **transcription start site** ($^{+}1$) is shown in Figure 11.7, but it should be emphasized that the distance between any of these elements and the transcription start site can vary.

Even more variation is observed in the position and orientation of the second major type of *cis*-regulatory element in eukaryotes, which are called **enhancers**. Regulatory *trans*-factor proteins called **transcription factors** bind to enhancers, then while still bound to DNA, these proteins interact with RNApol and other proteins at the promoter to activate transcription. Specific transcription factors recognize the DNA sequence of specific enhancers to activate gene expression in specific circumstances. Because DNA is a flexible molecule, enhancers can be located near or far, and either upstream or downstream, from the promoter (Figure 11.8).

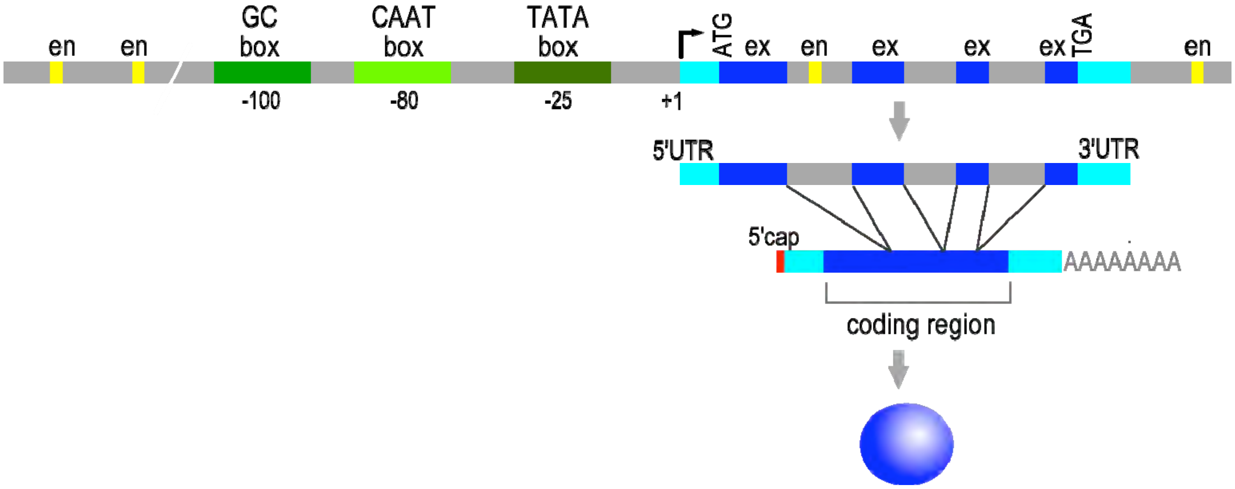


Figure 11.7 Structure of a typical eukaryotic gene. RNA polymerase binding may involve one or more *cis*-elements within the region of a promoter (green boxes). Enhancers (yellow boxes) may be located any distance upstream or downstream of the promoter.

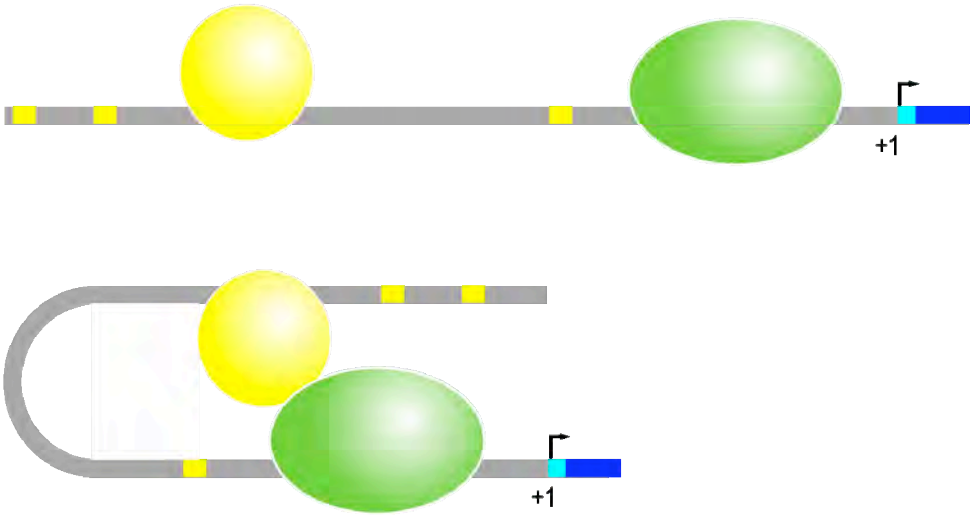


Figure 11.8 A transcription factor (yellow) bound to an enhancer that is located far from a promoter. Because of the flexibility of the DNA molecule, the transcription factor and RNApol are able to interact physically, even though the *cis*-elements to which they are bound are located far apart. In eukaryotic cells, RNApol is actually part of a large complex of proteins (not shown here) that assembles at the promoter.

The *yellow* gene of Drosophila demonstrates the modular nature of enhancers. This gene is part of a pathway that produces a black pigment (recall that genes are sometimes named after their mutant phenotypes). Each tissue in Drosophila that makes the black pigment produces a specific transcription factor that binds to a corresponding enhancer in *yellow* to activate its transcription (Figure 11.9). Thus, specific combinations of *cis*-elements and *trans*-factors control the tissue-specific expression of genes. This is typical of the transcriptional activation of almost any eukaryotic gene: specific transcription factors activate (or in some cases repress) the transcription of target genes under specific conditions.

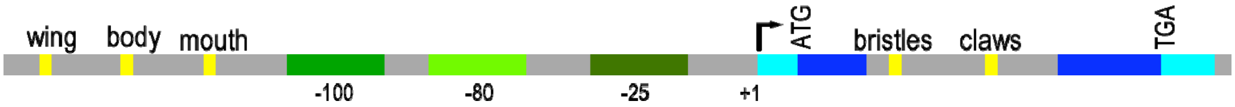


Figure 11.9 Tissue-specific *cis*-regulatory elements within a simplified representation of the *yellow* gene of Drosophila.

REGULATORY ELEMENTS IN EVOLUTION

Mutations can occur in both *cis*-elements and *trans*-factors, resulting in altered patterns of gene expression. If these altered patterns of gene expression produce a selective advantage (or at least do not produce a major disadvantage), they may be maintained and even contribute to evolution of new species.

