Chapter 8

Morgan and Linkage

Thomas Hunt Morgan was a famous geneticist who, in the initial years of the 20th century, studied *Drosophila*, the fruit fly, in his lab at New York City’s Columbia University. Morgan’s choice of *Drosophila* was both fortuitous and prescient not only for his own historic findings but also for developing a model organism that has evolved into a major workhorse in the science of genetics. These tiny flies reproduce quickly and leave a large number of progeny. Mendel had to wait months to plant and harvest two generations of peas. Morgan could study half a dozen generations in the same time. Moreover, *Drosophila* have only three chromosomes.

Two of Morgan’s many findings stand out. Despite all the complicated looping of the DNA around chromosomal proteins, Morgan found that the genes on a chromosome have a remarkable statistical property—namely, mathematically genes appear as if they are linearly arranged along the chromosome. Thus, one can draw a schematic of a chromosome as a single straight line, with the genes in linear order along that straight line, even though the actual physical construction of the chromosome is a series of looped and folded DNA. Morgan’s second finding was no less important. He discovered that chromosomes recombine and

Figure 8.1: Thomas Hunt Morgan.

From http://www.nobelprize.org/nobel_prizes/medicine/laureates/1933/morgan-bio.html
8.1 Recombination

We humans are diploid organisms. That means that we have two copies of every gene, except for those genes on the X and the Y chromosomes. We inherit one gene from our mother and the other from our father. Because the chromosome (and not the gene) is the physical unit of inheritance, it is more appropriate to say that we inherit one maternal chromosome (that happens to have that gene on it) and one paternal chromosome (that also happens to have a sequence of DNA that codes for the same kind of polypeptide as the gene on the maternal chromosome). Those two chromosomes that have the same ordering of genes on them are termed homologous chromosomes and are illustrated in panel (A) of Figure 8.2.

Here the color of the chromosome indicates the parent of origin (blue = father, pink = mother). The loci of interest are denotes by the letters A through F in the upper case versus lower case letters at the gene.

Recombination is a biological phenomenon that effectively “shuffles” parts of homologous chromosomes for transmission to the next generation. The process begins when homologous chromosomes pair up as in panel (A) of Figure 8.2. The chromosomes then join and exchange genetic material (panels B and C). In meiosis, each recombined chromosome goes into a separate gamete (the
two schematic sperm in panel D). As a result, the offspring usually inherits a combination of a parent’s paternal and maternal chromosomes.

The probability that a recombination event occurs between two loci is a function of the distance between the two loci. The alleles at two loci that are far apart on a chromosome are more likely to encounter a recombination event than the alleles for two loci that are close together on that chromosome.

A convenient mnemonic to remember this principal is a dart game. Imagine that the target for a dart game consisted of the two chromosomes in panel (A) of Figure 8.2. The place where a dart hits one of these chromosomes signifies a recombination event at that spot. What is the probability of placing a dart somewhere between the A and the F loci as compared to the probability of hitting somewhere between the A and the B locus? The probability of hitting somewhere between the A and the B locus is much lower than the probability of placing a dart between the A and the F locus. Hence recombination is less likely for two genes located close together on the same chromosome than it is for two genes that are far apart on the same chromosome.

A second mnemonic deals with relationships. It is not unusual for a couple who live apart in different cities to develop separate interests and break apart. At least, this outcome is more probable than it is for couples who reside close to each other. Hence, “close together, stay together, far apart, break apart.” Genes close to each other will tend to stay together. Those far apart will tend to break apart by recombination.

Recombination is not even across the whole genome. In general, there is more recombination near the telomeres of a chromosome than at the centromere (Chowdhury et al., 2009). There are also recombination “hotspots” and “coldspots” where the frequency of recombination is, respectively, increased and decreased.

In most mammals that have been studied, recombination differs as a function of sex. In the generation of a single human egg, females average between 20 and 60 recombinations. Human males, on the other hand, average between 15 and 35 recombination events per sperm (Chowdhury et al., 2009). Although there are various theories about the source of this sex difference, the reason is still not known (Hedrick, 2007).

