CHAPTER 1

Basic Concepts in Genetics and Linkage Analysis

ELIZABETH C. MELVIN and MARCY C. SPEER

This chapter explores the underpinnings for observational and experimental genetics. Concepts ranging from laws of Mendelian inheritance through molecular and chromosomal aspects of deoxyribonucleic acid (DNA) structure and function are defined; their ultimate utilization in linkage mapping of simple Mendelian disease and common and genetically complex disease is presented. The chapter concludes with clinical examples of the various types of DNA mutation and their implications for human disease.

INTRODUCTION

For centuries, the hereditary basis of human disease has fascinated both scientists and the general public. The Talmud gives behavioral proscriptions regarding circumcision in sons born after a male sibling who died of a bleeding disorder, suggesting the ancient Hebrews knew of hemophilia; nursing students in Britain in the 1600s tracked the recurrence of spina bifida in families; and questions as to whether Abraham Lincoln and certain celebrity sports figures had Marfan syndrome sometimes arise in casual dinner conversation.

In many respects, the study of the genetic factors in disease today remains, as it has for centuries, dependent on careful description of human pedigree data in which patterns of transmission from parent to offspring are characterized. For example, Gregor Mendel provided the groundwork for the study of human genetics by carefully constructing quantified observations of the frequency of variable characteristics in the pea plant. The importance of detailed pedigree analysis was exemplified recently in the delineation of patterns of transmission of the fragile X syndrome: the most common genetic cause of mental retardation. Careful documentation of pedigrees...
from families with more than one person with fragile X syndrome led to description of the aptly named Sherman paradox (Sherman et al., 1985), in which different recurrence risks for relatives of various types were described. From this, the complicated workings of unstable DNA harbored in expanding trinucleotide repeats were later elucidated (e.g., Fu et al., 1991; Burke et al., 1994).

HISTORICAL CONTRIBUTIONS

Segregation and Linkage Analysis

In 1865, Gregor Mendel, an Austrian monk (Fig. 1.1), published his findings on the inheritance of a series of traits in the pea plant, including seed texture (round or wrinkled), seed color (yellow or green), and plant height (tall or short). He described three properties of heritable factors that explained his quantified observations of these scorable (discontinuous or qualitative) traits. The first property was unit inheritance, which is now considered the basis for defining the gene. He hypothesized that a factor was transmitted from parent to offspring in an unchanged form. Such a factor produced an observable trait. This idea represented a radical departure

Figure 1.1. Mendel’s garden at the old monastery at Brno. On the right is the door to the Mendel museum, which contains exhibits celebrating his life and research. It is located in what used to be the monastery refectory (dining hall). Mendel’s apartment window(s) overlooked the garden. The garden is planted in ornamental red and white flowers, with the first two (farthest from the camera) labeled P (parental generation), the next single red one is labeled F1, the next row of four (three red and one white) is labeled F2, and the next nine labeled F3. At the far left, under a tree, is a large statue of Mendel that once stood in the town square just outside the garden gates. The square is called Mendelovaplatz. (Photo, taken in 1986, courtesy of Arthur S. Aylsworth.)
from the scientific thinking at the time, which suggested parental characteristics were blended in the offspring.

Mendel also described the behavior of factors controlling observable traits, such as flower color or plant height as a single unit. He proposed that these factors were transmitted ("segregated") independently and with equal frequency to germ cells (egg and sperm), and this observation is referred to as Mendel’s first law. In experiments for a variety of traits, Mendel crossbred the offspring (the F₁ generation) of two phenotypically different, pure-breeding parental strains with one another. The offspring of these matings (the F₂ generation) expressed the grandparental traits in a 3 : 1 ratio (Fig. 1.2). Serendipitously, some of the traits Mendel had chosen to study were simple dominant traits, such that the presence of one factor was sufficient to express the trait, with the other trait being recessive (expressed only in the absence of the dominant factor); later work showed that the factor defining a characteristic need not be dominant to the other. For instance, if each factor contributes to the trait equally, as in codominant systems, then three different classes of offspring from the same cross described above are possible: the two parental traits and a third intermediate trait. These classes occur in proportions parental to intermediate to parental of 1 : 2 : 1.

Figure 1.2. Principles of Mendel’s first law of segregation of heritable characters for a dominant trait.
Mendel extended his observations from the transmission of a single trait from parent to offspring to the interaction of two traits. Mendel’s law of independent assortment, also referred to as Mendel’s second law, predicts that factors controlling different traits will segregate to offspring independently from one another. For instance, seed texture will segregate to offspring independently of seed color. In one experiment Mendel crossed pure-breeding round, green seed plants to pure-breeding wrinkled, yellow seed plants. Since round is dominant to wrinkled and green is dominant to yellow, the resulting seeds (F₁ generation) yielded entirely round, green seed plants. These F₁ plants were then crossed with one another, confirming the predictions from his theory of independent assortment: The seed (offspring) types in the F₂ generation were nine round/green, three round/yellow and three wrinkled/green, one wrinkled/yellow (Fig. 1.3). Any observed departure from these expected ratios using identical parental crossing strategies suggests the two traits fail to segregate independently and may be physically linked. We will

![Figure 1.3. Principles of Mendel’s second law of independent assortment with a dominant trait.](image-url)
describe later how the failure of two traits to segregate independently can be exploited to find genes and diagnose genetic disorders.

Mendel’s observations remained largely obscure until the early 1900s, when they were independently rediscovered by plant geneticists and by Sir Archibald Garrod, who was studying the human hereditary disorder alkaptonuria (Garrod, 1902). Garrod’s work provided the basis of our understanding of alleles and genetic linkage. Mendel’s observations remain one of the most important contributions of critical descriptive science in the history of genetics.

**Hardy–Weinberg Equilibrium**

Another historical landmark in genetics occurred in the early 1900s as evolutionary biologists attempted to explain why the frequency of a dominant trait or disease in the population did not increase until, over many generations, everyone in the population was affected. The answer to this question was provided independently by Hardy (1908) and Weinberg (1908), who predicted the behavior of alleles in a population using the binomial theorem. Their proof, now called the Hardy–Weinberg theorem, shows that in a large, randomly mating population, trait (genotypic) frequencies for autosomal traits will achieve and remain in a state of equilibrium after one generation. Several evolutionary forces can alter equilibrium frequencies, including selection for or against a phenotype, migration into or out of a population, new mutation, and genetic drift. For a sex-linked trait, the attainment of equilibrium will require more than one generation.

Specifically, in a two-allele autosomal system with alleles A and a (having frequencies $p$ and $q$, respectively), $p + q = 1$. The Hardy–Weinberg theorem predicts the frequencies of genotypes AA, Aa, and aa are $p^2$, $2pq$, and $q^2$. Various manipulations of these algebraic formulas allow many useful calculations, such as carrier frequencies of diseases, disease prevalence, and gross estimates of penetrance. Some example applications of the Hardy–Weinberg theorem are shown in Table 1.1.

**DNA, GENES, AND CHROMOSOMES**

**Structure of DNA**

When Mendel described his genetic factor, he did not know what the underlying biological factor was. It was 90 years later when the actual genetic molecule was identified. Mendel’s fundamental unit of inheritance is termed the gene. A gene contains the information for synthesizing proteins necessary for human development, cellular and organ structure, and biological function. Deoxyribonucleic acid is the molecule that comprises the gene and encodes information for synthesizing both proteins and RNA (ribonucleic acid). Deoxyribonucleic acid is present in the nucleus of virtually every cell in the body. It is made up of three components: a sugar, a phosphate, and a base. In DNA, the sugar is deoxyribose, while in RNA the sugar is ribose. The four bases in DNA are the pyrimidines adenine (A) and
guanine (G) and the purines cytosine (C) and thymine (T). A DNA sequence is often described as an ordered list of bases, each represented by the first letter of its name (e.g., ACTGAAACTTGATT). A nucleoside is a molecule made of a base and a sugar; a nucleotide is made by adding a phosphate to a nucleoside. A single strand of DNA is a polynucleotide, consisting of nucleotides bonded together.

A single strand of DNA is, however, unstable. The double-helical nature of DNA, which confers stability to the molecule, was hypothesized in 1953 by J. D. Watson and F. H. C. Crick. Their cohesive theory of the structure of DNA accounted for some of the previously identified properties of DNA. A fascinating account of the internecine struggles in science surrounding this discovery was provided later by Watson (1968).

Specifically, Watson and Crick postulated that DNA is a double-stranded structure and that the two strands of DNA are arranged in an antiparallel orientation. In the central portion of the molecule, hydrogen bonds link a base with its complement, such that a purine always bonds with a pyrimidine (e.g., adenine always bonds with thymine and guanine always bonds with cytosine). The conformation of the

---

**TABLE 1.1. Useful Applications of Hardy–Weinberg Theory**

Recall that \( p + q = 1 \) and \( p^2 + 2pq = q^2 = 1 \).

**Example 1.** Cystic fibrosis (CF), an autosomal recessive disease, has an incidence of \( \frac{1}{400} \). What is the frequency of CF carriers in the general population?

The frequency of the CF allele \( (q) \) is calculated as \( \sqrt{\frac{1}{400}} = \frac{1}{20} \).

The frequency of CF carriers is calculated as \( 2pq = 2 \left( \frac{19}{20} \right) \left( \frac{1}{20} \right) = 0.095 \).

**Example 2.** The frequency of the allele \( (q) \) for an autosomal dominant disorder is \( \frac{1}{100} \). What is the frequency of the disease itself in the population?

Since the frequency of the disease allele is \( \frac{1}{100} \), the frequency of the normal allele \( (p) = 1 - \frac{1}{100} = \frac{99}{100} \).

Since the disease is dominant, both heterozygous carriers and homozygous individuals are affected with the disease:

\[
q^2 + 2pq = \left( \frac{1}{100} \right)^2 + 2 \left( \frac{99}{100} \right) \left( \frac{1}{100} \right) = 0.0199
\]

**Example 3.** An autosomal dominant disorder with incomplete penetrance \( (f) \) has a population prevalence of \( \frac{16}{1000} \). If the allele frequency for the normal allele \( p \) is 0.99, what is the estimated penetrance of the disease allele?

Since \( p = 0.99 \), then \( q = 0.01 \).

As in the Example 2, both heterozygous and homozygous gene carriers are affected (assuming no difference in penetrance) between homozygotes and heterozygotes. Therefore,

\[
f(q^2) + f(2pq) = 0.016
\]

\[
f(q^2 + 2pq) = 0.016
\]

\[
f(0.0199) = 0.016
\]

\[
f = 0.804
\]
resultant molecule is the double helix, which undergoes several levels of compacting to fit within the cell (Fig. 1.4).