The three-spined **stickleback** (Figure 11.1) provides an example of natural selection of a mutation in a *cis*-regulatory element. This fish occurs in two forms: populations that naturally inhabit deep, open water have a spiny pelvic fin that deters larger predator fish from feeding on the stickleback. However, populations of stickleback from shallow water environments lack this spiny pelvic fin. In shallow water, it appears that a long, spiny pelvic fin would be a disadvantage because it frequently contacts the sediment at the bottom of the pond and allows parasitic insects in the sediment to invade the stickleback. Researchers compared genomic DNA fragments of individuals from both deep and shallow water environments as shown in Figure 11.10. They observed that in embryos from the deep-water population, a gene called *Pitx* was expressed in several groups of cells, including those that developed into the pelvic fin. Embryos from the shallow-water population expressed *Pitx* in the same groups of cells as the other population, with an important exception: *Pitx* was not expressed in the pelvic fin **primordium** in the shallow-water population. Further analysis showed that the absence of *Pitx* gene expression from the developing pelvic fin of shallow-water stickleback was due to the absence of a particular enhancer element upstream of *Pitx*.

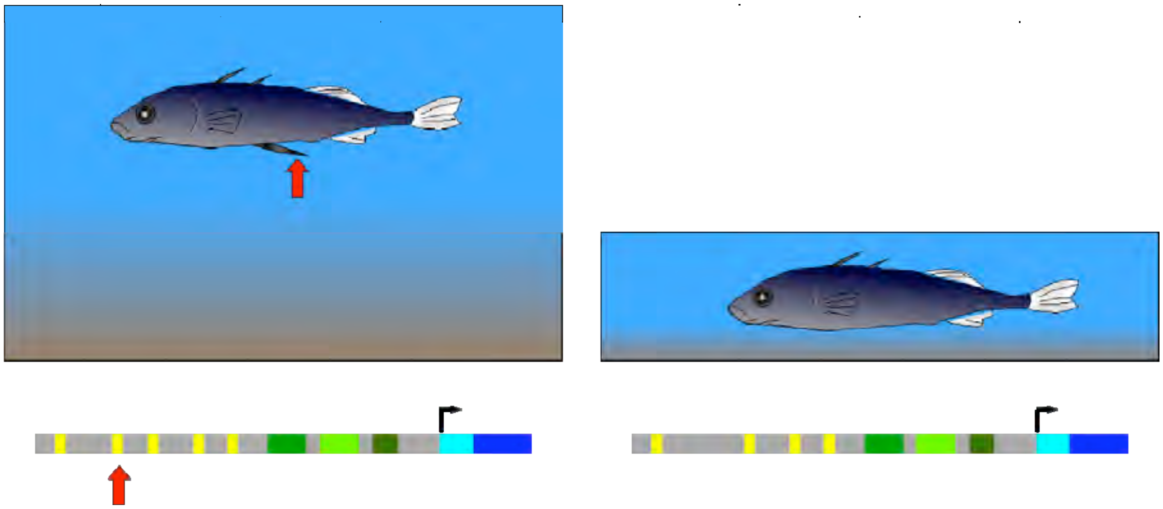


Figure 11.10 Development of a large, spiny pelvic fin in stickleback depends on the presence of a particular enhancer element upstream of a gene called *Pitx*. Mutants lacking this element (and therefore the large pelvic fin) have been selected for in shallow-water environments.

OTHER REGULATORS OF TRANSCRIPTION

Despite the simplified way in which we often represent DNA in figures such as those in this chapter, DNA is almost never entirely separated from chromatin proteins during interphase, and histones remain associated with the DNA at many positions along the molecule even during transcription. The rate of transcription is therefore also controlled by the accessibility of DNA to RNApol and regulatory proteins. If the chromatin that contains a particular gene is highly compacted, it is unlikely that the gene will be transcribed, even if all of the necessary *cis*- and *trans*- factors are present. Cells regulate the local structure of chromatin through the action of proteins called **chromatin remodeling** proteins. These include enzymes that add or remove chemical tags such as methyl or acetyl groups. **Acetylated** histones, for example, tend to be associated with actively transcribed genes, whereas deacetylation can causes associated genes to be silenced. Methylation of DNA itself also regulates transcription. Cytosine bases, particularly when followed by a guanine (**CpG sites**) are important targets for methylation. Methylated cytosine within clusters of CpG sites is often associated with transcriptionally inactive DNA. Reversible histone modification and DNA methylation are thus another layer by which eukaryotic cells control the transcription of specific genes.

Interestingly, some apparent changes in gene expression are heritable. For example, the grandchildren of famine victims are known to have lower birth weight than children without a family history of famine. The heritability of altered states of gene expression is surprising, since gene regulation does not usually involve changes in the sequence of DNA. The term **epigenetics** describes any heritable change in phenotype that is associated with a change in something other than chromosomal DNA sequence. Some epigenetic information is inherited transgenerationally, while in other cases, the epigenetic state is inherited following mitotic divisions, but not following meiosis. The basis of at least some types of epigenetic inheritance appears to be the replication of patterns of histone modifications and DNA methylation in parallel with the replication of the primary DNA sequence. It is becoming clear that epigenetics is an important part of biology, and can serve as a type of cellular memory, sometimes within an individual, or sometimes across a few generations.

Figure 11.11 A winter wheat crop (green) in early spring in the English countryside.

VERNALIZATION IS AN EXAMPLE OF EPIGENETIC CELLULAR MEMORY

Many plant species in temperate regions are **winter annuals**, meaning that their seeds germinate in the late summer, and grow vegetatively through early fall before entering a dormant phase during the winter, often under a cover of snow. In the spring, the plant resumes growth and is able to produce seeds before other species that germinated in the spring. In order for this life strategy to work, the winter annual must not resume growth or start flower production until winter has ended. **Vernalization** is the name given to the requirement to experience a long period of cold temperatures prior to flowering.