Finally, the frequency of human recombination as well as its location appears to be a heritable trait (Chowdhury et al., 2009; Fledel-Alon et al., 2011). Recombination occurs more frequently in some families than in others. The reason for this is unknown.

8.1.1 Linkage: DNA in chunks

Bear with me for a few minutes to develop an important concept in genetics—we inherit and pass on large “chunks” of DNA. That is, alleles that are close together on the same chromosome tend to be inherited as a unit (or not inherited at all). If you get lost in the math, do not despair. Skip to the last three paragraphs to get the bottom line.
With 3 billion nucleotides and 23 chromosomes, the number of nucleotides on one strand of the average chromosome is about 140 million. The number of recombinations expected to occur for our mythical average chromosome is around 1 for males and 2 for females. To make life easy, let’s just consider males and fix the average recombination frequency at 1.

Suppose that you are a male and have a dominant allele on one “average” chromosome and a recessive allele on the other. Pretend that nucleotide A at a certain position characterizes the dominant allele while nucleotide C characterizes the recessive one. We want to calculate the probability that a recombination will occur someplace downstream of the A/C polymorphism.

The probability that a recombination event will separate this nucleotide from its adjacent downstream partner is 1 divided by the number of nucleotides on this average chromosome, i.e. 1 divided by 140 million or $7 \times 10^{-9}$. The probability that the recombination event will occur somewhere between our selected nucleotide and a base pair $k$ nucleotides downstream can be accurately approximated as $7k \times 10^{-9}$.

It is easy to think of $k$ in units of 1,000 base pairs, i.e. a kilobase or kb. The probability that recombination will occur within 1 kb downstream of our chosen nucleotide is $7(1000) \times 10^{-9} = 7 \times 10^{-6}$. The probability of a recombination 10 kb downstream of the nucleotide is $7 \times 10^{-5}$; 100 kb downstream is $7 \times 10^{-4}$; and 1,000 kb or a million base pairs (a megabase or Mb), $7 \times 10^{-3}$.

The probability that recombination will not occur between the A/C site and a nucleotide $k$ bases away equals 1 minus the probability that a recombination will occur, i.e., the quantity $1 - (7k \times 10^{-9})$. Some examples will help. The probability that you will pass on that one million base pair section as a “chuck” is over 99%. With some math that need not concern us, the probability that you will pass on a “chuck” that starts 10 million base pairs upstream of the site and ends 10 million base pairs downstream is 0.86.

Let’s return to the topic, namely, alleles that are close together on the same chromosome tend to be inherited as a unit (or not inherited at all). If you transmit the A allele to an offspring, then there is an excellent chance that you will also transmit the “chunk” of DNA several megabases upstream to several megabases downstream of that nucleotide. Of course, if you transmit the A allele, then you will not transmit the C allele or any of the nucleotides that flank it within several millions of base pairs.

Some more math and we are finished. With 3 billion nucleotides and 20,000 protein-coding genes, the average distance between genes is roughly 160,000 nucleotides. Hence, a megabase will contain around six genes. Hence, if you transmit the A allele, you will also transmit all the spelling variations for over a dozen genes that flank that nucleotide.

You should now see how linkage works. If you transmit (or inherit) a certain section of DNA, with a high probability you will also transmit (or inherit) the DNA that surrounds that section. The alleles in this “chunk” are said to be linked.
8.1.1.1 A disclaimer

The calculations above are ballpark estimates. Moreover, they hide the variability in linkage and recombination that occurs throughout the human genome. In some areas, recombination occurs frequently while in others it is rare. Hence, the calculation of the probability of a recombination within a 10 megabase flanking area of a nucleotide will be accurate for some areas but not for others.

Similarly, protein coding genes are not evenly distributed throughout the genome. They are dense in some chromosomal segments. Other chromosomal areas are genetic deserts.