The sequence of DNA bases represents a code for synthesizing proteins. The fundamental unit of this genetic code is termed a codon, which consists of three nucleotides. Since there are four different nucleotides (one made with each of the four bases) and a codon is made of three nucleotides, there are \(4^3 = 64\) different codons. However, these 64 codons specify only 20 different amino acids, which are the building blocks of proteins. Thus, the genetic code is degenerate: Different codons may code for the same amino acid. In addition, some codons act as punctuation. For instance, one specific codon in a string of DNA signals the molecular code “interpreter” to start, and then the reading of the DNA strand proceeds in three base-pair chunks; several other specific codons signal the reading process to cease. These reading signals are called start and stop codons, respectively.

Not all of the DNA in a cell actually codes for a protein product; in fact, the vast majority of the DNA sequence does not carry the information for protein formation. Within a gene, exons are the portions utilized (transcribed) to make proteins. Introns are the sequences between exons that are not transcribed. The size and number of introns and exons vary dramatically between genes (Fig. 1.5).

The central dogma of genetics is that the utilization of DNA is unidirectional such that DNA → RNA → protein (Fig. 1.6). Specifically genes are encoded in the DNA. Then, in the nucleus, messenger RNA (mRNA) is transcribed (produced) from the DNA. Subsequently mRNA undergoes a series of posttranscriptional modifications: The introns are spliced out, a cap is added at the 5’ end of the
The molecule, and a string of adenylate residues (poly-A tail) is added to the 3' end. The mRNA is then transported out of the nucleus into the cytoplasm, where it is translated into protein by means of cellular machinery called the ribosomes. Many excellent resources describe the very complicated process of transcription and translation (e.g., Strachan and Read, 1996).

Figure 1.5. Intron and exon sizes vary between genes.

Figure 1.6. Central dogma of genetics: DNA → RNA → protein. (Reprinted by permission from Jorde et al., eds., Medical Genetics, C. W. Mosby, St. Louis, MO, 1995.)
Genes and Alleles

The physical site or location of a gene is called its locus. At any particular gene site, or locus, there can exist different forms of the gene, called alleles. Except on the sex chromosomes of males, an individual has two alleles at each locus. These alleles are analogous to the factors identified in the 1800s by Mendel. Homozygosity is defined as the presence of two alleles that are indistinguishable from one another. In heterozygotes, the two alleles can be distinguished from another. Males with a normal chromosome complement are hemizygous for all X chromosome loci, since they have only one copy of the X chromosome.

The difference between two alleles may be as subtle as a single base-pair change, such as the thymine-to-alanine substitution that alters the B chain of hemoglobin A from its wild type to its hemoglobin sickle cell state. Some base-pair changes have no deleterious effect on the function of the gene; nevertheless, these functionally neutral changes in the DNA still represent different forms of a gene. Alternatively, allelic differences can be as extensive as large, multicodon deletions, such as those observed in Duchenne muscular dystrophy. Any locus having two or more alleles, each with a frequency of at least 1% in the general population, is considered to be polymorphic (i.e., having many forms).

Differences in alleles can be scored via laboratory testing. The ability to score allele differences accurately within families, between families, and between laboratories is critically important for linkage analysis in both simple Mendelian and genetically complex common disorders. Allele scoring strategies may be as simple as the presence (+) or absence (−) of a deletion or point mutation or as complicated as assessing the allele size in base pairs of DNA. The latter application is common when highly polymorphic microsatellite repeat markers are used in linkage analysis.

A measure frequently utilized to quantitate the extent of polymorphism of a gene or marker system is the heterozygosity value $H$, which is calculated as

$$H = 1.0 - \sum_{i=1}^{n} p_i^2$$

where $n$ is the number of alleles at the locus and $p_i$ is the frequency of the $i$th allele at the locus. For example, the heterozygosity value of a three-allele marker with frequencies of allele 1 = 0.25, allele 2 = 0.30, and allele 3 = 0.45 is calculated as $1 - [(0.25 \times 0.25) + (0.30 \times 0.30) + (0.45 \times 0.45)] = 0.645$. This measure can be interpreted as the probability that an individual randomly selected from the general population will be heterozygous at the locus. Polymorphic loci or genetic markers are critically important to linkage analysis because they allow each individual in a family a high probability of being heterozygous for the locus. The investigator may then be able to deduce the parental origin of each allele and identify recombinant and nonrecombinant gametes, which is important for linkage analysis, as described later.

In general, genetic marker systems with high heterozygosities are desirable because there is a high probability that an individual in a genetic linkage analysis
will be heterozygous for that marker and thus be likely to contribute information to a genetic linkage study. However, recent emphasis has been on utilizing single-nucleotide polymorphisms (SNPs) as genetic markers, despite their limited heterozygosity, because of their ubiquity in the genome.

The PIC (or polymorphism information content) is a modification of the heterozygosity measure that subtracts from the $H$ value an additional probability that an individual in a linkage analysis does not contribute information to the study. Formally, the PIC value subtracts from the $H$ value the probability of obtaining a heterozygous offspring from an intercross mating (i.e., when a mother whose marker genotype is $1$ and a father whose marker genotype is $1$ have an offspring whose marker genotype is also $1$, it is impossible to tell whether the $1$ allele came from the mother or from the father; thus, this combination contributes essentially no information to the linkage analysis).

**Genes and Chromosomes**

Genes are organized as linear structures called chromosomes, with many thousands of genes on each chromosome. Each chromosome has distinguishable sites that aid in cell division and in the maintenance of chromosome integrity. The centromere is visualized as the central constriction on a chromosome and it separates the p (short) and the q (long) arms from one another. The centromere enables correct segregation of the duplicated chromosomal material during meiosis and mitosis. Telomeres are present at both ends of the chromosome and are required for stability of the chromosomal unit.

Using appropriate staining techniques, the chromosomes in a cell can be analyzed under the microscope following cell culture and the arrest of cell division at metaphase (when the chromosomes have duplicated and condensed). At this stage of the cell cycle, a chromosome has two double-stranded DNA molecules. Together, the strands are called *sister chromatids*. The sister chromatids are held together by the centromere. Photographs are magnified and the chromosomes are arranged into a karyotype. The normal human chromosome complement consists of 46 chromosomes arranged in 23 pairs, with one member of each pair inherited from each parent (Fig. 1.7). The first 22 pairs, called autosomes, are arranged according to size and are the same in males and females. The pair of sex chromosomes generally predicts an individual’s gender. Most females have two X chromosomes, while males have one X inherited from the mother and one Y chromosome inherited from the father. Therefore, the gender of an individual is determined by the father.

Because two copies of each chromosome are present in a normal somatic (body) cell, the human organism is diploid. In contrast, egg and sperm cells have haploid chromosomal complements, consisting of a single member of each chromosome pair. The correct number of chromosomes in the normal human cell was finally established in 1956, three years after the double-helical structure of DNA was described, when Tjio and Levan (1956) demonstrated unequivocally that the chromosomal complement is 46.
Regions of chromosomes are defined by patterns of alternating light and dark regions called bands, which become apparent after a chemical treatment has been applied. One of the most common types of banding process, called Giemsa or G banding, involves digesting the chromosomes with trypsin and then staining with a Giemsa dye. G banding identifies late-replicating regions of DNA; these are the dark bands. Other chemical processes will produce different banding patterns and identify unique types of DNA.

A specific genetic locus can then be defined quite precisely along a chromosome, such as the gene FRAXA (fragile X syndrome), which is located on the X chromosome at band q27.3. Alternatively, its localization may be specified as an interval flanked by two genetic markers. Any two loci that occur on the same chromosome are considered to be syntenic or physically linked. Two genes may be syntenic yet far enough apart on the chromosome to segregate independently from one another. Thus, two syntenic genes may be genetically unlinked. Two syntenic genes that fail to be transmitted to gametes independently from one another are genetically linked (Fig. 1.8). The location of two loci on the same arm of the chromosome is specified by their positions relative to each other and to the centromere. The gene closer to the centromere is termed centromeric or proximal to the other; similarly, the gene further from the centromere is distal or telomeric to the other (Fig. 1.9).
Figure 1.9. The myotonic dystrophy (DM) and insulin receptor (INSR) genes are distal (telomeric) to the ryanodine receptor 1 and CADASIL, respectively; RYR1 and CADASIL are proximal (centromeric) to DM and INSR, respectively.
The X and Y chromosomes are very different in their genetic composition except for an area at the distal end of the p arm of each, termed the pseudoautosomal region. The pseudoautosomal region behaves similarly to the autosomes during meiosis by allowing for segregation of the sex chromosomes. Just proximal to the pseudoautosomal region on the Y chromosome are the SRY (sex-determining region on the Y) and TDF (testes-determining factor) genes, which are critical for the normal development of male reproductive organs. When crossing over extends past the boundary of the pseudoautosomal region and includes one or both of these genes, sexual development will most likely be adversely affected. For instance, the rare occurrences of chromosomally XX males and XY females are due to such aberrant crossing over.

INHERITANCE PATTERNS IN MENDELIAN DISEASE

Alleles whose loci are on an autosome can be transmitted in a dominant or recessive (or codominant) fashion; similarly, alleles having loci on the X chromosome are expressed and transmitted as either X-linked recessive or X-linked dominant disorders. These well-known patterns of inheritance are shown in Figure 1.10.

The hallmark of dominant inheritance, regardless of whether the underlying gene is located on an autosome or on an X chromosome, is that only a single allele is necessary for expression of the phenotype. In an autosomal recessive trait, two copies of a trait allele must be present for it to be expressed. In most cases, it is correctly assumed that each parent of an offspring with the trait carries a recessive
allele at the locus. Rarely, an individual expressing an autosomal recessive disorder has inherited both abnormal recessive alleles from one parent; this phenomenon is called uniparental disomy.

To express an X-linked recessive disorder, a male needs only one abnormal allele; a female usually needs two abnormal alleles (one on each X chromosome) and this is generally rare. Thus, mostly males are affected. For X-linked dominant traits, both men and women can be affected since only one copy of the trait allele is necessary for phenotypic expression. Because females have two X chromosomes, twice as many females are affected as males.

Alleles located on the Y chromosome are transmitted from affected males to all sons, and in each case the son’s Y-linked phenotype will be identical to that of the father; daughters of males with a Y-linked trait will not inherit the trait, since they receive their father’s X chromosome. Very few expressed genes have been localized to the Y chromosome. Characteristics of each of the different Mendelian inheritance patterns are summarized in Table 1.2.

**GENETIC CHANGES ASSOCIATED WITH DISEASE/TRAIT PHENOTYPES**

Alterations or mutations in the genetic code can be neutral, beneficial, or deleterious. Changes in the genetic code can lead to trait and/or disease phenotypes; the pathology can be the result of either loss or gain of function of the gene product. Such changes can occur in a number of different ways.