How does a plant sense that winter has passed? The signal for resuming growth cannot simply be warm air temperature, since occasional warm days, followed by long periods of freezing, are common in temperate climates. Researchers have discovered that winter annuals use epigenetic mechanisms to sense and "remember" that winter has occurred

Fortunately for the researchers who were interested in vernalization, some varieties of Arabidopsis are winter annuals. Through mutational analysis of Arabidopsis, researchers found that a gene called *FLC* (*FLOWERING LOCUS C*) is a repressor of the transcription of several of the genes involved in early stages of flowering (Figure 11.12). In the fall and under other warm conditions, the histones associated with *FLC* are acetylated and so *FLC* is transcribed at high levels; expression of flowering genes is therefore entirely repressed. However, in response to cold temperatures, enzymes gradually deacetylate the histones associated with *FLC*. The longer the cold temperatures persist, the more acetyl groups are removed from the *FLC*-associated histones, until finally the *FLC* locus is no longer transcribed and the flowering genes are free to respond to other environmental and hormonal signals that induce flowering later in the spring. Because the deacetylated state of *FLC* is inherited as cells divide and the plant grows in the early spring, this is an example of a type of cellular memory mediated by an epigenetic mechanism

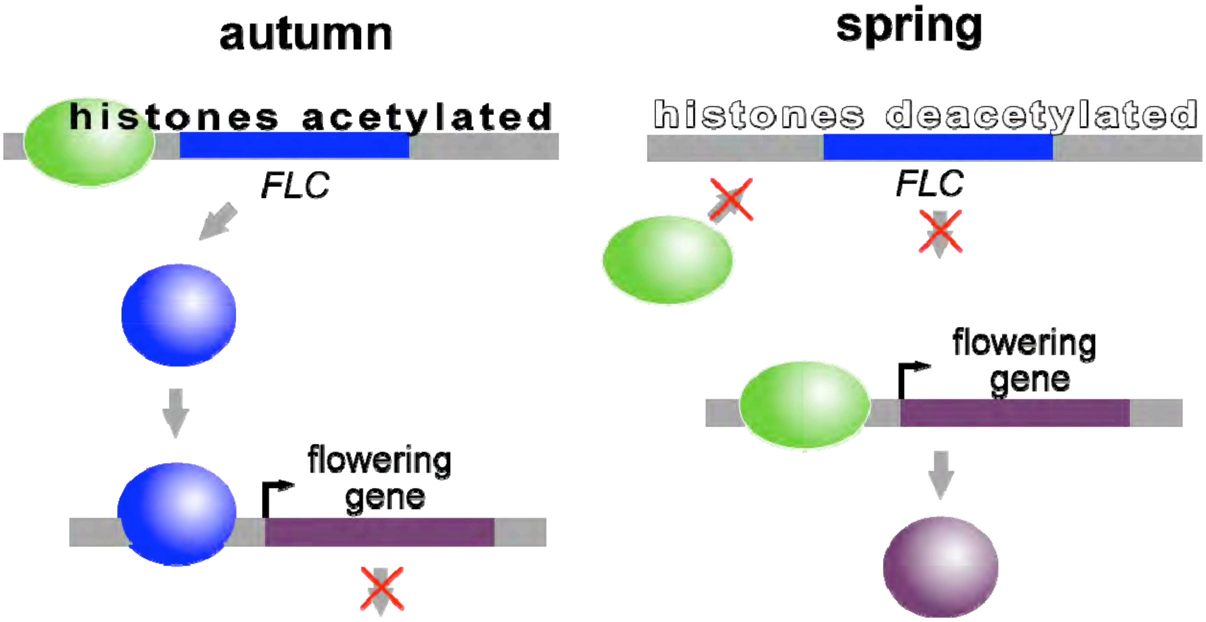


Figure 11.12 In the autumn, histones associated with FLC are acetylated, allowing this repressor of flowering genes to be expressed. During winter, enzymes progressive deacetylate FLC, preventing it from being expressed, and therefore allowing flowering genes to respond to other signals that induce flowering.

SUMMARY:

- Regulation of gene expression is essential to the normal development and efficient functioning of cells

- Gene expression may be regulated by many mechanisms, including those affecting transcript abundance, protein abundance, and post-translational modifications

- Regulation of transcript abundance may involve controlling the rate of initiation and elongation of transcription, as well as transcript splicing, stability, and turnover

- The rate of initiation of transcription is related to the presence of RNA polymerase and associated proteins at the promoter.

- RNApol may be blocked from the promoter by repressors, or may be recruited or stabilized at the promoter by other proteins including transcription factors

- The *lac* operon is an important paradigm demonstrating both positive and negative regulation through allosteric effects on *trans*-factors.

- In eukaryotes, *cis*-elements that are usually called enhancers bind to specific *trans*-factors to regulate transcriptional initiation.

- Enhancers may be modular, with each enhancer and its transcription factor regulating a distinct component of a gene's expression pattern, as in the *yellow* gene.

- Sticklebacks provide examples of recent evolutionary events in which mutation of an enhancer produced a change in morphology and a selective advantage.

- Chromatin structure, including reversible modifications such as acetylation of histones, and DNA methylation CpG sites also regulates the initiation of transcription.

- Chromatin modifications or DNA methylation of some genes are heritable over many mitotic, and sometimes even meiotic divisions.

- Heritable changes in phenotype that do not result from a change in DNA sequence are called epigenetic. Many epigenetic phenomena involve regulation of gene expression by chromatin modification and/or DNA methylation.

KEY TERMS:

gene expression
transcriptional regulation
operon
lactose
glucose
lac operon
lacZ
lacY
lacA
galactosidase
permease
trans-acetylase
P
promoter
O
operator
CBS
CAP-binding site
cis-elements
trans-regulators
lacI
homotetramer
repressor
allosteric
cAMP binding protein
CAP
CAP binding sequence
CBS
adenylate cyclase
constitutive
O^c
I^-
I^s
GC boxes
CAAT boxes
TATA boxes
transcription start site
enhancers
transcription factors
stickleback
primordium
chromatin remodeling
acetylation
deacetylation
methylation
CpG
epigenetics
winter annual
vernalization
FLC

STUDY QUESTIONS:

11.1 List all the mechanisms that can be used to regulate gene expression in eukaryotes.

11.2 With respect to the expression of β-galactosidase, what would be the phenotype of each of the following strains of *E. coli*?

