# Linkage Disequilibrium

2

Montgomery Slatkin

**Abstract**

Linkage disequilibrium (LD) is the nonrandom association between alleles at closely linked loci. LD is created by genetic drift and natural selection, and it decays exponentially with time at a rate proportional to the recombination rate. This chapter reviews the theory of LD between pairs of loci and the use of LD for detecting past episodes of selection and for gene mapping.

## 2.1 Introduction

Although population genetics largely focuses on one locus at a time, much is to be learned from considering two or more loci together. The reason is that alleles at different loci are transmitted together, creating the opportunity for correlations that reflect their common history. This correlation is important for gene mapping, where the goal is to identify loci that affect a trait, and when considering the effects of natural selection. Loci that are selected influence nearby neutral loci. Therefore, the study of sets of loci together can provide more insight into evolutionary processes and give additional information about gene action than can be obtained by focusing on each locus separately.

For simplicity, I will start by presenting results for two loci. Assume that the two loci are on the same pair of homologous chromosomes. If there are two alleles at each of the two loci (*A/a* and *B/b*), there are four combinations, called haplotypes, on a chromosome, *AB*, *Ab*, *aB*, and *ab*. These haplotypes can be thought of as the four kinds of gametes that can be produced by an individual.

M. Slatkin (✉)
Department of Integrative Biology, University of California, Berkeley, CA, USA
e-mail: slatkin@berkeley.edu



K. E. Lohmueller, R. Nielsen (eds.), *Human Population Genomics*,
https://doi.org/10.1007/978-3-030-61646-5_2

A population is characterized by the frequencies of the four haplotypes, $f_{AB}$, $f_{Ab}$, $f_{aB}$, and $f_{ab}$. Allele frequencies can be recovered from the haplotype frequencies:$f_A = f_{AB} + f_{Ab}$, etc. Given the haplotype frequencies, we can determine whether an allele's presence in a haplotype is independent of the allele present at the other locus. If they are independent, then the haplotype frequency is the product of the allele frequencies. For example, $f_{AB} = f_A f_B$. If that is the case, the two loci are said to be in linkage equilibrium. If not, they are in linkage disequilibrium, often abbreviated LD. If they are in LD, the extent of LD is quantified by the difference between the actual haplotype frequency and the frequency expected at linkage equilibrium:

$$D = f_{AB} - f_A f_B. \tag{2.1}$$

This calculation is illustrated in Fig. 2.1 in a sample of eight chromosomes. $f_A = 1/2, f_B = 5/8$, and $f_{AB} = 3/8$, which gives $D = 1/16$.

The quantity $D$ is called the coefficient of linkage disequilibrium. $D = 0$ implies there is linkage equilibrium. $D$ can be regarded as a covariance in the allelic state at the two loci.

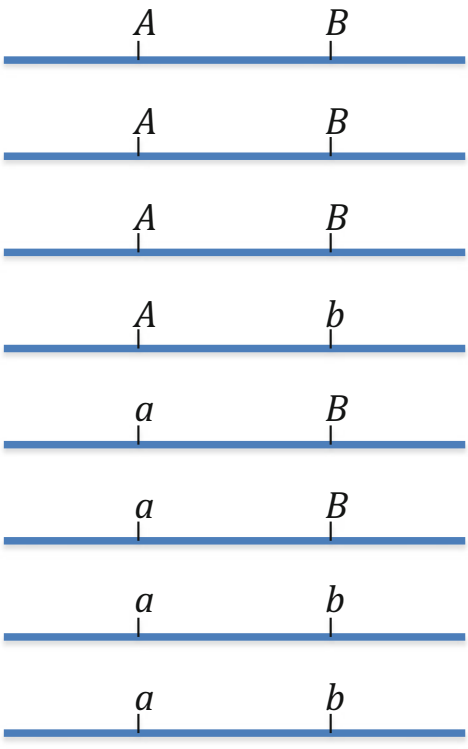


**Fig. 2.1** Illustration of haplotype counts in a hypothetical sample of 8 chromosomes

Although there are four haplotypes, there is only a single $D$ needed to describe the extent of LD. That is clear when the relationship between haplotype and allele frequencies is used. For example, if $D$ is defined by Eq. (2.1), then

$$f_{Ab} - f_A f_b = f_A - f_{AB} - f_A f_b = f_A (1 - f_b) - f_{AB} = f_A f_B - f_{AB} = -D.$$

It is possible, then, to express all the haplotype frequencies in terms of the allele frequencies and $D$:

$$\begin{aligned} f_{AB} &= f_A f_B + D \\ f_{Ab} &= f_A f_b - D \\ f_{aB} &= f_a f_B - D \\ f_{ab} &= f_a f_b + D. \end{aligned} \tag{2.2}$$

Notice that, if there are more $AB$ haplotypes than expected at linkage equilibrium ($D > 0$), there have to be more $ab$ haplotypes and fewer $Ab$ and $aB$ haplotypes.