**Point Mutations**

A point mutation is defined as an alteration in a single base pair in a stretch of DNA, thereby changing the 3-bp codon. Since the genetic code is degenerate, many such changes do not necessarily alter the resulting amino acid; however, if the single-base-pair change leads to the substitution of one amino acid for another, the result can be devastating. Point mutations can be classified as transition mutations (a purine → purine or a pyrimidine → pyrimidine) or as the less common transversion mutations (purine → pyrimidine or a pyrimidine → purine). In general, transitions are less likely than transversion mutations to change the resulting amino acid. Five effects of point mutations have been defined: Synonymous or silent mutations are single-base-pair changes in the DNA that do not affect the resultant amino acid; nonsense mutations result in a premature stop codon, leading to a polypeptide of reduced length; missense mutations lead to the substitution of one amino acid for another; splice site mutations affect the correct processing of the mRNA strand by eliminating a signal for the excision of an intron; and mutations in regulatory genes alter the amount of material produced. Several examples of point mutations in human diseases are discussed in the sections that follow.
TABLE 1.2. Hallmarks of Mendelian Inheritance Patterns of Different Types

<table>
<thead>
<tr>
<th>Inheritance Pattern</th>
<th>Examples</th>
<th>Gender Differences in Proportion of Affecteds?</th>
<th>Transmission Features</th>
<th>Recurrence Risks</th>
<th>Prevalence in Population</th>
<th>Other Critical Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>Autosomal dominant</td>
<td>Marfan syndrome; neurofibromatosis; myotonic dystrophy</td>
<td>No</td>
<td>Transmitted from affected parent to affected offspring (vertical transmission); male-to-male transmission</td>
<td>For each offspring of affected parent, risk to child to inherit disease gene is 50%</td>
<td>$p^2 + 2pq$</td>
<td>Reduced penetrance frequent; for a &quot;true dominant,&quot; individuals heterozygous for trait allele are no more severe than individuals homozygous for trait allele</td>
</tr>
<tr>
<td>Autosomal recessive</td>
<td>Sickle cell anemia; cystic fibrosis</td>
<td>No</td>
<td>Carrier parents generally unaffected</td>
<td>For parents who have one affected child, risk for each subsequent child is 25%</td>
<td>$q^2$</td>
<td>Consanguinity frequent</td>
</tr>
<tr>
<td>Sex-linked recessive</td>
<td>Duchenne muscular dystrophy; hemophilia</td>
<td>Males more frequently affected; carrier females generally unaffected; rare cases of nonrandom X inactivation can lead to affected females</td>
<td>Gene transmitted from unaffected carrier mother to affected son; no male-to-male transmission</td>
<td>Carrier mother has 25% chance to have affected son and 25% chance to have carrier daughter; all daughters of affected males are carriers and no sons of affected males are affected</td>
<td>Females: $q^2$; males: $q$</td>
<td>Females affected in rare cases of nonrandom X inactivation</td>
</tr>
<tr>
<td>Inheritance Pattern</td>
<td>Examples</td>
<td>Gender Differences in Proportion of Affecteds?</td>
<td>Transmission Features</td>
<td>Recurrence Risks</td>
<td>Prevalence in Population</td>
<td>Other Critical Features</td>
</tr>
<tr>
<td>---------------------</td>
<td>----------</td>
<td>-----------------------------------------------</td>
<td>----------------------</td>
<td>------------------</td>
<td>-------------------------</td>
<td>------------------------</td>
</tr>
<tr>
<td>Sex-linked dominant</td>
<td>Hypophosphatemic rickets; fragile X syndrome</td>
<td>No</td>
<td>Vertical transmission from mothers to both sons and daughters; fathers transmit to daughters only; no male-to-male transmission</td>
<td>50% of offspring of affected mothers are affected (unless mother is homozygous for disease allele); all daughters of affected males are affected and no sons of affected males are affected</td>
<td>Females: ( p^2 + 2pq ), males: ( p )</td>
<td>Females affected three times more frequently than males</td>
</tr>
<tr>
<td>Y linked</td>
<td>Genes <em>SRY</em> and <em>TDF</em>, important in sex determination, are on the Y chromosome; no known diseases are located on Y</td>
<td>Yes; only males would express trait</td>
<td>Exclusively male-to-male transmission</td>
<td>All sons of affected males are affected; no daughters of affected males are affected</td>
<td>Females: 0; males: ( q )</td>
<td>Male-determining genes are located just proximal to pseudoautosomal region on Y chromosome; faulty recombination in pseudoautosomal region can lead to errors in sex determination</td>
</tr>
<tr>
<td>Autosomal codominant</td>
<td>MN blood group; microsatellite repeat markers</td>
<td>No</td>
<td>Each allele confers measurable component to phenotype</td>
<td>Varies according to mating type</td>
<td>Genotypes expected to occur in Hardy-Weinberg proportions of ( p^2 ), ( 2pq ), and ( q^2 )</td>
<td></td>
</tr>
</tbody>
</table>
**Amyotrophic Lateral Sclerosis.** Approximately 10–15% of amyotrophic lateral sclerosis (ALS) patients have a positive family history consistent with autosomal dominant inheritance. This rapidly progressive neurodegenerative disorder has an average age of onset in the mid-40s and is usually fatal within a few years after onset. The gene responsible for about 15–20% of familial ALS cases has been identified as the cytosolic form of superoxide dismutase (Cu, Zn SOD) at 21q22.1. To date, 38 different point mutations have been identified. A recent study of clinical correlations associated with different mutations within the SOD1 gene (Juneja et al., 1997) found evidence for a significantly faster rate of progression (1.0 ± 0.4 year vs. 5.1 ± 5.1 years) in patients with the most common mutation, a transition mutation in which a thymine at codon 4 is substituted for a cytosine. This substitution changes the resultant amino acid from an alanine to a valine.

**Sickle Cell Anemia.** Sickle cell anemia, an autosomal recessive disorder with a carrier frequency in African Americans of approximately 1/12, is a classic example of a point mutation leading to disease. Affected patients have the familiar phenotype of chronic anemia, sickle cell crises, and debilitating pain. Sickle cell anemia results from a single nucleotide substitution of an adenine to a thymine at position 6 in the B chain of hemoglobin. This changes the resultant amino acid from glutamine to valine. Interestingly, the carrier state for sickle cell may lead to a selective advantage in certain environments: Carriers have a resistance to malaria that is useful in tropical climes.

**Achondroplasia.** Achondroplasia, the most common type of short-limbed dwarfism, is an autosomal dominant disorder. About 85% of cases are the result of a new mutation. It has been observed that the rate of new dominant mutations increases with advancing paternal age (Penrose, 1955; Stoll et al., 1982). Achondroplasia is now known to result from mutations in the fibroblast growth receptor 3 gene (FGFR3), located on chromosome 4p16.3. Interestingly, over 95% of the mutations are the identical G-to-A transition at nucleotide 1138 on the paternal allele (Rousseau et al., 1994; Shiang et al., 1994; Bellus et al., 1995). Other mutations in FGFR3 are also responsible for hypochondroplasia and thanatophoric dysplasia, types of dwarfism that are clinically distinct from achondroplasia.

**Deletion/Insertion Mutations**

Another class of mutations involves the deletion or insertion of DNA into an existing sequence. Deletions or insertions may be as small as 1 bp or they may involve one or many exons or even the entire gene. Even single-base-pair deletions or insertions can have devastating effects, frequently by altering the reading frame of the DNA strand.

**Neurofibromatosis.** Neurofibromatosis type 1 (NF1) is an autosomal dominant disorder with variable expression; penetrance of the disorder is high, and some clinicians consider penetrance of this gene to be complete. The most common phenotypic manifestations are multiple café-au-lait spots and peripheral
neurofibromatous skin tumors. Approximately 50% of all cases are due to new mutations, most frequently on the paternally inherited allele.

The gene for NF1, coding for a protein called neurofibromin, is located at chromosome 17q11.2 and has been cloned. Its function is thought to involve tumor suppression (DeClue et al., 1992). To date, only a fraction of the mutations responsible for NF1 have been identified. Those identified include entire deletions of the gene (Wu et al., 1995), insertion mutations, and small and large deletions. Single-base-pair mutations leading to premature stop codons (Valero et al., 1994) and deletions, both leading to the production of a truncated protein, account for the majority of NF1 mutations.

**Duchenne and Becker Forms of Muscular Dystrophy.** Duchenne muscular dystrophy (DMD) is a severe, childhood-onset X-linked muscular dystrophy; Becker muscular dystrophy (BMD) is its allelic, clinically milder variant. Boys with DMD develop normally for the first few years of life, after which rapidly progressive muscle deterioration becomes obvious. Affected males lose the ability to walk by age 10–12 years. The eventual loss of muscle strength in the cardiac and respiratory muscles leads to death in early adulthood. The gene coding for the protein dystrophin, which is abnormal in DMD/BMD, has been cloned (Koenig et al., 1988). Approximately two-thirds of mutations in this very large gene have been identified; the majority are deletions and duplications, although point mutations have also been identified. Correlations of the clinical phenotype with the molecular mutation have been complicated in DMD/BMD. The general deletions and point mutations leading to alterations in the reading frame of the DNA molecule (frame-shift mutations) are more severe than those that do not alter the reading frame (in-frame mutations).

**Cystic Fibrosis.** Cystic fibrosis (CF), an autosomal recessive disorder, is the most common hereditary disease among Caucasians, with a carrier frequency of between $\frac{1}{20}$ and $\frac{1}{30}$. The function of the pancreas, lungs, and sweat glands, among other organ systems, is affected. In American Caucasians, a single mutation called $\Delta F508$ accounts for about 70% of the abnormal CF alleles. Three base pairs (codon 508) are deleted, and the resulting amino acid sequence is missing a phenylalanine. Over 900 other deleterious mutations have been identified throughout the world. The frequency of specific mutations differs among populations.

**Novel Mechanisms of Mutation: Unstable DNA and Trinucleotide Repeats**

Dynamic mutations, or unstable DNA, have received considerable attention of late. Some loci of the genome have variable numbers of dinucleotide or trinucleotide repeats. Most are not associated with expressed genes but can be exploited as markers, since they are highly polymorphic. A few loci with trinucleotide repeats are near or within genes, and by expansion beyond a certain threshold, these disrupt gene expression and cause disease. To date, 11 disorders (7 autosomal and 4 X
linked) have been shown to be the result of expansion of these unstable triplet repeats (Table 1.3).

The phenotypes of myotonic dystrophy, Huntington disease, and Machado–Joseph disease, among others, are associated with anticipation, a clinical phenomenon in which disease severity worsens in each successive generation (see also Chapter 2). Because disease expression can be quite variable and difficult to measure, age of onset is frequently utilized as an analogue of severity, and anticipation is then observed as a decreasing mean age of onset with each passing generation. Since the discovery that expanding trinucleotide repeats may explain anticipation, investigators have reported clinical evidence for anticipation in many various disorders, including bipolar affective disorder (McInnis et al., 1993), limb–girdle muscular dystrophy (Speer et al., 1998), familial spastic paraplegia (Raskind et al., 1997; Scott et al., 1997), and facioscapulohumeral muscular dystrophy (Tawil et al., 1996; Zatz et al., 1995). None of these disorders is proven to be caused by trinucleotide repeat expansions, and elucidation of their underlying defect will shed additional light on the phenomenon of anticipation.