a) I^+, O^+, Z^+, Y^+ (no glucose, no lactose)
b) I^+, O^+, Z^+, Y^+ (no glucose, high lactose)
c) I^+, O^+, Z^+, Y^+ (high glucose, no lactose)
d) I^+, O^+, Z^+, Y^+ (high glucose, high lactose)
e) I^+, O^+, Z^-, Y^+ (no glucose, no lactose)
f) I^+, O^+, Z^-, Y^+ (high glucose, high lactose)
g) I^+, O^+, Z^+, Y^- (high glucose, high lactose)
h) I^+, Oc, Z^+, Y^+ (no glucose, no lactose)
i) I^+, Oc,Z^+, Y^+ (no glucose, high lactose)
j) I^+, Oc, Z^+, Y^+ (high glucose, no lactose)
k) I^+, Oc, Z^+, Y^+ (high glucose, high lactose)
l) I^-, O^+, Z^+, Y^+ (no glucose, no lactose)
m) I^-, O^+, Z^+, Y^+ (no glucose, high lactose)
n) I^-, O^+, Z^+, Y^+ (high glucose, no lactose)
o) I^-, O^+, Z^+, Y^+ (high glucose, high lactose)
p) I^s, O^+, Z^+, Y^+ (no glucose, no lactose)
q) I^s, O^+, Z^+, Y^+ (no glucose, high lactose)
r) I^s, O^+, Z^+, Y^+ (high glucose, no lactose)
s) I^s, O^+, Z^+, Y^+ (high glucose, high lactose)

11.3 In the *E. coli* strains listed below, some genes are present on both the chromosome, and the extrachromosomal F-factor episome. The genotypes of the chromosome and episome are separated by a slash. What will be the β-galactosidase phenotype of these strains? All of the strains are grown in media that lacks glucose.

a) $I^+, O^+, Z^+, Y^+ / O^-, Z^-, Y^-$ (high lactose)
b) $I^+, O^+, Z^+, Y^+ / O^-, Z^-, Y^-$ (no lactose)
c) $I^+, O^+, Z^-, Y^+ / O^-, Z^+, Y^+$ (high lactose)
d) $I^+, O^+, Z^-, Y^+ / O^-, Z^+, Y^+$ (no lactose)
e) $I^+, O^+, Z^-, Y^+ / I^-, O^+, Z^+, Y^+$ (high lactose)
f) $I^+, O^+, Z^-, Y^+ / I^-, O^+, Z^+, Y^+$ (no lactose)
g) $I^-, O^+, Z^+, Y^+ / I^+, O^+, Z^-, Y^+$ (high lactose)
h) $I^-, O^+, Z^+, Y^+ / I^+, O^+, Z^-, Y^+$ (no lactose)
i) $I^+, Oc, Z^+, Y^+ / I^+, O^+, Z^-, Y^+$ (high lactose)
j) $I^+, Oc, Z^+, Y^+ / I^+, O^+, Z^-, Y^+$ (no lactose)
k) $I^+, O^+, Z^-, Y^+ / I^+, Oc, Z^+, Y^+$ (high lactose)
l) $I^+, O^+, Z^-, Y^+ / I^+, Oc, Z^+, Y^+$ (no lactose)
m) $I^+, O^+, Z^-, Y^+ / I^s, O^+, Z^+, Y^+$ (high lactose)
n) $I^+, O^+, Z^-, Y^+ / I^s, O^+, Z^+, Y^+$ (no lactose)
o) $I^s, O^+, Z^+, Y^+ / I^+, O^+, Z^-, Y^+$ (high lactose)
p) $I^s, O^+, Z^+, Y^+ / I^+, O^+, Z^-, Y^+$ (no lactose)

11.4 What genotypes of *E. coli* would be most useful in demonstrating that the *lacO* operator is a *cis*-acting regulatory factor?

11.5 What genotypes of E. coli would be useful in demonstrating that the *lacI* repressor is a *trans*-acting regulatory factor?

11.6 What would be the effect of the following loss-of-function mutations on the expression of the lac operon?

a) loss-of-function of adenylate cyclase
b) loss of DNA binding ability of CAP
c) loss of cAMP binding ability of CAP
d) mutation of CAP binding site (CBS) *cis*-element so that CAP could not bind

11.7 How are eukaryotic and prokaryotic gene regulation systems similar? How are they different?

11.8 Deep-water sticklebacks that are heterozygous for a loss-of-function mutation in the coding region of *Pitx* look just like homozygous wild-type fish from the same population. What phenotype or phenotypes would be observed if a wild-type fish from a deep-water population mated with a wild-type fish from a shallow-water population?

11.9 Some varieties of Arabidopsis, including those adopted for lab use, do not require vernalization before flowering. How might these varieties have evolved?

11.10 Histone deacetylase (HDAC) is an enzyme involved in gene regulation. What might be the phenotype of a winter annual plant that lacked HDAC function?

Chapter 12 CANCER GENETICS

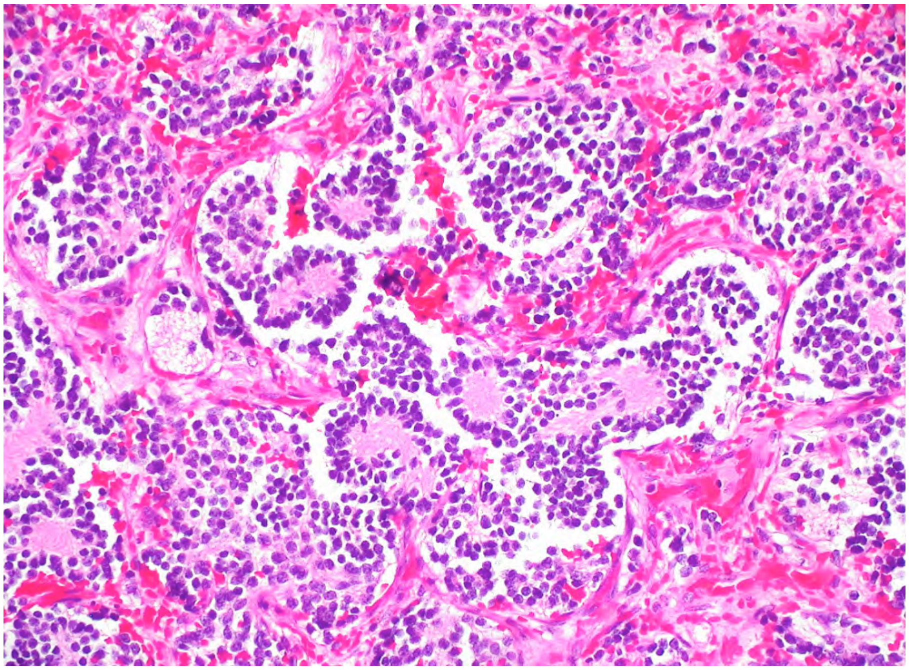


Figure 12.1 Stained histological section of a neuroblastoma in an adrenal gland. Photo Credit Ed Uthman, M.D.