8.2 Haplotypes

In genetics, a haplotype is defined as the ordered alleles on a (sometimes short) segment of the same chromosome. Like many definitions, examples of haplotypes can be more informative than abstract definitions. So examine Figure 8.3 which depicts a short segment of the paternal and maternal chromosomes for two hypothetical individuals, Smith and Jones. Three loci occur in this segment, the A, B, and C loci. Both Smith and Jones have the same genotypes at these loci—\(Aa\) at the A locus, \(Bb\) at the B locus and \(Cc\) at the C locus.

But Smith and Jones have different haplotypes. Smith has haplotype \(AbC\) on his paternal (blue) chromosome and \(aBc\) on his maternal (pink) chromosome. Hence, one would denote Smith’s haplotypes as \(AbC/aBc\). Despite having the same genotypes as Smith, Jones’ haplotypes are \(ABc/abC\). The two have the same genotypes but different haplotypes because the order of the alleles on their chromosomes is different.

8.3 Linkage disequilibrium and haplotype blocks

The terms linkage equilibrium and linkage disequilibrium deal with the ability to predict the alleles in a haplotype. Ask yourself the question “If I know one allele in a haplotype, can I predict the other allele(s) in that haplotype better than chance?” If the answer is “No,” then the alleles are said to be in linkage equilibrium. If the answer is “Yes” (i.e., you can predict better than chance), then the alleles are in linkage disequilibrium. As you might suspect, linkage
disequilibrium can range from weak (prediction is better than chance but is not very accurate) to strong (prediction is very accurate).

A haplotype block is a haplotype in which all of the alleles are in strong disequilibrium. Haplotype blocks characterize the human genome at short spans, say, DNA regions of several kilobases to tens of kilobases. A moment's reflection on the calculations in Section 8.1.1 can tell us why. Alleles arise because of mutation. When a new mutation occurs close to an existing allele, the initial haplotype will be in disequilibrium. After generation upon generation of recombinations between the initial polymorphism and mutation, the two loci into equilibrium. But how long would this be in practical terms?

Let's consider a haplotype of two alleles that are exactly 1 kb apart. The probability that a recombination event will occur between them in the generation of a gamete is roughly 7E-6, so the probability that the two alleles will be transmitted together is 1 - 7E-6 = 0.999993. The probability that this haplotype will be transmitted intact (i.e., not broken up by recombination) across n
Anatomically modern humans emerged about 200,000 years ago. With an average of, say, 20 years per generation we humans have been around for 10,000 generations. Hence, the probability that our haplotype, had it originated with the initial members of our species, would be transmitted intact until the present generation is $0.999993^{10000} = 0.93$. Were the alleles 10 kb apart, that probability becomes 0.49, imperceptibly different from the flip of a coin. Hence, at short genomic distances, we humans today have haplotype blocks in which the alleles are in strong disequilibrium.

Figure 8.4, taken from Walton et al. (2005), illustrates the haplotype blocks in a genomic region that codes for some enzymes responsible for oxidative metabolism. Most of the information in this figure is of a highly technical nature, so let us concentrate on the highlights. The column immediately to the left of the part with the red triangle lists the 66 polymorphisms in this area in linear order starting at the top. Every red triangle to the right of this list that has a black arrow pointing to it indicates a major haplotype block. Hence, the first five polymorphisms are in such strong disequilibrium that if one know just one allele of these five, one can predict the other four with a great degree of accuracy. Loci numbers 6 through 12 form the second block, and so on.

Considerable research has gone into the identification of haplotype blocks in the human genome. Why? We want to detect which of the many millions of human polymorphisms are associated with a medical condition. In the past, it was not feasible to genotype people on all of the polymorphisms. One could, however, optimize genotyping information by identifying haplotype blocks and then genotyping only one locus per block. Reconsider Figure 8.4. There are 66 polymorphisms in the region. One could genotype only one or two polymorphisms within the five strong haplotype blocks and greatly reduce the need for genotyping this region.