**SUSCEPTIBILITY VERSUS CAUSATIVE GENES**

As the study of common and genetically complex human diseases identifies the significant contribution of heredity in their development, it is likely that more genes or genetic risk factors will be found to affect susceptibility to disease rather than the more traditionally considered causative genes. Historically, one of the most widely investigated examples of susceptibility loci is the human leukocyte antigen (HLA) system on the p arm of chromosome 6. Specific HLA antigens have been associated with various human diseases; for instance, the Bw47 antigen confers an 80–150-fold increased risk for congenital adrenal hyperplasia; the B27 antigen confers an 80–100-fold increased risk for ankylosing spondylitis; and the DR2 antigen confers a 30–100-fold increased risk for narcolepsy, a 3-fold increased risk for systemic lupus erythematosus, and a 4-fold increased risk for multiple sclerosis.

A recent and well-characterized example of a susceptibility locus is that of the apolipoprotein E (APOE) gene and Alzheimer’s disease (AD). The APOE gene on chromosome 19 has three different alleles, scored as 2, 3, and 4, which occur with frequency 6, 78, and 16% in most European populations, respectively (e.g., Saunders et al., 1993). These alleles differ in their DNA sequence by only one base at codons 112 and/or 158 (Fig. 1.11). The APOE 4 allele increases risk and decreases age of onset in familial and sporadic late-onset AD and early-onset sporadic AD. The 2 allele has been shown to be protective to some extent for risk to develop AD (Corder et al., 1994, 1995a,b; Farrer et al., 1997). Interestingly, the 4 allele has been shown to exist at lower frequency in the Indiana Amish, at least partially explaining the decreased frequency of AD in this inbred population (Pericak-Vance et al., 1996). It is important to note that for APOE and AD, the 4 allele is not by itself sufficient or necessary for the development of AD but has
<table>
<thead>
<tr>
<th>Condition</th>
<th>Gene Symbol</th>
<th>Chromosome Location</th>
<th>Repeat Type</th>
<th>Repeat Localization</th>
<th>Repeat Number Abnormal Range</th>
<th>Clinical Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fragile X syndrome</td>
<td>FRAXA</td>
<td>Xq27.3</td>
<td>CGG</td>
<td>5' Untranslated region (premutations in range of 52–200)</td>
<td>200–1000</td>
<td>Moderate to severe mental retardation, macroorchidism, large ears, and prominent jaw; FRAXA accounts for about one-half of all X-linked mental retardation</td>
</tr>
<tr>
<td>Fragile site mental retardation-2</td>
<td>FRAXE</td>
<td>Xq28</td>
<td>GCC</td>
<td>?</td>
<td>200–1000</td>
<td>Similar to fragile X syndrome phenotypically; cytogenetic evidence for Xq fragile site; negative for expansion in FRAXA</td>
</tr>
<tr>
<td>Fragile site F</td>
<td>FRAXF</td>
<td>Xq28</td>
<td>(GCCGTC)$_n$ (GCC)$_n$</td>
<td>?</td>
<td>300–500</td>
<td>Cytogenetic evidence for Xq fragile site without molecular expansion at FRAXA or FRAXE; whether aberrant phenotype is associated with expansion at FRAXF is uncertain.</td>
</tr>
<tr>
<td>Condition</td>
<td>Chromosome</td>
<td>Repeat Unit</td>
<td>Disease Manifestation</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>----------------------------------------------</td>
<td>------------</td>
<td>-------------</td>
<td>---------------------------------------------------------------------------------------</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fragile site 16q22</td>
<td>16q22</td>
<td>CCG</td>
<td>1000–2000 Expansion is molecular explanation for cytogenetic observation of fragile site at 16q22; expansion has been associated with infertility and spontaneous abortions</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kennedy spinal and bulbar muscular atrophy (SBMA)</td>
<td>Xq11–q12</td>
<td>CAG Open reading frame</td>
<td>40–52 Caused by defect in androgen receptor gene, SBMA usually presents in midlife with bulbar signs and facial fasciculations</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Huntington disease (HD)</td>
<td>4p16.3</td>
<td>CAG Open reading frame</td>
<td>37–100 Caused by a defect in Huntingtin gene, HD is characterized by choreiform movements, rigidity, and dementia</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spinocerebellar ataxia, type 1</td>
<td>6p23</td>
<td>CAG Open reading frame</td>
<td>&lt;100 Autosomal dominant ataxia with onset in 30s; upper motor neuron signs and extensor planton responses</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(continued)
<table>
<thead>
<tr>
<th>Condition</th>
<th>Gene Symbol</th>
<th>Chromosome Location</th>
<th>Repeat Type</th>
<th>Repeat Localization</th>
<th>Repeat Number Abnormal Range</th>
<th>Clinical Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dentatorubropallidolusian atrophy, Haw River syndrome</td>
<td><em>DRPLA</em></td>
<td>12p13.31</td>
<td>CAG</td>
<td>Open reading frame</td>
<td>&lt;100</td>
<td>Myoclonus epilepsy, dementia, ataxia, and choreoathetosis transmitted as autosomal dominant</td>
</tr>
<tr>
<td>Machado–Joseph (spinocerebellar ataxia type 3)</td>
<td><em>MJD; SCA3</em></td>
<td>14q24.3–q31</td>
<td>CAG</td>
<td>Open reading frame</td>
<td>61–84</td>
<td>Ataxia with onset usually in 40s; frequent dystonia and facial fasciculations</td>
</tr>
<tr>
<td>Myotonic dystrophy</td>
<td><em>DM</em></td>
<td>19q13.2–q13.3</td>
<td>CTG</td>
<td>3′ Untranslated region</td>
<td>200–4000</td>
<td>Myotonia, ptosis, characteristic cataracts, testicular atrophy, and frontal balding</td>
</tr>
<tr>
<td>Spinocerebellar ataxia type 2</td>
<td><em>SCA2</em></td>
<td>12q24.1</td>
<td>CAG</td>
<td>5′ Coding region</td>
<td>36–59</td>
<td>Abnormalities of balance due to cerebellar dysfunction or pathology; clinically identical to <em>SCA1</em></td>
</tr>
</tbody>
</table>
been shown to be associated with increased susceptibility to AD. The underlying biological mechanism for this observation has not yet been precisely identified.

The most frequent successes to date in the localization of genes underlying disease linkage analysis have been with diseases whose mode of inheritance is known (as illustrated above). These disorders are often highly or completely penetrant and are due to a defect in a single gene, yet these Mendelian disorders are often relatively rare in the population. However, some of the most common and deadly diseases of society such as cardiovascular disease and obesity have significant genetic components. These diseases are termed “complex” because they are likely due to the interaction of multiple factors, both environmental and genetic. Susceptibility genes for such complex disorders are substantially harder to identify than genes responsible for Mendelian disorders.

**GENES, MITOSIS, AND MEIOSIS**

A cell’s ability to reproduce itself is critical to the survival of an organism. This cell duplication process, utilized by somatic cells, is called mitosis. Similarly, an organism’s ability to reproduce itself is critical to the survival of the species. In sexual organisms, the reproductive process involves the union of gametes (sperm and egg cells), which are haploid. Meiosis is the process by which these haploid gametes are formed and is the biological basis of linkage analysis.

Meiosis consists of two parts: meiosis I and meiosis II. In meiosis I, which is called a reduction division stage, each chromosome in a cell is replicated to yield two sets of duplicated homologous chromosomes. During meiosis I, physical contact between chromatids may occur, resulting in the formation of chiasmata. Chiasmata are thought to represent the process of crossing over or recombination, in which an exchange of DNA between two (of the four) chromatids occurs (Fig. 1.12). A chiasma occurs at least once per chromosome pair. Thus, a parental haplotype (the arrangement of many alleles along a chromosome) may not remain intact upon transmission to an offspring. When two loci are unlinked to one another, the recombination fraction ($\theta$) between them is 0.50. The upper limit for observed recombination between two unlinked loci is set at 50% because the frequency with which odd numbers of recombination events between a pair of loci occur

**Figure 1.11.** Single base-pair changes in exon 4 of APOE define the 2, 3, and 4 alleles at this locus. (Modified from M. A. Pericak-Vance and J. L. Haines, Trends Genet 11, 1995.)
should equal the frequency with which even numbers of recombination events occur; when an even number of recombination events occurs between two loci, the resultant gametes appear to be nonrecombinant and hence these recombination events are unobserved.

Following crossing over, at least two of the four chromatids become unique, unlike those of the parent. The cellular division process that occurs ensures that one paternal homologue and one maternal homologue are transmitted to each of two diploid daughter cells. This cell division marks the end of meiosis I.

The process of genetic recombination helps to preserve genetic variability within a species by allowing for virtually limitless combinations of alleles in the transmission from parent to offspring. Estimates of genetic recombination can also predict distance between two loci: The closer two loci are to one another, the less chance for recombination between them. The frequency of recombination is not uniform through the genome. Some areas of some chromosomes have increased rates of recombination (hot spots), while others have reduced rates of recombination (cold

**Figure 1.12.** Genetic results of crossing over: (a) no crossover: A and B remain together after meiosis; (b) crossover between A and B results in a recombination (A and B are inherited together on a chromosome and A and B are inherited together on another chromosome); (c) double crossover between A and B results in no recombination of alleles. (Reprinted by permission from Jorde et al., eds., *Medical Genetics*, C. W. Mosby, St. Louis, MO, 1995.)
spots). For instance, recombination frequencies may vary between sexes or may vary depending on whether the loci are at the telomere or centromere of the chromosome.

The second phase of meiosis is identical to a mitotic (somatic cell) division, in which genetic material is transmitted equally, identically, and without recombination to daughter cells. However, in contrast to a mitotic division, which yields two identical diploid daughter cells, the end result of the entire meiotic process in sperm cells is four haploid daughter cells with chromosomal haplotypes different from those originally present in the parent; in egg cells, the final outcome is a single haploid daughter cell, with the remainder of the genetic material lost because of the formation of nonviable polar bodies.

The fundamental differences between meiosis and mitosis are summarized in Table 1.4.

When Genes and Chromosomes Segregate Abnormally

Failure of meiosis at either phase (meiosis I or II) is termed nondisjunction and leads to aneuploidy, or abnormal chromosomal complements. The most well known aneuploidy is Down syndrome, caused by an extra copy of chromosome 21. Down syndrome is often called trisomy 21 because patients have a total of 47 chromosomes, with three copies of chromosome 21. A monosomy, or the absence of a second member of a chromosome pair, is rarely viable. A noted exception is Turner syndrome, in which a female has a total of 45 chromosomes, only one of which is an X chromosome.

Triploidy and tetraploidy are the terms for the presence of one or two entire extra sets of chromosomes, leading to a total of 69 or 92 chromosomes. These anomalies, which usually are inviable in humans, are due to errors in fertilization such as dispermy (two sperm fertilizing an ovum) or failure of the ovum’s polar body to separate.
Segregation distortion, a phenomenon so far observed only rarely in humans, is characterized by a departure from the 50:50 segregation ratio expected from normal meiosis. Specifically, one allele at a locus is transmitted to the gamete more than 50% of the time. Segregation distortion, also termed meiotic drive, has been described in many experimental systems. Myotonic dystrophy (Beri et al., 1994), an autosomal dominant muscular dystrophy, was among the first human disease alleles suggesting preferential transmission of an allele to offspring.