Cancer is any one of a group of diseases that exhibit uncontrolled growth, invasion of adjacent tissues, and sometimes **metastasis** (the movement of cancer cells through the blood or lymph). In cancer cells, the regulatory mechanisms that control cell division and limit abnormal growth have been disrupted, usually by the accumulation of several mutations. Cancer is therefore essentially a genetic disease. Although some cancer-related mutations may be heritable, most cancers are sporadic, meaning they arise from new mutations that occur in the individual who has the disease. In this chapter we will examine the connection between cancer and genes.

CANCER CELL BIOLOGY

Cancer is a progressive disease that usually begins with increased frequency of cell division (Figure 12.2). Under the microscope, this may be detectable as increased cellular and nuclear size, and an increased proportion of cells undergoing mitosis. As the disease progresses, cells typically lose their normal shape and tissue organization. Tissues with increased cell division and abnormal tissue organization exhibit **dysplasia**. Eventually a tumor develops, which can grow rapidly and expand into adjacent tissues. As cellular damage accumulates and additional control mechanisms are lost, some cells may break free of the primary tumor, pass into the blood or lymph system, and be transported to another organ, where they develop into new tumors (Figure 12.3). The early detection of tumors is important so that they can be treated or removed before the onset of metastasis.

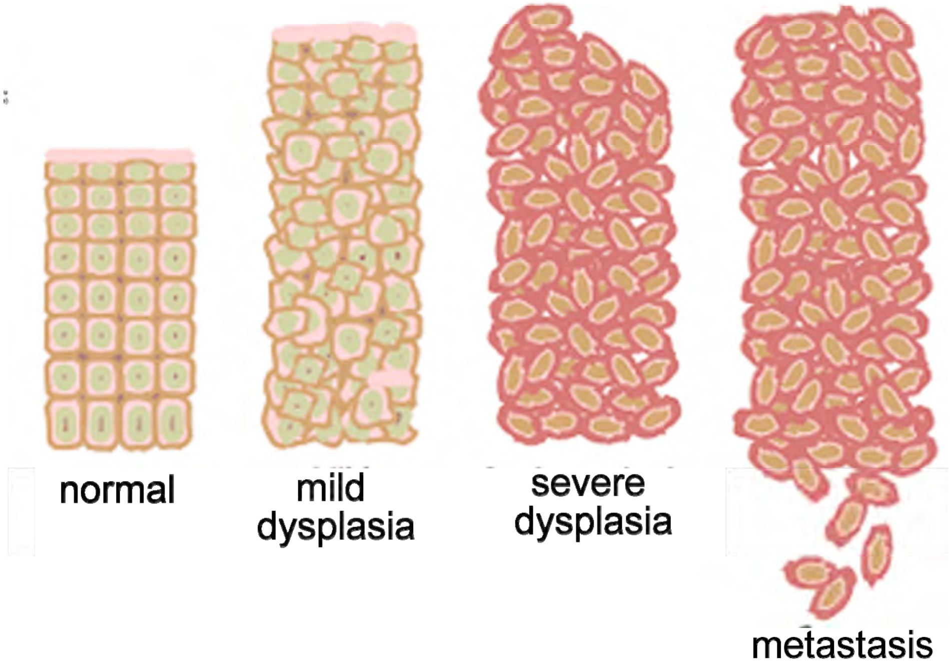


Figure 12.2 Progressive increases in cell division and abnormal cell morphology associated with cancer

MUTAGENS AND CARCINOGENS

A **carcinogen** is any agent that directly increases the incidence of cancer. Most, but not all carcinogens are mutagens. Carcinogens that do not directly damage DNA include substances that accelerate cell division, thereby leaving less opportunity for cell to repair induced mutations, or errors in replication. Carcinogens that act as mutagens may be biological, physical, or chemical in nature, although the term is most often used in relation to chemical substances.

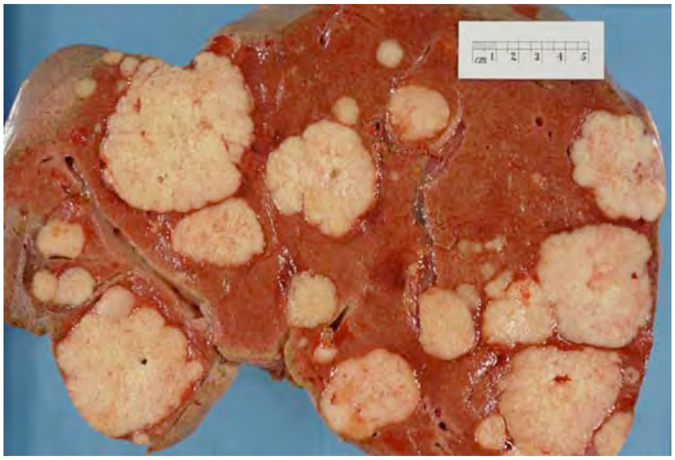


Figure 12.3 Secondary tumors (white) develop in the liver from cells of a metastatic pancreatic cancer.

.

Human Papilloma Virus (**HPV**, Figure 12.6) is an example of a biological carcinogen. Almost all cervical cancers begin with infection by HPV, which contains genes that disrupt the normal pattern of cell division within the host cell. Any gene that leads to an uncontrolled increase in cell division is called an **oncogene**. The HPV E6 and E7 genes are considered oncogenes because they inhibit the host cell's natural tumor suppressing proteins (include p53, described below). The product of the E5 gene mimics the host's own signals for cell division, and these and other viral gene products may contribute to dysplasia, which is detected during a Pap smear (Figure 12.5). Detection of abnormal cell morphology in a Pap smear is not necessarily evidence of cancer. It must be emphasized again that cells have many regulatory mechanisms to limit division and growth, and for cancer to occur, each of these mechanisms must be disrupted. This is one reason why only a minority of individuals with HPV infections ultimately develop cancer. Although most HPV-related cancers are cervical, HPV infection can also lead to cancer in other tissues, in both women and men.

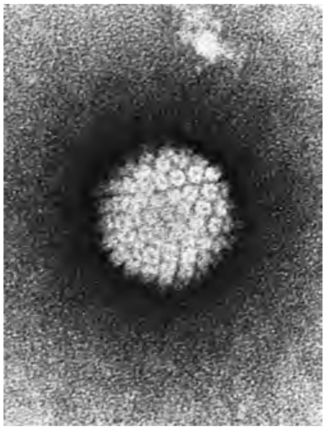


Figure 12.4 Electron micrograph of HPV.