In 2003, several genetics groups throughout the world initiated a haplotype mapping project that became known as HapMap (The International HapMap Consortium, 2003). Within several years and considerable lab work, they several millions of polymorphisms in linkage disequilibrium scattered throughout the human genome (International HapMap Consortium, 2007). Biotech firms followed up and developed efficient genotyping “chips” based on these results. We will examine the results of this technology later in this book (Sections X.X).

8.4 Measures of genetic distance (graduate)

There are several measures the quantify the distance between two linked loci. The first, and most straightforward, is simply the number of nucleotides that separate the loci. Usually, the units here are expressed in base pairs (or bp), thousands of base pairs (kilobases or kb), or millions of base pairs (megabases or Mb). To account for insertions and deletions, the number of nucleotides separating loci is based on the consensus human genome sequence.

The second unit is the recombination fraction, usually denoted by the greek
lower case theta ($\theta$). Often mistakenly defined as the probability of a recombination, $\theta$ is actually a *conditional probability* for two loci that equals the probability that a gamete will contain an allele from the *opposite* chromosome given that it contains an allele from the original chromosome of interest.

The third unit is the *centimorgan* or $cM$. One centimorgan is defined as the physical distance corresponding to a value of $\theta$ equalling 0.01. In other words, it is the distance such that the probability is .01 that a gamete will contain an allele from the first locus on one chromosomal strand but an allele at the second locus from the *opposite* chromosomal strand.

When the loci are close together, then $\theta \approx cM$. This holds for values of $\theta$ between 0 and about 0.10. As the distance increases, however, one must account for the probability that more than one cross over may occur between the loci. The famous geneticist, Haldane (1919)\(^1\) developed a mapping function that related the recombination fraction to centimorgans

$$\theta = \frac{1 + \exp(-2cM/100)}{2}$$

(8.1)

or conversely,

$$cM = 50 \left( \frac{1}{1 - 2\theta} \right)$$

(8.2)

It is crucial to realize that a centimorgan does not refer to a constant number of base pairs throughout the whole human genome. Recall that recombination does not occur uniformly throughout the genome (Petes, 2001). Consequently, one cM will equal more base pairs in one region that it does in other regions.

Finally, the terms developed here are mostly—but not entirely—of historical relevance and can aid the student in reading the literature. Today, most distance measures are expressed in terms of base pairs. Both the recombination fraction and centimorgans were used when the era of gene mapping (i.e., finding the exact chromosomal location of DNA regions) was in full swing.

### 8.5 Calculating gametes and genotypes under linkage (graduate)

To calculate gametes under linkage, first review Section X.X on the Punnett rectangle. Linkage involves a similar setup. It will only differ in applying the *recombination fraction* instead of the Mendelian probability of 0.5 to the elements of the rectangle.

An example can help to illustrate the situation. Assume two linked loci, the first with alleles $A$ and $a$ and the second with alleles $B$ and $b$ and consider the gametes that may be be generated from a person with the genotype depicted in Figure X.X. (HINT: when you are learning about linkage, it is helpful to color-code the chromosomes according to parental origin. In the figure, we use a traditional

\(^1\)There are a number of other mapping functions; see Zhao and Speed (1996) for details.
Table 8.1: Gametes under linkage

<table>
<thead>
<tr>
<th>First Locus</th>
<th>prob</th>
<th>Second Locus</th>
<th>prob</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.5</td>
<td>b</td>
<td>0.5(1-\theta)</td>
</tr>
<tr>
<td>a</td>
<td>0.5</td>
<td>B</td>
<td>0.5\theta</td>
</tr>
</tbody>
</table>

The recombination fraction is a statistical measure of the distance between two loci. Technically, it equals the conditional probability that, given a gamete that contains \(A\) from the paternal chromosome, the gamete will also contain \(B\) from the maternal chromosome, i.e., \(\text{prob} (\text{gamete has } B | \text{gamete has } A)\). It is assumed that the probability is symmetric in the sense that it will also equal the conditional probability that, given a gamete with \(a\), the gamete will also contain \(b\). The value of \(\theta\) will range from 0 (the two loci are so close together that the linked loci are always transmitted together) to 0.5 (the upper limit of the probability under Mendel’s law of segregation). A value of \(\theta = 0.5\) implies that the loci are very far apart on the same chromosome or are on different chromosomes.