Although $D$ is defined by Eq. (2.1), there is an equivalent but different expression:

$$\begin{aligned} D &= f_{AB} - f_A f_B = f_{AB} - (f_{AB} + f_{Ab})(f_{AB} + f_{aB}) \\ &= f_{AB}(1 - f_{AB} - f_{Ab} - f_{aB}) - f_{Ab} f_{aB} = f_{AB} f_{ab} - f_{aB} f_{Ab} \end{aligned} \tag{2.3}$$

Equation (2.2) tells us that, because none of the haplotype frequencies can be negative, there is a limit on the magnitude of $D$ imposed by the allele frequencies. If $D > 0$, $D$ must be no greater than the smaller of $f_A f_b$ and $f_a f_B$, and if $D < 0$, $D$ must be larger than $-f_A f_B$ and $-f_a f_b$. That is,

$$-\min(f_a f_b, f_A f_B) \le D \le \min(f_A f_b, f_a f_B). \tag{2.4}$$

One question that arises when computing $D$ for different pairs of loci is whether a particular value is large or small. For example, does $D = 0.006$ indicate a small or large amount of LD for a pair of loci? Inequality (2.4) tells us that the answer depends on the allele frequencies and suggests that it is useful to express $D$ relative to its maximum or minimum possible value. We can do this by defining

$$\begin{aligned} D' &= \frac{D}{\min(f_A f_b, f_a f_B)} \text{ if } D > 0 \\ &= \frac{D}{-\min(f_a f_b, f_A f_B)} \text{ if } D < 0 \end{aligned} \tag{2.5}$$

which indicates how close $D$ is to its maximum or minimum value (Lewontin 1964). In the example with $D = 0.006$, if $f_A = 0.4$ and $f_B = 0.01$, then $D' = 1$, while if $f_A = 0.4$ and $f_B = 0.3$, then $D' = 0.0333$.

## 2.2 Tests of whether $D = 0$

The quantity $D'$ is, as will be seen below, useful for some purposes, but its value alone does not tell us whether there is statistically significant LD between a pair of loci, that is, whether the hypothesis that $D = 0$ can be rejected. Two tests are commonly used. One is the standard $\chi^2$ test of significance for a 2×2 contingency table whose entries are the numbers of each of the four haplotypes. Closely related to the $\chi^2$ test is the $r^2$ statistic,

$$r^2 = \frac{D^2}{f_A f_a f_B f_b} \tag{2.6}$$

which provides another way to quantify the extent of LD. $r^2$ is formally a correlation coefficient. Although the upper bound of $r^2$ is 1, it does in general not take the value 1 even when $D' = 1$. It is convenient to use $r^2$ because, when testing for significance in the contingency table, $\chi^2 = nr^2$ where $n$ is the number of haplotypes sampled. This result follows from the fact that $D$ is the difference between the observed haplotype frequency and the haplotype frequency expected if alleles at the two loci are randomly associated. Such a difference arises naturally when doing a $\chi^2$ test of statistical significance in a 2×2 contingency table. When sample sizes are large and neither allele is rare, the $\chi^2$ test is powerful and easy to use. When sample sizes are small or at least one of the haplotypes is rare (<5 in count), then Fisher's exact test is preferable (Weir 1996).

For most practical purposes, $D'$ and $r^2$ are equally useful, and in real data sets, their values are highly correlated. One important feature of $D'$ is that when its value is 1, at least one of the four haplotype frequencies is 0. That situation is particularly important because, when a new mutation arises at a previously monomorphic locus, $D' = 1$ with all polymorphic loci on the same chromosome. The single copy of the new mutant arises on only one genetic background. After this time, $D'$ between the mutant and another locus becomes less than 1 only if there is recombination between the two loci.