ORDERING AND SPACING OF LOCI BY MAPPING TECHNIQUES

The segregation of loci in meiosis provides the opportunity for assessment of Mendel’s law of independent assortment. When two loci are unlinked to one another \( (\theta = 0.50) \), this law holds true; however, the law is violated when two loci are linked \( (\theta < 0.50) \) to one another such that the transmission of one is not independent of transmission of the other. Estimating the distance between linked loci by assessing the frequency of recombination between them allows the development of an order of the markers relative to one another.

Once a gross localization for a disease or trait locus has been identified, either through linkage analysis or from some other approach (e.g., clues from chromosomal rearrangements), it is necessary to home in on the actual gene. This process is always complicated, but it can be simplified by the use of mapping resources, many of which were developed as a direct result of the Human Genome Initiative. No single mapping resource is best for all situations. Regardless of the mapping approach utilized, the resultant locus order along a chromosome should be identical.

Genetic maps order polymorphic markers by specifying the amount of recombination between markers, whereas physical maps quantify the distances among markers in terms of the number of base pairs of DNA. For small recombination fractions (usually \(<10\text{–}12\%\) ), the estimate of the recombination fraction provides a very rough estimate of the physical distance. In general, 1% recombination corresponds to one crossover per 100 meioses and is equivalent to about a million base pairs of DNA and is defined as one centimorgan (cM). Physical measurements of DNA are often described in terms of thousands of kilobases (10 kb of DNA is equivalent to 10,000 bp). A specific type of physical map, the radiation hybrid (RH) map (see below), also allows quantification of the length of a segment of DNA. The RH map distance is measured in centirays (cR), and on average throughout the genome 1 cR is equivalent to about 30,000 bp of DNA, although this estimate varies according to radiation dose. Estimates of distance from physical and genetic maps of the identical region may vary dramatically (Table 1.5) throughout the genome. A summary of the characteristics of different types of maps is shown in Table 1.6.

Physical Mapping

The purpose of a physical map of the genome is identical to that of a genetic map: to order pieces of DNA and, subsequently, genes. However, the materials utilized and
the average resolutions of various mapping methods differ. Some of the oldest available physical maps of the genome are restriction maps, which identify sites at which an enzyme cuts (digests) a specific sequence of DNA. Contig maps are developed by cloning pieces of DNA into vectors, such as yeast artificial chromosomes (YACs), bacterial artificial chromosomes (BACs), or cosmids, and then ordering them by their overlapping sequences. Sequence-tagged site (STS) maps utilize unique stretches of DNA to identify particular clones. Expressed sequence tag (EST) maps order the coding stretches of DNA. The DNA sequence maps order specific stretches of DNA at the level of single base pairs. Radiation hybrid maps are developed by exposing DNA to high doses of radiation, thereby breaking the DNA into small pieces. The frequency with which particular markers are retained in a piece of DNA is scored, and this provides an estimate of the relative order and distance between markers in centirays. As with genetic maps, the basis of the RH maps is statistical: The relative order of a marker is determined in a manner analogous to calculating LOD (logarithm of the odds of linkage) scores.

### TABLE 1.5. Estimated Physical and Genetic Lengths of Selected Chromosomes

<table>
<thead>
<tr>
<th>Chromosome</th>
<th>Physical Length (Mb)</th>
<th>Genetic Length (cM)</th>
<th>Length (cR)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>263</td>
<td>305</td>
<td>7894 (31.4)</td>
</tr>
<tr>
<td>2</td>
<td>255</td>
<td>271</td>
<td>6973 (34.4)</td>
</tr>
<tr>
<td>3</td>
<td>214</td>
<td>237</td>
<td>7785 (25.9)</td>
</tr>
<tr>
<td>4</td>
<td>203</td>
<td>244</td>
<td>2867 (24.4)</td>
</tr>
<tr>
<td>5</td>
<td>194</td>
<td>224</td>
<td>5611 (32.6)</td>
</tr>
<tr>
<td>6</td>
<td>183</td>
<td>207</td>
<td>6095 (28.3)</td>
</tr>
<tr>
<td>7</td>
<td>171</td>
<td>178</td>
<td>6606 (24.3)</td>
</tr>
<tr>
<td>8</td>
<td>155</td>
<td>172</td>
<td>3996 (36.6)</td>
</tr>
<tr>
<td>9</td>
<td>145</td>
<td>146</td>
<td>4513 (30.0)</td>
</tr>
<tr>
<td>10</td>
<td>144</td>
<td>181</td>
<td>5423 (25.1)</td>
</tr>
<tr>
<td>11</td>
<td>144</td>
<td>150</td>
<td>4858 (27.9)</td>
</tr>
<tr>
<td>12</td>
<td>143</td>
<td>160</td>
<td>5002 (26.9)</td>
</tr>
<tr>
<td>13</td>
<td>114</td>
<td>130</td>
<td>3306 (27.8)</td>
</tr>
<tr>
<td>14</td>
<td>109</td>
<td>122</td>
<td>3513 (25.0)</td>
</tr>
<tr>
<td>15</td>
<td>106</td>
<td>154</td>
<td>2822 (29.8)</td>
</tr>
<tr>
<td>16</td>
<td>98</td>
<td>157</td>
<td>2735 (34.4)</td>
</tr>
<tr>
<td>17</td>
<td>92</td>
<td>208</td>
<td>3039 (28.6)</td>
</tr>
<tr>
<td>18</td>
<td>85</td>
<td>143</td>
<td>2977 (26.8)</td>
</tr>
<tr>
<td>19</td>
<td>67</td>
<td>148</td>
<td>2122 (29.6)</td>
</tr>
<tr>
<td>20</td>
<td>72</td>
<td>122</td>
<td>2010 (33.8)</td>
</tr>
<tr>
<td>21</td>
<td>50</td>
<td>114</td>
<td>1562 (23.7)</td>
</tr>
<tr>
<td>22</td>
<td>56</td>
<td>81</td>
<td>1522 (26.9)</td>
</tr>
<tr>
<td>X</td>
<td>164</td>
<td>220</td>
<td>3644 (42.5)</td>
</tr>
</tbody>
</table>

*From Morton (1991).*

*Numbers in parenthesis give kilobases per centiray.*

*Data extracted from Stanford Human Genome Center (see Appendix for website).*
## TABLE 1.6. Summary Characteristics of Genome Maps of Selected Types

<table>
<thead>
<tr>
<th>Map Type</th>
<th>Measurement of Distance</th>
<th>Material Needed</th>
<th>Caveats and Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genetic maps</td>
<td>Recombination frequency (θ) or centimorgan (cM); 1% recombination approximately equal to 1 cM</td>
<td>Reference pedigrees; polymorphic markers</td>
<td>Extremely sensitive to genotyping error; usually not useful in areas &lt;2 cM</td>
</tr>
<tr>
<td>Meiotic breakpoint maps</td>
<td>Not applicable</td>
<td>Reference pedigrees; polymorphic markers</td>
<td>Good mechanism for minimizing genotyping to determine marker order using statistical techniques to minimize number of recombination events</td>
</tr>
<tr>
<td>Radiation hybrid maps</td>
<td>Centiray (cR): 1 cR represents 1% breakage between two markers</td>
<td>Somatic cell hybrid panel; markers not necessarily polymorphic</td>
<td>Maps developed using statistical techniques that assess frequency of chromosome breakage</td>
</tr>
<tr>
<td>Sequence-tagged site (STS) maps</td>
<td>Not applicable</td>
<td>Clones from which STSs are derived must be ordered</td>
<td>Resulting maps have landmarks assayed by polymerase chain reaction (PCR), but markers are often not polymorphic</td>
</tr>
<tr>
<td>Restriction maps</td>
<td>Tens of thousands of base pairs of DNA; 1000 bp of DNA is termed a kilobase</td>
<td>Genomic DNA</td>
<td></td>
</tr>
<tr>
<td>Expressed sequence tag (EST) maps</td>
<td>Not applicable</td>
<td>Genomic DNA</td>
<td></td>
</tr>
</tbody>
</table>
(Falk, 1991; Lange and Boehnke, 1992) and is reported in terms of odds in favor of a placement relative to another placement. All these maps were generated as steps in the process of obtaining the human genome sequence, which is the ultimate physical map. Even when declared completed, there will be some small regions of the genome that are difficult to sequence completely. The various physical maps will still be helpful in spanning these regions.

**Genetic Mapping**

The study of human inherited disease has benefited from numerous experiments in other organisms. Although mapping in humans has a relatively recent history, the idea of a linear arrangement of genes on a chromosome was first proposed in 1911 by T. H. Morgan from his work with the fruit fly *Drosophila melanogaster*. The possibility of a genetic map was first formally investigated by A. H. Sturtevant, who ordered five markers on the X chromosome in *Drosophila* and then estimated the relative spacing among them.

In experimental organisms, genetic mapping of loci involves counting the number of recombinant and nonrecombinant offspring of selected matings (Table 1.7).

**TABLE 1.7. Example Development of Genetic Map Using Four Linked Loci, A, B, C, and D, Scored in 100 Offspring**

<table>
<thead>
<tr>
<th>Loci Scored</th>
<th>Number of Recombinants</th>
<th>Frequency of Recombination</th>
</tr>
</thead>
<tbody>
<tr>
<td>A–B</td>
<td>10</td>
<td>0.10</td>
</tr>
<tr>
<td>A–C</td>
<td>3</td>
<td>0.03</td>
</tr>
<tr>
<td>A–D</td>
<td>15</td>
<td>0.15</td>
</tr>
<tr>
<td>B–C</td>
<td>7</td>
<td>0.07</td>
</tr>
<tr>
<td>B–D</td>
<td>5</td>
<td>0.05</td>
</tr>
<tr>
<td>C–D</td>
<td>12</td>
<td>0.12</td>
</tr>
</tbody>
</table>

**B. Process**

1. Determine which two loci have the highest frequency of recombination between them; these two loci are the farthest apart on the map.
2. Fit the other loci into the map like pieces of a puzzle.

**C. Resulting Genetic Map**

The percent recombination loci order is A, C, B, D.