Table 8.1 gives the template for calculating the probability of gametes under linkage. Label the rows with the “given” locus and set the probability to 0.5 because the probability of transmitting either of the two alleles follows Mendel’s law of segregation. Then add two columns, the first for the allele at the second locus and the second for the other allele at the second locus. Once again, color coding can diminish the amount of confusion here.

Finally, if the first and second allele are on the same chromosome (i.e., they have the same colors), then the probability of transmitting gamete with these alleles will be \(0.5(1-\theta)\). If the first and second allele are on opposite chromosomes (i.e., different colors), then the probability of that gamete equals \(0.5\theta\). To arrive at an actual number, one must have a numerical value for \(\theta\). If \(\theta = 0.12\), then the probability that a gamete contains \(AB\) equals the probability that the gamete has \(aB\) and will equal \(0.5(1-.12) = 0.44\). The probability that the gamete is \(AB\) equals the probability that it is \(ab\) and is \(0.5(.12) = 0.06\).
8.5. CALCULATING GAMETES AND GENOTYPES UNDER LINKAGE (GRADUATE)

To say the same things in different terms, the probability that a haplotype involving two loci will be transmitted intact is the sum of the probabilities for intact (i.e., non recombinant) gametes in Table 8.1 or \((1 - \theta)\). The probability that a haplotype involving two loci will be broken up by recombination is the sum of the probabilities in that table for recombinant gametes or \(\theta\). This perspective gives another definition for the recombination fraction, \(\theta\), albeit one that is mathematically equivalent to its definition as a conditional probability–\(\theta\) is the probability of a recombinant haplotype involving two loci.

8.5.1 Genotypes (graduate)

The probabilities for the genotypes of offspring under linkage is a straightforward application of the rules of the Punnett rectangle outlined in Section X.X but substituting the gametic probabilities under linkage for those under Mendel’s law of independent assortment. Hence, we will have a contingency table of the, say, gametes from the female parent and their probabilities as the rows and the gametes and probabilities of the other—in this case, male—parent as the columns.

As an example, consider a mother with haplotypes \(AB\) and \(ab\) and her mate with haplotypes \(aB\) and \(Ab\). The generic contingency table for the genotypes of this mating are given in Table 8.2 where \(\theta_{fe}\) is the probability of a recombination between these loci for a female-generated gamete and \(\theta_{ma}\) for a male gamete.

The very first step in constructing this table is to list the color-coded gametes that can be generated by mother. These label the rows of the table. Then do the same for the father’s gametes which will determine the columns of the table. Next use the rules outlined above to calculate the probability of mother’s gametes (see Section 8.5). These are the algebraic quantities alongside the alleles from the maternal gametes listed in Table 8.2. The subsequent step consists of calculating the probability of the male gametes and listing them below the alleles contained in his gametes. Finally, multiply the row probability and the column probability to determine the probability of the genotype of the offspring.

For example, assume that the recombination frequency for females between these loci is \(\theta_{fe} = .04\) and the recombination frequency for males is \(\theta_{ma} = 0.11\). Then the probability of mother’s gamete \(AB\) equals \(0.5(1 - .04) = 0.48\). Similarly, the probability of father’s gamete containing \(AB\) is \(.5*.11 = 0.055\). Then the probability of an offspring from these two gametes is their product or \(0.48*0.055 = 0.0264\).

The final step is to collect all similar genotypes for the offspring and then add their probabilities together. These are listed in Table 8.3. For example, consider an offspring with genotype \(AABb\). There are two ways to observe such a genotype. The first occurs when mother transmits \(AB\) and father transmits \(Ab\). The probability of this event equal \(0.48*0.445 = 0.2136\). The second way to observe this genotype in the offspring happens when mother transmits \(XX\) and father transmits \(XX\). The probability of this event equals \(0.02*0.055 = 0.0011\). Hence, the probability of observing genotype \(AABb\) in the offspring is \(0.2136 + 0.0011 = 0.2147\).
### Table 8.2: Offspring genotypes for two linked loci.