Radiation is a well-known physical carcinogen, because of its potential to induce DNA damage within the body. The most damaging type of radiation is **ionizing**, meaning waves or particles with sufficient energy to strip electrons from the molecules they encounter, including DNA or molecules that can subsequently react with DNA. Ionizing radiation, which includes x-rays, gamma rays, and some wavelengths of ultraviolet rays, is distinct from the non-ionizing radiation of microwave ovens, cell phones, and radios. As with other carcinogens, mutation of multiple, independent genes that normally regulate cell division is required before cancer develops.

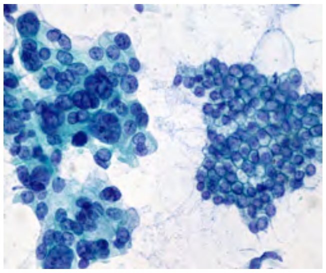


Figure 12.5 Dysplastic (left) and normal (right) cells from a Pap smear. Photo Credit Ed Uthman, M.D.

.

Chemical carcinogens (Table 12.1) can be either natural or synthetic compounds that, based on animal feeding trials or **epidemiological** (i.e. human population) studies, increase the incidence of cancer. The definition of a chemical as a carcinogen is problematic for several reasons. Some chemicals become carcinogenic only after they are metabolized into another compound in the body; not all species or individuals may metabolize chemicals in the same way. Also, the carcinogenic properties of a compound are usually dependent on its dose. It can be difficult to define a relevant dose for both lab animals and humans. Nevertheless, when a correlation between cancer incidence and chemical exposure is observed, it is usually possible to find ways to reduce exposure to that chemical.

ONCOGENES

The control of cell division involves many different genes. Some of these genes act as signaling molecules to activate normal progression through the cell cycle. One of the pre-requisites for cancer occurs when one or more of these activators of cell division become mutated.

Table 12.1 Known human carcinogens (International Agency for Research on Cancer "Carcinogenic to humans" (Group 1)). Source: American Cancer Society.

Exposure circumstances

- Aluminum production
- Arsenic in drinking-water
- Auramine production
- Boot and shoe manufacture and repair
- Chimney sweeping
- Coal gasification , Coal-tar distillation, Coke production
- Furniture and cabinet making
- Hematite mining (underground) with exposure to radon
- Involuntary smoking (exposure to secondhand or 'environmental' tobacco smoke)
- Iron and steel founding
- Isopropyl alcohol manufacture (strong-acid process)
- Magenta production
- Painter (occupational exposure as a)
- Paving and roofing with coal-tar pitch
- Rubber industry
- Strong-inorganic-acid mists containing sulfuric acid (occupational exposure to)
- Tobacco smoking and tobacco smoke

Mixtures

- Aflatoxins (naturally occurring mixtures of)
- Alcoholic beverages
- Areca nut
- Betel quid
- Coal-tar pitches
- Coal-tars
- Herbal remedies containing plant species of the genus Aristolochia
- Household combustion of coal, indoor emissions from
- Mineral oils, untreated and mildly treated
- Phenacetin, analgesic mixtures containing
- Salted fish (Chinese-style)
- Shale-oils
- Soots
- Tobacco, smokeless
- Wood dust

Agents and groups of agents

- 4-Aminobiphenyl
- Arsenic and arsenic compounds (Note: This applies to the group of compounds as a whole)
- Asbestos
- Azathioprine
- Benzene
- Benzidine
- Benzo[a]pyrene
- Beryllium and beryllium compounds
- N,N-Bis(2-chloroethyl)-2-naphthylamine (Chlornaphazine)
- Bis(chloromethyl)ether and chloromethyl methyl ether (technical-grade)
- 1,3-Butadiene
- 1,4-Butanediol dimethanesulfonate (Busulphan; Myleran)
- Cadmium and cadmium compounds

- Chlorambucil
- 1-(2-Chloroethyl)-3-(4-methylcyclohexyl)-1-nitrosourea (Methyl-CCNU; Semustine)
- Chromium[VI]
- Ciclosporin
- Cyclophosphamide
- Diethylstilbestrol
- Dyes metabolized to benzidine
- Epstein-Barr virus
- Erionite
- Estrogen-progestogen menopausal therapy (combined)
- Estrogen-progestogen oral contraceptives (combined) (Note: There is also convincing evidence in humans that these agents confer a protective effect against cancer in the endometrium and ovary)
- Estrogen therapy, postmenopausal
- Ethanol in alcoholic beverages
- Ethylene oxide
- Etoposide in combination with cisplatin and bleomycin
- Formaldehyde
- Gallium arsenide
- Helicobacter pylori (infection with)
- Hepatitis B virus (chronic infection with)
- Hepatitis C virus (chronic infection with)
- Human immunodeficiency virus type 1 (infection with)
- Human papillomavirus types 16, 18, 31, 33, 35, 39, 45, 51, 52, 56, 58, 59 and 66
- Human T-cell lymphotropic virus type I
- Melphalan
- 8-Methoxypsoralen (Methoxsalen) plus ultraviolet A radiation
- Methylenebis(chloroaniline) (MOCA)
- MOPP and other combined chemotherapy including alkylating agents
- Mustard gas (Sulfur mustard)
- 2-Naphthylamine
- Neutrons
- Nickel compounds
- N'-Nitrosonornicotine (NNN) and 4-(N-Nitrosomethylamino)-1-(3-pyridyl)-1-butanone (NNK)
- *Opisthorchis viverrini* (infection with)
- Oral contraceptives, combined estrogen-progestogen
- Oral contraceptives, sequential
- Phosphorus-32, as phosphate
- Plutonium-239 and its decay products (may contain plutonium-240 and other isotopes), Radioiodines, short-lived isotopes, including iodine-131
- Radionuclides, a-particle-emitting, internally deposited
- Radionuclides, b-particle-emitting, internally deposited
- Radium-224, Radium-226, Radium-228, Radium-222 and their decay products
- *Schistosoma haematobium* (infection with)
- Silica, crystalline (inhaled in the form of quartz or cristobalite from occupational sources)
- Solar radiation
- Talc containing asbestiform fibres
- Tamoxifen (Note: tamoxifen reduces the risk of contralateral breast cancer)
- 2,3,7,8-Tetrachlorodibenzo-para-dioxin
- Thiotepa
- Thorium-232 and its decay productsortho-Toluidine
- Treosulfan
- Vinyl chloride
- X- and Gamma (g)-radiation

The mutation may involve a change in the coding sequence of the protein, so that it is more active than normal, or a change in the regulation of its expression, so that it is produced at higher levels than normal, or persists in the cell longer than normal. Genes that are a part of the normal regulation of cell division, but which after mutation contribute to cancer, are called **proto-oncogenes**. Once a proto-oncogene has been abnormally activated by mutation, it is called an oncogene. More than 100 genes have been defined as proto-oncogenes. These include genes at almost every step of the signaling pathways that normally induce cell to divide, including growth factors, **receptors**, **signal transducers**, and transcription factors.

ras is an example of a proto-oncogene. *ras* acts as a switch within signal transduction pathways, including the regulation of cell division. When a receptor protein receives a signal for cell division, the receptor activates *ras*, which in turn activates other signaling components, ultimately leading to activation of genes involved in cell division. Certain mutations of the *ras* sequence causes it to be in a permanently active form, which can lead to constitutive activation of the cell cycle. This mutation is dominant as are most oncogenes.