---

*This example demonstrates no evidence for interference (see text). Positive interference would be manifested as a decrease in overall map length from what would be expected by adding the pairwise distances. In this example, if (A–C) + (C–B) + (B–D) > A–D, positive interference may be present.*
Genetic mapping in humans is usually more complicated than in experimental organisms for many reasons, including our inability to design specific matings of individuals, which limits the unequivocal assignment of recombinants and nonrecombinants. Therefore, maps of markers in humans are developed by means of one of several statistical algorithms used in computer programs such as CRIMAP and MAPMAKER (Lander and Green, 1987), CLINKAGE and MULTIMAP (Matise et al., 1994), and MAP-O-MAT (Matise and Gitlin, 1999). Genetic maps can assume equal recombination between males and females or can allow for sex-specific differences in recombination since it has been well established that there are substantial differences in recombination frequencies between men and women; on average, the female map is twice as long as the male map (Li et al., 1998). These maps are generally produced utilizing a single set of reference pedigrees, such as the those developed by the Centre d’Étude du Polymorphisme Humain (CEPH) (Dausslet et al., 1990), which are mostly comprised of three-generation pedigrees with a large number of offspring (average 8.5). Both sets of maternal and paternal grandparents are usually available, so linkage phase frequently can be established. The collection of CEPH pedigrees, in its entirety, consists of more than 60 pedigrees and includes more than 600 individuals; DNA from this valuable resource is available through the Coriell Institute for Medical Research. The complexity of the underlying statistical methods used to generate genetic maps renders them sensitive to marker genotyping errors, particularly in small intervals (Buetow, 1991), and these maps are less useful in small regions of less than about 2 cM. While marker order is usually correct, genotyping errors can result in falsely inflated estimates of map distances.

Interference and Genetic Mapping

Another factor complicating genetic mapping is interference, where the probability of a crossover in a given chromosomal region is influenced by the presence of an already existing crossover. In positive interference, the presence of one crossover in a region decreases the probability that another crossover will occur nearby. Negative interference, the opposite of positive interference, implies the formation of a second crossover in a region is made more likely by the presence of a first crossover. Most documented interference has been positive, but some reports of negative interference exist in experimental organisms. Interference is very difficult to measure in humans because exceedingly large sample sizes, usually on the order of 300–1000 fully informative meiotic events, are required to detect it (Weeks et al., 1994).

The investigation of interference is important because accurate modeling of interference will provide better estimates of true genetic map length and intermarker distances and more accurate mapping of trait loci. Interference ($I$) is frequently measured in terms of the coefficient of coincidence (c.c.) in genetic crosses where three separate linked markers can be scored. The coefficient of coincidence is the ratio of the observed number of double crossovers to the expected number of double crossovers assuming no interference. When $I > 0$, interference is present (positive); when $I < 0$, interference is negative; when $I = 0$, there is no evidence for interference and recombination fractions across intervals are additive.
For example, assume three loci whose order is A–B–C. If the distance between A and B is 10 cM and between B and C is 5 cM, when \( I = 0 \), the distance between A and C is 15 cM. As noted earlier, the frequency of recombination in humans is generally decreased near the centromeric region of chromosomes, tends to be greater near the telomeric regions, and is increased in females when compared to males. It should be further noted that genetic map distance is not tied directly to physical map distance.

Several mathematical formulas have been developed to account for interference in predicting an additive measure of genetic map distance from recombination frequencies in human linkage studies. These mapping functions include those developed by Kosambi (1944), Rao (Watzke et al., 1977), and Haldane (1919). Haldane’s map function assumes the absence of interference, while Kosambi’s map function assumes interference is large at small genetic distances but decreases as the genetic distance between two loci increases. A program in the LINKAGE utility package, MAPFUN, translates recombination frequencies into map distances and vice versa under a variety of mapping functions. This package, along with a comprehensive listing of other available linkage analysis programs, is available at http://www.linkage.rockefeller.edu.

**Meiotic Breakpoint Mapping**

Meiotic breakpoint maps, an outgrowth of genetic maps, are graphical descriptions of critical, confirmed recombination events within reference pedigrees (Fig. 1.13).

<table>
<thead>
<tr>
<th>Family</th>
<th>ID</th>
<th>DSS818</th>
<th>DSS804</th>
<th>DSS816</th>
<th>DSS812</th>
</tr>
</thead>
<tbody>
<tr>
<td>1331</td>
<td>9</td>
<td>●</td>
<td>○</td>
<td>●</td>
<td>○</td>
</tr>
<tr>
<td>1331</td>
<td>10</td>
<td>●</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>1332</td>
<td>3</td>
<td>○</td>
<td>●</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>1332</td>
<td>12</td>
<td>○</td>
<td>●</td>
<td>●</td>
<td>○</td>
</tr>
<tr>
<td>1333</td>
<td>4</td>
<td>○</td>
<td>○</td>
<td>●</td>
<td>○</td>
</tr>
<tr>
<td>1333</td>
<td>9</td>
<td>○</td>
<td>○</td>
<td>●</td>
<td>○</td>
</tr>
<tr>
<td>1344</td>
<td>5</td>
<td>○</td>
<td>○</td>
<td>●</td>
<td>○</td>
</tr>
<tr>
<td>1347</td>
<td>6</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>●</td>
</tr>
<tr>
<td>1347</td>
<td>10</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>●</td>
</tr>
<tr>
<td>1362</td>
<td>9</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>●</td>
</tr>
<tr>
<td>1362</td>
<td>11</td>
<td>○</td>
<td>○</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>884</td>
<td>3</td>
<td>○</td>
<td>○</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>884</td>
<td>11</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>●</td>
</tr>
</tbody>
</table>

**Figure 1.13.** Sample meiotic breakpoint map from data at the CHLC. Shaded circles indicate markers that recombine with those having open circles. Squares indicate marker is either uninformative or not genotyped. To test whether a marker is located between DSS818 and DSS804, an investigator would need to genotype individuals 9 and 10 in family 1331 and 3 in 1332, in addition to their parents. In typical practice, the laboratory is blinded to which individuals are recombinants to avoid potential bias. Consequently, two or three nonrecombinant siblings for each known recombinant individual are also genotyped.
Once a meiotic breakpoint map for a region has been developed, a marker whose location is nearby can be genotyped in a subset of pedigrees in which critical recombination events have occurred. Limiting the genotyping efforts to a small number of specific individuals within pedigrees greatly increases genetic mapping efficiency.

**Disease Gene Discovery**

Disease gene discovery is greatly facilitated by the availability of dense genetic maps. Linkage analysis for the localization of disease genes boils down to the “simple” idea of counting recombinants and nonrecombinants, but in humans this process is complicated for a variety of reasons. For example, the generation time is long in humans such that large, multigenerational pedigrees in which a disease or trait is segregating are rare; scientists cannot dictate matings or exposures, nor can they require participation of specific individuals in a study. Thus, the process of linkage analysis in humans requires a statistical framework in which various hypotheses about the linkage of a trait locus and marker locus can be considered. How far apart are the disease and marker, and how certain is the conclusion of linkage?

Linkage analysis has traditionally been performed using either a parametric approach or a nonparametric approach. Parametric approaches require the assumption of a genetic model and the specification of various parameters such as disease allele frequency and penetrance. In contrast, nonparametric methods do not require specification of a genetic model and thus do not suffer from the potentially hazardous effects of model misspecification. When the genetic model for a disease is clearly known, parametric- (likelihood-) based methods are more powerful than the nonparametric approaches.

**Example 1: Linkage Analysis in Pedigree with Unlinked Marker.** For illustrative purposes, consider the pedigree in Figure 1.14a. In this pedigree, shaded individuals are affected with disease and unshaded individuals are unaffected. Results for a genetic marker are shown underneath each individual. In pedigrees in which the genetic model can be deduced with a high degree of certainty, linkage analysis can be broken down into five steps.

It should be emphasized that this example is a simplified version of an extremely complicated process.

**Step 1: State Components of Genetic Model.** For parametric linkage analysis, the genetic model must be specified. Components of the genetic model include the inheritance pattern of the disease locus (autosomal or sex linked; dominant, recessive, or codominant), disease allele frequency and penetrance, and frequency of phenocopies and new mutation. Rough estimates of the disease allele frequency and penetrance can often be obtained from the literature or from computer databases such as Online Mendelian Inheritance in Man (http://www3.ncbi.nlm.nih.gov/Omim/); estimates of the rate of phenocopies and new mutation are frequently guesses, included as a nuisance parameter in some cases to allow for the fact that these can exist. Linkage analysis using LOD scores is relatively robust to modest
Figure 1.14. (a) Pedigree in which a rare, fully penetrant autosomal dominant disease is segregating. Genotypes are shown beneath pedigree symbols. (b) Putative disease genotype listed beneath pedigree figure. (c) Assignment of disease-associated haplotypes. Results on the left of the bar were transmitted from the male parent and those on the right of the bar were transmitted from the female parent. In the third generation, meioses are scored as recombinant (R) or nonrecombinant (NR). Since the number of recombinant meioses is equal to the number of nonrecombinant meioses, this pedigree is consistent with nonlinkage between the disease and marker loci.
misspecification of the disease allele frequency and penetrance, but misspecification of whether the disease is dominant or recessive can lead to incorrect conclusions of linkage or nonlinkage (Clerget-Darpoux et al., 1986). In this example, the disease allele will be assumed to be rare and to function in an autosomal dominant fashion with complete penetrance; the disease locus will be assumed to have two alleles, \( N \) (for normal or wild type) and \( A \) (for affected or disease). In addition, mutation and phenocopies are assumed to be absent. These assumptions allow substantial simplification of the problem, as outlined further below.

In addition to specifying the parameters of the disease locus, frequencies for the alleles at the marker locus are required. In pedigrees in which genotypes are missing in founding individuals (due either to an unsampled individual or to laboratory complications), the misspecification of allele frequencies can have substantial impact, leading to incorrect conclusions of linkage and nonlinkage and biased estimates of the recombination fraction (Ott, 1992; Knowles et al., 1992). Other factors can also wreak havoc in a linkage analysis. Such factors include linkage heterogeneity and scoring errors in the pedigree structure, diagnostic status, or marker genotyping.

**Step 2: Assign Underlying Disease Genotypes Given Information in Genetic Model** (Fig. 1.14). In preparation for scoring recombinant and nonrecombinant individuals in a pedigree, the relationship between genotype and phenotype as defined by the genetic model can be used to assign the underlying genotype of pedigree members at the disease locus. The assumption of complete penetrance of the disease allele allows all unaffected individuals in the pedigree to be assigned a disease genotype of \( NN \). Since the disease allele is assumed rare, the disease genotype for affected individuals can be assigned a disease genotype of \( AN \). In other words,
since the disease allele is rare, the chance that an affected individual is homozygous for the disease allele is so small that, for the purposes of this example, it can be considered to be zero.

Again, it is important to emphasize that this example is a simplified version; most linkage analysis is performed by computer analysis that allows the consideration of the small probability that a founder in such a pedigree (e.g., individual I-1) is homozygous for the disease allele. When a computer program performs this analysis, it assigns probabilities for genotypes $AA$ and $AN$ in individuals like I-1 by using user-specified information on disease allele frequencies.

Step 3: Determine Putative Linkage Phase. Individual II-1 has inherited the disease trait together with marker allele 2 from his affected mother. Thus, the $A$ allele at the disease locus and the 2 allele at the marker locus were inherited in the gamete transmitted to II-1. There are two mutually exclusive hypotheses to consider: The null hypothesis is that the disease and marker loci are unlinked to one another. If the loci are genetically unlinked, there will be an approximately equal number of recombinant and nonrecombinant gametes among the offspring of II-1. The alternate hypothesis is that the disease and marker loci are linked to one another. If the loci are genetically linked, there will be more nonrecombinants than recombinants among the offspring ("meiotic events") of II-1. The basic idea is to consider these two competing hypotheses of linkage versus nonlinkage and determine which hypothesis best describes the available data from the pedigree.