<table>
<thead>
<tr>
<th>Maternal gametes and probabilities:</th>
<th>$AB$</th>
<th>$ab$</th>
<th>$AB$</th>
<th>$Ab$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$AB$ $.5(1 - \theta_{fe})$</td>
<td>.25(1 - \theta_{fe})(1 - \theta_{ma})</td>
<td>.25(1 - \theta_{fe})\theta_{ma}</td>
<td>.25(1 - \theta_{fe})\theta_{ma}</td>
<td>.25(1 - \theta_{fe})(1 - \theta_{ma})</td>
</tr>
<tr>
<td>$Ab$ $.5\theta_{fe}$</td>
<td>.25\theta_{fe}(1 - \theta_{ma})</td>
<td>.25\theta_{fe}\theta_{ma}</td>
<td>.25\theta_{fe}\theta_{ma}</td>
<td>.25\theta_{fe}(1 - \theta_{ma})</td>
</tr>
<tr>
<td>$aB$ $.5\theta_{fe}$</td>
<td>.25\theta_{fe}(1 - \theta_{ma})</td>
<td>.25\theta_{fe}\theta_{ma}</td>
<td>.25\theta_{fe}\theta_{ma}</td>
<td>.25\theta_{fe}(1 - \theta_{ma})</td>
</tr>
<tr>
<td>$ab$.5(1 - \theta_{fe})</td>
<td>.25(1 - \theta_{fe})(1 - \theta_{ma})</td>
<td>.25(1 - \theta_{fe})\theta_{ma}</td>
<td>.25(1 - \theta_{fe})\theta_{ma}</td>
<td>.25(1 - \theta_{fe})(1 - \theta_{ma})</td>
</tr>
</tbody>
</table>
8.6. **Statistics for Linkage Equilibrium (Graduate)**

There are several statistics used to quantify linkage disequilibrium (Devlin and Risch, 1995; Neale, 2010), but two are of prime importance. The data are first organized in a two by two table, a generic form of which is given in Table 8.4. Here it is assumed that the first allele in the haplotype can be either \( A \) or \( G \) and at the the second locus, \( C \) or \( T \) (after the nucleotides). The notation in Table 8.4 is in terms of proportions. Hence, \( p_1 \) is the frequency of allele \( A \), \( q_1 \), the frequency of \( G \), etc.

### Table 8.3: Offspring genotypes from a mating involved linked loci.

<table>
<thead>
<tr>
<th>Offspring Genotype</th>
<th>Matern. Gamete</th>
<th>Patern. Gamete</th>
<th>Probability</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>AABB</td>
<td>( AB )</td>
<td>( AB )</td>
<td>( .25(1 - \theta_{fe})\theta_{ma} )</td>
<td>0.0264</td>
</tr>
<tr>
<td>AABB</td>
<td>( AB )</td>
<td>( Ab )</td>
<td>( .25(1 - \theta_{fe})(1 - \theta_{ma}) )</td>
<td>0.2147</td>
</tr>
<tr>
<td>AAbb</td>
<td>( Ab )</td>
<td>( Ab )</td>
<td>( .25\theta_{fe}\theta_{ma} )</td>
<td>0.0089</td>
</tr>
<tr>
<td>AaBB</td>
<td>( AB )</td>
<td>( aB )</td>
<td>( .25(1 - \theta_{fe})\theta_{ma} )</td>
<td>0.2147</td>
</tr>
<tr>
<td>AaBb</td>
<td>( Ab )</td>
<td>( aB )</td>
<td>( .25\theta_{fe}\theta_{ma} )</td>
<td>0.0706</td>
</tr>
<tr>
<td>Aabb</td>
<td>( Ab )</td>
<td>( ab )</td>
<td>( .25\theta_{fe}\theta_{ma} )</td>
<td>0.2147</td>
</tr>
<tr>
<td>aaBB</td>
<td>( aB )</td>
<td>( aB )</td>
<td>( .25\theta_{fe}\theta_{ma} )</td>
<td>0.0089</td>
</tr>
<tr>
<td>aaBb</td>
<td>( ab )</td>
<td>( ab )</td>
<td>( .25\theta_{fe}\theta_{ma} )</td>
<td>0.2147</td>
</tr>
<tr>
<td>aabb</td>
<td>( ab )</td>
<td>( ab )</td>
<td>( .25(1 - \theta_{fe})(1 - \theta_{ma}) )</td>
<td>0.0264</td>
</tr>
</tbody>
</table>