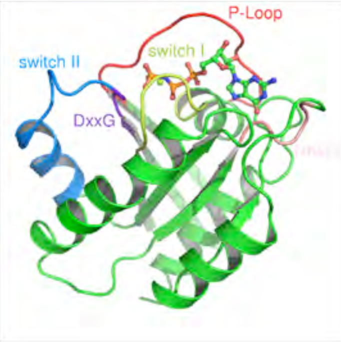


Figure 12.6 Structure of the *ras* protein.

TUMOR SUPPRESSOR GENES

More than 30 genes are classified as **tumor suppressors**. The normal functions of these genes include repair of DNA, induction of programmed cell death (**apoptosis**) and prevention of abnormal cell division. In contrast to proto-oncogenes, in tumor suppressors it is loss-of-function mutations that contribute to the progression of cancer. This means that tumor suppressor mutations tend to be recessive, and thus both alleles must be mutated in order to allow abnormal growth to proceed. It is perhaps not surprising that mutations in tumor suppressor genes, are more likely than oncogenes to be inherited. An example is the tumor suppressor gene, ***BRCA1***, which is involved in DNA-repair. Inherited mutations in *BRCA1* increase a woman's lifetime risk of breast cancer by up to seven times, although these heritable mutations account for only about 10% of breast cancer. Thus, sporadic rather than inherited mutations are the most common sources of both oncogenes and disabled tumor suppressor genes.

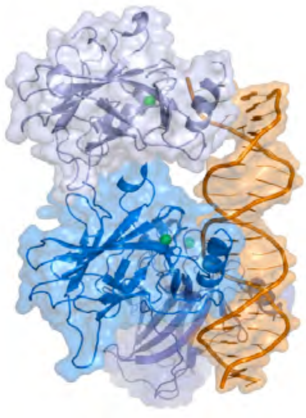


Figure 12.7 p53 bound to its target site on a DNA molecule.

An important tumor suppressor gene is a transcription factor named **p53**. Other proteins in the cell sense DNA damage, or abnormalities in the cell cycle and activate p53 through several mechanisms including **phosphorylation** (attachment of phosphate to specific site on the protein) and transport into the nucleus. In its active form, p53 induces the transcription of genes with several different types of tumor suppressing functions, including DNA repair, cell cycle arrest, and apoptosis. Over 50% of human tumors contain mutations in p53. People who inherit only one function copy of p53 have a greatly increased incidence of early onset cancer. However, as with the other cancer related genes we have discussed, most mutations in p53 are sporadic, rather than inherited.

Summary:

- Cancer is the name given to a class of different diseases that share common properties.
- Most cancers require accumulation of mutations in several different genes.
- Most cancer causing mutations are sporadic, rather than inherited, and most are caused by environmental carcinogens, including virus, radiation, and certain chemicals.
- Oncogenes are hyperactivated regulators of cell division, and are often derived from gain-of-function mutations in proto-oncogenes.
- Tumor suppressor genes normal help to repair DNA damage, arrest cell division, or to kill over proliferating cells. Loss-of-function of these genes contributes to the progression of cancer.

Key terms:

metastasis
dysplasia
carcinogen
HPV
oncogene
ionizing
epidemiology
proto-oncogene
receptor
signal transduction
ras
apoptosis
BRC1A
p53
tumor suppressor
phosphorylation

Study Questions:

12.1 Why do oncogenes tend to be dominant, but mutations in tumor suppressors tend to be recessive?

12.2 What tumor suppressing functions are controlled by p53? How can a single gene affect so many different biological pathways?

12.3 Are all carcinogens mutagens? Are all mutagens carcinogens? Explain why or why not.

12.4 Imagine that a laboratory reports that feeding a chocolate to laboratory rats increases the incidence of cancer. What other details would you want to know before you stopped eating chocolate?

12.5 Do all women with HPV get cancer? Why or why not? Do all women with mutations in *BRCA1* get cancer? Why or why not?