## 2.3 More than Two Alleles per Locus

Describing LD for a pair of biallelic loci is relatively simple because a single quantity $D$, combined with the allele frequencies, provides a complete characterization. If there are more than two alleles at either or both loci, more coefficients of LD are needed, one for the difference between the frequency of each haplotype and the frequency expected under random association. If the alleles at one locus are $A_1$, $A_2$, $A_3$ ... and those at the other are $B_1$, $B_2$, $B_3$ ..., then

$$D_{ij} = f_{A_i B_j} - f_{A_i} f_{B_j}. \tag{2.7}$$

The $D_{ij}$ are not independent because of the requirement that allele frequencies at each locus sum to 1. If there are $n_1$ alleles at the first locus and $n_2$ at the second, there are $(n_1 - 1)(n_2 - 1)$ independent values of the $D_{ij}$.

The theory of LD for multiple alleles has been important, particularly in the application to the major histocompatibility complex (MHC) region in humans and other vertebrates. MHC loci often have dozens and even hundreds of alleles, and there is abundant LD among most of the loci (Hedrick et al. 1986). Furthermore, there is evidence of balancing selection which could contribute to the extent of LD. For genomic data, nearly all single-nucleotide polymorphisms (SNPs) and insertion/deletion polymorphisms are biallelic, so a single value of $D$ is sufficient.

## 2.4 More than Two Loci: Haplotype Blocks and the HapMap Project

When more than two polymorphic loci are analyzed together, as is always the case with genomic data, the analysis can become quite complicated. There are analogs of $D$ defined for three or more loci that represent higher-order linkage disequilibria. For example, with three diallelic loci (*A/a, B/b*, *C/c*), the third-order disequilibrium coefficient is defined to be

$$D_{ABC} = f_{ABC} + f_A D_{BC} + f_B D_{AC} + f_C D_{AB} - f_A f_B f_C \tag{2.8}$$

where $D_{AB}$, $D_{AC}$, and $D_{BC}$ are the pairwise coefficients of LD defined above (Geiringer 1944). The higher-order coefficients of LD are analogous to higher-order interaction terms in the analysis of contingency tables with more than two dimensions.

These higher-order coefficients are well defined, and their theoretical properties have been studied extensively, but they are difficult to estimate and interpret. Furthermore, there are many of them because the number of higher-order coefficients grows as an exponential function of the number of loci. Higher-order coefficients of LD have been used primarily in the study of the human MHC loci (Robinson et al. 1991) because there are strong multilocus patterns of LD that are of clinical significance.

The approach much more commonly taken to analyzing numerous polymorphic loci is to compute $D'$ and $r^2$ for all pairs of polymorphic sites. In the human genome, it is often found that relatively large values of $D'$ and $r^2$ are found between sets of closely linked sites (Daly et al. 2001). Sets of closely linked sites in strong LD are called haplotype blocks, and they have played an important role in human genetics in the past several years. The discovery that much of the human genome is made up of haplotype blocks was part of the impetus for the HapMap project, which had the goal, now achieved, of finding nearly all common SNPs in several human populations (Consortium 2003, 2005, 2007; HapMap3 2010). The emphasis on common SNPs, i.e., those with allele frequencies in the range (0.05, 0.95) was motivated by the hypothesis that alleles in this frequency range are

largely responsible for the genetic basis of complex inherited diseases. For complex diseases, such as most cancers, most forms of heart disease, and many psychiatric disorders and autoimmune diseases, disease risk in close relatives of an affected individual is higher than the average risk in a population, which strongly suggests there is a genetic basis, yet the genetic basis is not attributable to single Mendelian loci.

## 2.5 Dynamics of *D*

The term "linkage disequilibrium" is unfortunate for two reasons. First, it does not necessarily tell us something about the linkage of two loci. Two loci on different chromosomes might be in linkage disequilibrium, while two closely linked loci might be in linkage equilibrium. Second, the term implies that it describes a dynamic process, but it does not. Instead, $D$, $D'$, and $r^2$ quantify the relationship between haplotype and allele frequencies in a population at a given time. We can understand the dynamics of LD by determining how $D$ changes under the influence of various forces, including random mating, natural selection, recombination, and genetic drift.