### Table 8.4: Schematic two by two table for analyzing linkage disequilibrium.

<table>
<thead>
<tr>
<th>Second Locus:</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>First Locus:</td>
<td></td>
</tr>
<tr>
<td>( A )</td>
<td>( p_1 )</td>
</tr>
<tr>
<td>( X_{11} )</td>
<td></td>
</tr>
<tr>
<td>( X_{12} )</td>
<td></td>
</tr>
<tr>
<td>( G )</td>
<td>( q_1 )</td>
</tr>
<tr>
<td>( X_{21} )</td>
<td></td>
</tr>
<tr>
<td>( X_{22} )</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>( p_2 )</td>
</tr>
<tr>
<td>( q_2 )</td>
<td></td>
</tr>
</tbody>
</table>

The first measure of association is \( D' \), called “D prime” Lewontin (1988). Consider the \( AC \) cell. Under linkage equilibrium, the expected frequency of this cell is the product of the two base rates or \( p_1p_2 \). Hence, a measure of discrepancy from equilibrium can be calculated as the observed rate less its
expected frequency under chance, or

\[ D = X_{11} - p_1 p_2 \] (8.3)

The value of \( D \) is sensitive to the allele frequencies. For example, suppose that \( p_1 = p_2 = 0.5 \). Under complete disequilibrium, \( X_{11} \) takes on its maximum value of 0.5, giving \( D = 0.5 - 0.5 \times 0.5 = 0.25 \). If, on the other hand, \( p_1 = p_2 = 0.2 \), then under complete disequilibrium, \( D = 0.2 - 0.2 \times 0.2 = 0.16 \). The statistic \( D' \) overcomes this limitation by dividing \( D \) by its maximum value. The formula is

\[ D' = \begin{cases} D / \min(p_1 q_2, q_1 p_2) & D > 0 \\ D / \min(p_1 p_2, q_1 q_2) & D < 0 \end{cases} \] (8.4)

A second measure of association is the correlation. The quantity \( D \) in Equation 8.3 is also the covariance between the two alleles. The correlation divides the covariance by the product of the standard deviations of the two variables,

\[ R = \frac{D}{\sqrt{p_1 q_1 p_2 q_2}} \] (8.5)

In statistics, the quantity \( R \) in Equation 8.5 is called the phi coefficient (\( \phi \)). Many geneticists report the square of the correlation, \( R^2 \), because it removes the sign but still measures the magnitude of the association.

Both \( D' \) and \( R \) have advantages and disadvantages. The correlation is sensitive to differences in allelic frequencies between the two loci. For example, suppose that the frequency of \( \text{A} \) in Table 8.4 is 0.1 and the frequency of \( \text{C} \) is 0.5. The maximum value for \( R \) in this case is 0.33.

\( D' \), on the other hand, can give values of 1.0 when there is a lack of predictability. Once again, consider the case when \( p_1 = 0.1 \) and \( p_2 = 0.5 \). The maximum amount of disequilibrium occurs when the frequency of \( \text{AC} \) is 0.1 and \( \text{AT} \) is not observed. Here, \( D' \) equals 1.0. Hence, we know that if a person as two \( \text{A}s \), then that person must also have two \( \text{C}s \). But if the person is \( \text{AG} \) at the first locus, can I perfectly predict the genotype at the second locus? The answer is no. That person must have at least one \( \text{C} \) but one cannot predict perfectly whether the person has a \( \text{C} \) or a \( \text{T} \) at the second locus.