Photo and Illustration Credits

Ch	Fig	source	author	license[1]
1	0	flickr	eclectic echoes	CC: AND
1	1	original	unknown	PD
1	2	original	Deyholos	CC: AN
1	3	J. Exp. Med. 98:21, 1953	Austrian, Robert	pending
1	4	original	Deyholos	CC: AN
1	5	Wikipedia	Colm, Graham	PD
1	6	Wikipedia (modified)	Adenosine, Deyholos	CC: AS
1	7	Wikipedia	Ströck, Michael	GFDL
1	8	original	Deyholos	CC: AN
1	9	original	Deyholos	CC: AN
1	10	flickr	Westby, Max	CC: ANS
1	10	flickr	Joly, David	CC: ANS
1	10	Wikipedia	Altun, Zeynep F.	CC: AS
1	10	Wikipedia	Masur	GFDL
1	10	Wikipedia	Azul	GFDL
1	10	Wikipedia	unknown	GFDL
2	1	flickr: TheJCB	Zhang et al. (2007) J. Cell Biol. 177:231-242.	CC: ANS
2	2	Wikipedia	Wheeler, Richard	GFDL
2	3	unknown	unknown	pending
2	4	original	Deyholos	CC: AN
2	5	BMC	unknown	pending
2	6	PLoS Genetics	Somma MP et al. (2008) PLoS Genets 4(7): e1000126	PD
2	7	original	Deyholos	CC: AN
2	8	PLoS Genetics	Chelysheva, L. et al (2008) PLoS Genetics	PD
2	9	original	Deyholos	CC: AN
2	10	original	Deyholos	CC: AN
2	11	Wikipedia	NHGRI	PD
2	12	Wikipedia	Zephyris	GFDL
2	13	flickr	Bell, Darwin	CC: AN
2	14	flickr	Elissa Lei, Ph.D. @ NIH	CC: A
3	1	flickr	Guthier, Christian	CC: A
3	2	Wikipedia	Ruiz, Mariana	PD
3	3	original	Deyholos (Fireworks)	CC: AN
3	4	original	Deyholos	CC: AN
3	5	original	Deyholos	CC: AN
3	6	original	Deyholos	CC: AN
3	7	original	Deyholos	CC: AN
3	8	original	Deyholos	CC: AN
3	9	Wikipedia	PAR	PD
4	1	flickr	ecstaticist	CC: ANS
4	2	PLoS Biology	Zarbalis, K. et al (2004) PLoS Biology	PD
4	3	original	Deyholos	CC: AN
4	4	original	Deyholos	CC: AN
4	5	original	Deyholos	CC: AN
4	6	original	Deyholos	CC: AN
4	7	Wikipedia	Zephyris	GFDL
4	8	original	Deyholos	CC: AN
4	9	original	Deyholos	CC: AN
4	10	original	Deyholos	CC: AN
4	11	original	Deyholos	CC: AN
4	B1	original	Deyholos	CC: AN
4	B2	Wikipedia	unknown	PD
4	B2	flickr	windy234	CC: AN
5	1	Wikipedia	en:User:Drgnu23	GFDL
5	2	original	Deyholos	CC: AN
5	3	original	Deyholos	CC: AN
5	4	original	Deyholos	CC: AN
5	5	Wikipedia	Mrich	CC: AS
5	6	original	Deyholos	CC: AN
5	7	Wikipedia	U.S. Air Force photo/Staff Sgt Eric T. Sheler	PD
5	8	original	Deyholos	CC: AN
5	9	Wikipedia	unknown	PD
5	10	Wikipedia	NASA	PD
5	11	original	Deyholos (Fireworks)	CC: AN
5	12	flickr	Stern, Zach	CC: AND
6	1	flickr	Gossamer1013	CC: AND
6	2	original	Deyholos	CC: AN
6	3	original	Deyholos	CC: AN
6	4	original	Deyholos	CC: AN
6	5	flickr	Curley, John	CC: AN
6	5	flickr	Romans, Phil	CC: AND
6	5	flickr	Miss Chien	CC: AND
6	6	original	Deyholos	CC: AN
6	7	flickr	unknown	CC: AD
6	8	original	Deyholos	CC: AN
6	9	original	Deyholos	CC: AN
6	10	original	Deyholos	CC: AN
6	11	other	UN Lincoln	pending
6	12	original	Deyholos	CC: AN
6	13	original	Deyholos	CC: AN
7	1	Wikipedia	Abiyoyo	CC: AS
7	2	original	Deyholos	CC: AN
7	3	original	Deyholos	CC: AN

7	4	Wikipedia	Morgan	PD
7	5	original	Deyholos	CC: AN
7	6	original	Deyholos	CC: AN
7	7	original	Deyholos	CC: AN
7	8	original	Deyholos	CC: AN
7	9	original	Deyholos	CC: AN
7	10	NCBI	NIH	PD
7	11	original	Deyholos	CC: AN
7	12	original	Deyholos	CC: AN
7	13	original	Deyholos	CC: AN
8	1	flickr	estherase	CC: ANS
8	2	original	Deyholos	CC: AN
8	3	original	Deyholos	CC: AN
8	4	Wikipedia	madprime	GFDL
8	5	original	Deyholos	CC: AN
8	6	Wikipedia	madprime	GFDL
8	7	NCBI	--	PD
8	8	original	Deyholos	CC: AN
8	9	original	Deyholos	CC: AN
8	10	flickr	DeathByBokeh	CC: AN
8	11	flickr	457088634_585df11af5_o	pending
8	12	Wikipedia	Magnus Manske	PD
8	13	Wikipedia	Transcontrol	GFDL
8	14	original	Deyholos	CC: AN
9	1	flickr	Jaime Golombek	CC: AND
9	2	original	Deyholos	CC: AN
9	3	original	Deyholos	CC: AN
9	4	original	Deyholos	CC: AN
9	5	original	Deyholos	CC: AN
9	6	original	Deyholos	CC: AN
9	7	original	Deyholos	CC: AN
9	9	original	Deyholos	CC: AN
9	10	original	Deyholos	CC: AN
10	1	Goodsell, Scripps	Goodsell, Scripps	EDU
10	2	original	Deyholos	CC: AN
10	3	original	Deyholos	CC: AN
10	4	original	Deyholos	CC: AN
10	5	original	Deyholos	CC: AN
10	6	Wikipedia	Abizar Lakdawalla	PD
10	7	NCBI	unknown	PD
10	8	original	Deyholos	CC: AN
11	1	flickr	frenquency	CC: AND
11	2	original	Deyholos	CC: AN
11	3	original	Deyholos	CC: AN
11	4	original	Deyholos	CC: AN
11	5	original	Deyholos	CC: AN
11	6	original	Deyholos	CC: AN
11	7	original	Deyholos	CC: AN
11	8	original	Deyholos	CC: AN
11	9	original	Deyholos	CC: AN
11	10	original	Deyholos	CC: AN
11	11	flickr	Beardy Git	CC: AND
11	12	original	Deyholos	CC: AN
12	1	flickr	Uthman, Ed	CC: AS
12	2	Wikipedia	NIH	PD
12	3	Wikipedia	Hayman, J	PD
12	4	Wikipedia	unknown	PD
12	5	flickr	Uthman, Ed	CC: AS
12	6	Wikipedia	Mark 'AbsturZ'	PD
12	7	Wikipedia	Splettstoesser, Thomas based on Cho et al. Science 265 pp. 346, 1994	CC: AS

[1] License details:
CC: AD Creative Commons Attribution-No Derivative Works 2.0 Generic
CC: AN Creative Commons Attribution-Noncommercial 2.0 Generic
CC: AND Creative Commons Attribution-Noncommercial-No Derivative Works 2.0 Generic
CC: AS Creative Commons Attribution-Share Alike 2.0 Generic
EDU Educational use explicitly allowed by author, who retains copyright
GFDL Permission is granted to copy, distribute and/or modify this document under the terms of the GNU Free Documentation License
PD Public Domain because it was created by a US government agency or because the author has explicity released it into the public domain.