Thus, the putative linkage phase (the disease allele “segregates” with marker allele 2) has been established, and this phase can be tested in subsequent generations.

Step 4: Score Meiotic Events as Recombinant or Nonrecombinant. For this mating type, there are four possible gametes from the affected parent II-1: $N1$, $N2$, $A1$, and $A2$. Based on the putative linkage phase assigned above in step 3, gametes $A2$ and $N1$ are recombinant. Thus, all affected offspring of II-1 and II-2 who have inherited marker allele 2 from their father will be scored as nonrecombinant for the disease and marker; affected offspring who have inherited the $I$ allele will be scored as recombinant for the disease and marker. Similar reasoning applies to the unaffected offspring, except that the unaffected offspring who have inherited allele $I$ are nonrecombinant and those who have inherited allele 2 are recombinant.

In this pedigree (Fig. 1.14c), five offspring of II-1 are recombinant and five are nonrecombinant. Thus, out of 10 scorable meiotic events, the number of recombinant gametes is equal to the number of nonrecombinant gametes. These data are consistent with the hypothesis of nonlinkage between the disease and marker loci.

One frequent question arises—why is the transmission of the 2 allele from the affected grandmother to the affected son not counted as a meiotic event? Or, why are there not 11 instead of 10 meioses in this pedigree? The answer is that we do not know the linkage phase in individual II-1; we are just using the transmission from his affected mother to him to determine our hypothesis about what the linkage phase would be if the disease and marker loci were linked to one another.

Step 5: Calculate and Interpret LOD Scores. The pedigree data are used to consider both the hypothesis of linkage between the disease and marker locus and
nonlinkage between the disease and marker locus. Morton (1955) suggests a likeli-
hood ratio approach in which the likelihood of the pedigree and marker data is
calculated under the null hypothesis that assumes free recombination ($\theta = 0.50$)
between the disease and marker loci, where $\theta$ represents the recombination fraction,
and then compared to the likelihood of the hypothesis of linkage between the disease
and marker loci. This likelihood of the pedigree data under the hypothesis of linkage
between the disease and marker loci is calculated at various increments of $\theta < \frac{1}{2}$
within the range of allowable values (0.00–0.49), representing unique subhypoth-
eses of linkage between the disease and marker loci. The likelihood ratio (LR) is
constructed as $L(\text{pedigree}|\theta = x)/L(\text{pedigree}|\theta = 0.50)$, where $x$ is some value of $\theta$.

The likelihood of observing the pedigree data is just $Q R (1 - Q) NR$, where $R$ and $NR$
are the number of recombinants for the two phases and $N$ is the total number of scored
offspring ($R + NR = N$); this likelihood is the numerator in the ratio of likelihoods. The
denominator is the likelihood assuming that the two loci are unlinked (i.e., when $\Theta = 0.5$). Thus, the ratio of the likelihoods of the two hypotheses is constructed as

$$LR = \frac{\theta^R (1 - \theta)^{NR}}{(0.5)^R (0.5)^{NR}}$$

and reduces to

$$LR = \frac{\theta^R (1 - \theta)^{NR}}{0.5^N}$$

Typically, the base-10 logarithm of this ratio is taken to obtain a LOD score [Note: $\zeta(\theta)$
is sometimes used to denote a LOD score]. A LOD score of 3 indicates odds of $10^3 : 1$
(1000 : 1) in favor of linkage and is considered to be conclusive evidence for linkage
between two markers (or a marker locus and disease locus) in most cases. A LOD
score of $-2$ indicates odds of $10^{-2} : 1$ (0.01 : 1) in favor of linkage, or more
commonly, odds of 100 : 1 against linkage of the two markers. A LOD score of $-2$
or less is considered to be conclusive evidence that the two markers under study are
unlinked at the specified recombination fraction. It is important to note that this “exclu-
sion” of a disease gene from a region is only valid under the specific set of assumptions
made for the analysis. For example, if an analysis were made assuming an autosomal
dominant trait but the trait was actually autosomal recessive, it could be falsely
excluded from a region. The LOD scores between the values of $-2$ and 3 are considered
inconclusive and warrant additional study. A study is “expanded” by rendering the
currently available family data more informative (i.e., testing the family with different
or more informative markers) or by increasing the number of families under study.
A LOD score should always be considered in conjunction with its respective
estimate of $\Theta$.

It is important to note that using a LOD score of 3.0 (odds 1000 : 1 in favor of
linkage) as a test statistic does not equate with a type 1 error rate ($\alpha$) of 0.001,
which is much more stringent than the standardly used type I error rate of 0.05 in
most statistical analyses. Because of the prior probability of linkage between two
traits, the true $p$-value associated with a LOD score of 3.0 is approximately 0.04,
consistent with a false-positive rate of 1 in 25. In other words, when evidence for linkage between two loci is declared significant at a LOD score of 3.0, there is a 4% chance it is a spurious positive result (for more detail, see Ott, 1999).

Table 1.8, Example 1, shows the two-point LOD scores for the marker at a variety of hypotheses about the estimate of the recombination fraction between the disease and marker locus. In this example, the highest LOD score is $-0.09$ at $\theta = 0.40$ and at no value of $\theta$ is the LOD score positive, let alone $\geq 3.0$, so this pedigree demonstrates no evidence in favor of linkage between the disease and marker loci. However, the pedigree does provide important information about where the disease locus is not located. Visual inspection of the LOD score data suggests that the value of $\theta$ at which the LOD score exceeds $-2.0$ is between 0.10 and 0.15, so approximately 13 cM on either side of the marker locus can be excluded as harboring the disease gene, for a total exclusion of 26 cM as a result of typing this marker!

Several well-tested computer programs including the LINKAGE computer package (Lathrop et al., 1984), FASTLINK (Schaffer et al., 1994; Cottingham et al., 1993), VITESSE (O’Connell and Weeks, 1995), Genehunter (Kruglyak et al., 1996), and Allegro (Gudbjartsson et al., 1999) are often utilized for calculation of LOD scores.

### Example 2. Linkage Analysis in Pedigree with Linked Marker

Consider the pedigree in Figure 1.15a.

**Step 1: State Components of Genetic Model.** Assume the genetic model is the same as in Example 1 (rare, autosomal dominant disease allele with complete penetrance, no mutation, no phenocopies).

**Step 2: Assign Underlying Disease Genotypes Given Information in Genetic Model.** The assignment of disease genotypes to pedigree members is the same as in Example 1 (Fig.1.15b). Unaffected individuals have an underlying disease genotype of $NN$ and affected individuals have a disease genotype of $AN$.

**Step 3: Determine Putative Linkage Phase.** The assignment of putative linkage phase is identical to that in Example 1, so that the disease allele is transmitted in the same gamete as marker allele 2.

**Step 4: Score Meiotic Events as Recombinant or Nonrecombinant** (Fig. 1.15c). In this example, all five of the affected children have inherited marker allele 2 from their affected father and are thus nonrecombinant. In addition, four of the five unaffected children have inherited marker allele 1 from their affected father. These four offspring, too, are nonrecombinant with respect to the disease and marker loci. Individual III-6, however, is unaffected and has inherited marker allele.
Figure 1.15. (a) Pedigree in which a rare, fully penetrant autosomal dominant disease is segregating. Genotypes are shown beneath pedigree symbols. (b) Putative disease genotype listed beneath pedigree figure. (c) Assignment of disease-associated haplotypes. Results on the left of the bar were transmitted from the male parent and those on the right of the bar were transmitted from the female parent. In the third generation, meioses are scored as recombinant (R) or nonrecombinant (NR). Since the number of nonrecombinant gametes is larger than the number of recombinant gametes, this pedigree is consistent with linkage between the disease and marker loci.
2 from his affected father. This individual is scored as a recombinant individual (a recombinant meiosis). This nonrandom segregation of the disease allele and a marker allele within a pedigree is consistent with linkage between the disease and marker locus.

Step 5: Calculate and Interpret LOD Scores. The LOD scores for this pedigree at a variety of hypotheses about the recombination fraction between the disease and marker loci are shown in Table 1.8, Example 2. No LOD score at any tested recombination fraction is higher than 3.0, so this pedigree does not provide significant evidence in favor of linkage between the disease and marker locus. The highest LOD score in this pedigree is 1.6 so that more data from additional pedigrees need to be generated in order to interpret the results. However, this LOD score is probably “interesting” enough that it would indeed warrant such follow-up. Here, it is critical to note that the recombination event occurs in an unaffected individual. If the assumption of complete penetrance of the disease allele is incorrect, the assignment of this meiotic event as a “recombinant” could be in error. Thus, it is incumbent on the investigator to weigh carefully information about disease gene localization gleaned from unaffected family members.

The maximum-likelihood estimate (MLE) of the recombination fraction is that value of $\theta$ at which the LOD score is largest. It is this value of the LOD score (the highest LOD score) with which the determination of statistical significance is made. In this case, the maximum LOD score is 1.60, which occurs when $\theta = 0.10$, and so 0.10 is the MLE for $\theta$ in this example (note that this result is identical to the direct estimate for $\theta$ obtained earlier). This result does not meet the established criteria of odds $\geq 1000:1$ (a LOD score $\geq 3$) for concluding evidence for
linkage, yet it is interesting in and of itself and perhaps merits further investigation by the genotyping of additional tightly linked markers or additional families.

An intuitive, direct estimate of the recombination frequency (usually designated as $\theta$), which is related to the genetic distance between the disease and marker locus, is just the proportion of recombinant meioses counted among total meioses scored. In this example, the direct estimate of the recombination frequency, based on the observation of 1 recombinant out of 10 scorable meioses, is $\frac{1}{10}$ or 10%. This direct approach is possible in this example because these are phase-known events so that recombinants and nonrecombinants can be counted.

When significant evidence for linkage is found, a formal presentation of results of a linkage analysis is incomplete without an indication of a support interval for the MLE of $\theta$. A description of the calculation of these support intervals, usually performed using the “one LOD score down” method, can be found in either Ott (1999) or Conneally et al. (1985). In this approach, all values of the recombination fraction that fall within one LOD score below the highest attained LOD score are considered within the support interval.

**Example 3. Linkage Phase Unknown.** Next consider an example in which the grandparents are unavailable (Fig. 1.16). In this pedigree, the linkage phase cannot be established with certainty; however, only one of two linkage phases is possible. Under phase 1, the disease allele is segregating with the 1 allele. Alternatively, the disease allele could be segregating with the 2 allele (phase 2). Under each of these alternate scenarios, it can be determined whether or not the offspring—these 10 scorable meioses—are recombinant or nonrecombinant gametes. In this example the LOD scores are calculated a little differently:

$$
\text{LOD} = \log_{10} \left[ \frac{1}{2} \left( \frac{(\theta^R (1 - \theta)^{NR})}{(\theta = 0.5)^N} + \frac{(\theta^R (1 - \theta)^{NR})}{(\theta = 0.5)^N} \right) \right] 
$$

where $NR$ and $R$ are the number of nonrecombinants for the two phases.