### 8.7 Factors influencing disequilibrium (graduate)

Most of the factors that influence linkage disequilibrium are evolutionary forces that will be discussed later in Chapter X.X. It is obvious that natural selection can influence disequilibrium when some haplotypes have greater reproductive fitness than others. Random genetic drift (i.e., change in haplotype frequencies due to chance and chance alone) will also influence disequilibrium but only in small populations.

Population structure (i.e., all those factors that influence who mates with whom in a population) will also influence linkage disequilibrium, but there are
several different ways in which this can occur. First, consider a population that is subdivided into several smaller populations that have a tendency to mate within themselves. This is a phenomenon called population stratification. Differences in haplotype frequencies among the subpopulations will contribute to disequilibrium. Similarly, when stratification breaks down and mating becomes random, then the approach to equilibrium will be accelerated.

A second factor in population structure, most applicable to human populations, is assortative mating or a mating system that generates a correlation among the phenotypes of mates. (Think height. Tall people tend to marry tall people and short people, short people). The effect of assortment on disequilibrium will usually be small.

Finally, disequilibrium begins with mutation. Consider a population with several haplotypes at a region of the genome. When a novel mutation occurs, it must happen in one of those haplotypes, immediately inducing disequilibrium. The statistical effect, however, of the new mutant will be negligible when the population is large. If the haplotype with the mutation increases, then the statistical effect of disequilibrium may become considerable.

The one non evolutionary force influencing equilibrium is simply time. At each generation, linkage disequilibrium will have a tendency to become smaller when it is not counteracted by one of the evolutionary forces.

8.7.1 Approach to equilibrium (graduate)

The statistics outlined above in Section 8.6 can be used to view the approach to equilibrium. Here, it is assumes that the population is large, mating at random and not undergoing selection for the loci in question. We will also ignore mutation.

Consider haplotype AC from Table 8.4. Its frequency in the population is \( X_{11} \) and the frequency of the A allele is \( p_1 \) while that of the C allele is \( p_2 \). Let \( \theta \) denote the recombination fraction between the two loci and \( X^*_{11} \) the frequency of this haplotype in the next generation. With some math it can be shown that

\[
X^*_{11} = X_{11}(1 - \theta) + p_1 p_2 \theta
\]  

(8.6)

The two terms in this equation apply to two phenomenon. The first, \( X_{11}(1 - \theta) \), equals the frequency of haplotype AC in the initial generation times the probability that the haplotype does not recombine and “pick up” a different allele. The second term, \( p_1 p_2 \theta \), is the probability that by change a recombination event occurs and generates the AC haplotype.

Subtract the quantity \( p_1 p_2 \) from both sides of Equation 8.6 giving

\[
X_{11} - p_1 p_2 = X_{11}(1 - \theta) + p_1 p_2 \theta - p_1 p_2
\]  

\[
= (X_{11} - p_1 p_2)(1 - \theta)
\]  

(8.7)

Note that the assumptions of no selection, no genetic drift and random mating require that both \( p_1 \) and \( p_2 \) remain unchanged over time.
Recall from Equation 8.3 that, by definition, \( D = X_{11} - p_1 p_2 \). Hence, Equation 8.7 can be written as

\[
D^* = (1 - \theta)D
\]  

(8.8)

If \( D_0 \) is the value of \( D \) at generation 0, then the value of \( D \) at generation \( n \) is

\[
D_n = (1 - \theta)^n D_0
\]  

(8.9)

As \( n \) grows large the quantity \((1 - \theta)^n\) approaches 0, giving the equilibrium condition. When \( \theta \) is not close to 0, then the approach to equilibrium can be quite rapid as illustrated in Figure

8.8 References