![Figure 1.16. Pedigree for Example 3 for calculation of LOD score for linkage-phase-unknown pedigree.](image-url)
Note that the uncertainty associated with the unknown linkage phase reduces the overall linkage information. The highest LOD score obtained in this pedigree is 1.3 (Table 1.7), as opposed to the highest LOD score of 1.6 in the earlier pedigree. Although this may not seem like a substantial loss of information, it does influence the results when working with pedigree samples that are large enough to give evidence in favor of linkage. If that much information is lost on a series of 10 pedigrees, a substantial amount of potential linkage information has been lost.

**Additional Notes.** Note that in the pedigree in Figure 1.15c, the 2 allele is segregating with all affected individuals and with only one of five unaffected individuals. One might characterize this observation as an “association” between the 2 allele and the disease allele within this family. By definition, two traits that are linked to one another will show an association within a family, but the associated allele may vary between families linked to the same locus.

**Information Content in a Pedigree**

For the meiotic events of a parent to be scorable for linkage analysis, the parent must be heterozygous at both loci (disease and marker) under consideration (Fig. 1.17). The family studied in the above examples is large and relatively atypical for a human pedigree since all family members are available for sampling. How much information is required to obtain significant results from a linkage analysis? The answer to this question can be complicated and highly dependent on the inheritance pattern of the disease (i.e., dominant, recessive, or sex linked), the penetrance of the disease, whether sporadic cases of the disease (phenocopies) may be present, and other factors. As a guideline, each scorable, phase-known meiotic event (with no recombination between the trait and marker locus) contributes about 0.30 to the LOD score. Table 1.9 gives some additional information about the contribution of each meiotic event to the LOD score at a variety of recombination fractions. Thus, at a minimum, 10 phase-known meiotic events demonstrating no recombination between the trait and marker locus are required to obtain a LOD score of 3.0. In general, human pedigrees have few phase-known meiotic events since some individuals may be unable or unwilling to participate in the study. Computer simulation studies should be performed to assess the power of an available dataset to detect linkage under an assumed genetic model prior to initiating a screen of the entire genome to detect linkage. Several programs including SIMLINK and SLINK are available for performing these power studies (Boehnke, 1986; Ploughman and Boehnke, 1989; Ott, 1989; Weeks et al., 1990).

**Disease Gene Localization**

Once linkage between a disease locus and marker locus is established via two-point linkage analysis, the next step is to identify the smallest region of the genome that should contain the disease gene, the minimum-candidate region (MCR). Two approaches to disease gene localization are generally used: multipoint linkage analysis and haplotype analysis.

Multipoint linkage analysis is a statistical technique using available genetic maps in which the order of markers is relative to one another and the distances between
Figure 1.17. Pedigree examples demonstrating families that are informative and noninformative for linkage analysis. Both the autosomal dominant and autosomal recessive disease are assumed to be fully penetrant. Shaded individuals are affected with the disease, unshaded individuals are unaffected, and stippled individuals are asymptomatic gene carriers (in the recessive disease). Marker results are indicated beneath each individual. 

**Autosomal dominant diseases:** The affected parent, who is by definition homozygous at the disease locus, must also be heterozygous at the marker locus. (a) Fully informative pedigree: All three offspring can be scored for linkage analysis. (b) Uninformative pedigree: Cannot tell which allele is transmitted with the disease allele. (c) Partially informative pedigree (an “intercross”): Homozygous offspring of this mating contribute significantly to linkage analysis.

**Autosomal recessive diseases:** For pedigrees to be fully informative, both parents, who are heterozygous at the disease locus by virtue of the fact that they have at least one affected child with a recessive disease, must also be heterozygous for different alleles at the marker locus. (d) Fully informative pedigree: Both paternal and maternal gametes can be scored with respect to disease and marker locus. (e) Partially informative pedigree: Only paternally contributed gametes contribute to linkage analysis. (f) Uninformative pedigree: Both parents are homozygous at the marker locus. (g) Partially informative pedigree: Homozygous offspring contribute significantly to the linkage analysis.

<table>
<thead>
<tr>
<th>(\theta)</th>
<th>Number of Informative Meioses Needed for LOD = 3</th>
<th>Expected LOD per Meiosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.00</td>
<td>10</td>
<td>0.30</td>
</tr>
<tr>
<td>0.01</td>
<td>11</td>
<td>0.28</td>
</tr>
<tr>
<td>0.02</td>
<td>12</td>
<td>0.26</td>
</tr>
<tr>
<td>0.05</td>
<td>14</td>
<td>0.21</td>
</tr>
<tr>
<td>0.10</td>
<td>19</td>
<td>0.16</td>
</tr>
<tr>
<td>0.20</td>
<td>36</td>
<td>0.08</td>
</tr>
<tr>
<td>0.25</td>
<td>53</td>
<td>0.06</td>
</tr>
</tbody>
</table>
the markers are known. Multipoint linkage analysis allows the simultaneous consideration of genotypes from multiple linked loci. This technique is useful for localization of disease genes between two markers and for maximizing the informativeness of a series of markers. In a general sense, given a series of markers of known location, order, and spacing, the likelihoods of the pedigree data are sequentially calculated for the disease gene to be at any position within the known map of markers. The multipoint LOD scores are typically graphed as in Figure 1.18. The x axis represents the genetic distance between the markers and the y axis represents the LOD score. In this example the most likely location for the disease is in the 7-cM interval bounded by IL9 and D5S178. In this interval, the LOD score reaches its highest value of approximately 23.5. The next most likely interval is the interval outside IL9, where the maximum LOD score reaches approximately 21.0. Note that the interval between D5S178 and D5S210 has a LOD score that maximizes at approximately 18. This interval between D5S178 and D5S210 is outside the approximate 99% confidence interval for disease gene localization because it does not fall with the three-LOD unit support interval.

Haplotype analysis is a tool for ordering alleles on a chromosome. It complements multipoint linkage analysis in that it provides visual confirmation of statistical testing. Given the order of the genetic markers on a chromosome, we can establish with varying degrees of certainty the alleles that were passed from parent to offspring in any specific gamete by haplotype analysis and identify critical recombination events that flank the MCR. Haplotype analysis can be done by hand or with the use of a computer program such as SIMWALK2 (Weeks et al., 1995). Through haplotype analysis in Mendelian disease, critical recombination events may allow us to identify the upper and lower bounds of the disease gene interval. The process of haplotyping involves identifying which alleles were transmitted in the same

Figure 1.18. The most likely location for the disease gene is in the 7-cM interval bounded by IL9 and D5S178. The maximum LOD score is 23.50. Narrow support intervals for disease gene localization are calculated by subtracting 3 from the maximum LOD score; all locations within this three-LOD unit are within the approximate 99% confidence interval for disease gene localization.
gamete from each parent and identifying where recombination events between markers and/or between markers and the disease gene are located. The main goal is to identify markers that flank the disease gene. In the presence of reduced penetrance, recombination events that occur in affected individuals tend to be associated with a higher degree of certainty than recombination events that involve unaffected individuals because an unaffected individual could either be a non–gene carrier or a non–penetrant carrier. The confidence with which a recombination event defines a boundary of the MCR then is heavily dependent on the penetrance.

An example of haplotype analysis is shown in Figure 1.19. In the pedigree on the left, the unordered genotypes for three markers are shown underneath the pedigree symbols. The markers are arranged linearly and are tightly linked to one another such that 2 cM with intermarker distances. The task in haplotype analysis is to assign the allelic combinations that were transmitted in the gamete from parent to offspring. In the pedigree on the left one can see that from the affected grandfather the gamete with the combination of alleles 3/4/1 was transmitted to the affected son. The haplotype with the combination of alleles 2/3/2 was transmitted from the unaffected grandmother to the affected son. The haplotype 3/4/1 is now considered to represent the linkage phase that we test in the third generation. This haplotype was transmitted in its complete state from the affected father to his first affected son, individual 5. The unaffected individual 6 inherited the grand maternal combination of alleles at the three loci. Both individuals 5 and 6 represent nonrecombinant gametes. Individual 7 inherited the disease-associated haplotype from markers 1 and 2 but inherited the grand maternal derived allele 2 at marker 3.

Figure 1.19. Pedigree on the left shows unordered genotypes underneath pedigree symbol for three linked markers. Pedigree on the right shows ordered genotypes, with alleles transmitted from father on the left. Alleles included in the disease-associated haplotype are in the box. See text for details.
The combination of markers that individual 7 inherited demonstrates that a recombination event between the disease gene and the marker loci occurred distal to marker 2. The affected daughter, 8, inherited the disease-associated haplotype for markers 2 and 3. She inherited the grand maternal allele from her affected father at marker 1 so that a recombination event occurred between the disease and marker loci proximal to marker 2. These two recombination events suggest that the disease gene in this family is located between markers 1 and 3.

Extensions to Complex Disease

Typically, parametric LOD score analysis (two-point and multipoint) and haplotype analysis are used in Mendelian disease; however, Mendelian transmission of a “complex” trait may be apparent in a subset of families, such as in AD (Post et al., 1997; Pericak-Vance et al., 1991; Schellenberg et al., 1992), ALS (Rosen et al., 1993), and breast cancer (Hall et al., 1992). The study of complex disorders present special, yet not insurmountable, challenges in linkage analysis for gene localization. Elucidation of the genetic defect in a subset of families may provide insights into the pathogenesis of the non-Mendelian form of the disease.

SUMMARY

The study of genes, chromosomes, and patterns of transmission of human traits within families has led to remarkable discoveries that are useful in genetic counseling for recurrence risk, presymptomatic testing, and prenatal diagnosis (see Chapter 4) and in the understanding of the pathogenesis of diseases. The genetic basis of Mendelian disease is relatively straightforward and in many cases is well understood. The situation in common complex disorders is markedly different from the study of Mendelian disease, since more than one gene as well as various nongenetic factors are typically associated with trait phenotype expression. Yet, many of the same principles hold true in complex disease: Mendel’s laws regarding the transmission of genes and alleles at loci are as important to the study of resemblance between relatives in genetically complex disease as in Mendelian disease; the same holds true for the extent and result of the differing types of mutation. Because of the completion of the human genome sequence, the advances will accumulate more rapidly; human genomic study can only expand dramatically.

Linkage analysis is a powerful tool for the primary detection of genes leading to human disease. Methods for localizing genes in diseases that are clearly Mendelian have been highly successful. Extreme care and caution must be exercised when planning a linkage analysis, particularly with complex diseases, as the effects of model misspecification can be severe. Although linkage analysis of more complex diseases, in which the mode of inheritance is unclear, is less straightforward, methods are available to address such situations. The potential rewards of localizing and cloning genes causing the complex and most common diseases are abundant, especially with respect to public health and policy.
REFERENCES


REFERENCES