We will begin with random mating and recombination. We assume that zygotes are formed by randomly combining haplotypes and that the two loci have a recombination rate $c$ between them. We also assume the haplotype frequencies in generation $t$ are $f_{AB}$, $f_{Ab}$, $f_{aB}$, and $f_{ab}$. We can compute the haplotype frequencies in the next generation $(t + 1)$ by assuming gametes are randomly combined into genotypes. Table 2.1 shows the genotypes, the genotype frequencies, and the frequencies of gametes produced by each genotype. It will be necessary to distinguish haplotypes of the two parental gametes of doubly heterozygous individuals, meaning those with genotype *AaBb*, because the gametes produced depend on whether they are doubly heterozygous because their parental gametes are *AB* and *ab* or *Ab* and *aB*.

**Table 2.1** Two-locus genotypes and their frequencies in a randomly mating population, along with the frequencies of gametes produced by each genotype

| | | Gametes produced | | | |
|---|---|---|---|---|---|
| Genotype | Frequency | *AB* | *Ab* | *aB* | *ab* |
| *AB/AB* | $f_{AB}^2$ | 1 | 0 | 0 | 0 |
| *AB/Ab* | $2f_{AB}f_{Ab}$ | 1/2 | 1/2 | 0 | 0 |
| *AB/aB* | $2f_{AB}f_{aB}$ | 1/2 | 0 | 1/2 | 0 |
| *AB/ab* | $2f_{AB}f_{ab}$ | $(1–c)/2$ | $c/2$ | $c/2$ | $(1–c)/2$ |
| *Ab/ab* | $f_{Ab}^2$ | 0 | 1 | 0 | 0 |
| *Ab/aB* | $2f_{Ab}f_{aB}$ | $c/2$ | $(1–c)/2$ | $(1–c)/2$ | $c/2$ |
| *Ab/ab* | $2f_{Ab}f_{ab}$ | 0 | 1/2 | 0 | 1/2 |
| *aB/aB* | $f_{aB}^2$ | 0 | 0 | 1 | 0 |
| *aB/ab* | $2f_{aB}f_{ab}$ | 0 | 0 | 1/2 | 1/2 |
| *Ab/ab* | $f_{ab}^2$ | 0 | 0 | 0 | 1 |

From Table 2.1, it is straightforward to compute the haplotype frequencies in the next generation. For example,

$$\begin{aligned} f_{AB}(t+1) &= f_{AB}^2 + 2f_{AB}f_{Ab}(1/2) + 2f_{AB}f_{aB}(1/2) \\ &\quad +2f_{AB}f_{ab}(1-c)/2 + 2f_{Ab}f_{aB}(c/2) \\ &= f_{AB}^2 + f_{AB}f_{Ab} + f_{AB}f_{aB} + f_{AB}f_{ab} - c(f_{AB}f_{ab} - f_{Ab}f_{aB}) \\ &= f_{AB} - c(f_{AB}f_{ab} - f_{Ab}f_{aB}) \\ &= f_{AB} - cD. \end{aligned} \tag{2.9}$$

where the last step uses Eq. (2.3).

We conclude that under random mating, haplotype frequencies change each generation unless $D = 0$. That is in contrast to what happens at each locus separately. The Hardy-Weinberg law tells us that random mating does not change allele frequencies. It is also important that the extent of change in the haplotype frequencies depends on the recombination rate between the two loci. We can find the recursion equation for $D$ alone by using the fact that $f_{AB}(t+1) = f_A(t+1)f_B(t+1) + D(t+1)$, $f_A(t+1) = f_A$, and $f_B(t+1) = f_B$ to obtain

$$f_A f_B + D(t+1) = f_A f_B + D(t) - cD(t) \tag{2.10}$$

or

$$D(t+1) = (1-c)D(t). \tag{2.11}$$

Equation (2.11) tells us that $D$ decreases by a factor of (1–$c$) after one generation of random mating in a very large population. Because the decrease is by the same factor each generation, $D$ decreases exponentially with time from its initial value:

$$D(t) = (1-c)^t D(0). \tag{2.12}$$

The rate of decrease is determined by the recombination rate between the two loci. If $c$ is small, then $(1-c)^t \approx e^{-ct}$, and we conclude that $D$ will decrease to roughly 37% of its initial value after $1/c$ generations of random mating.

It is important that $D$ does not go to zero after one generation of random mating. Even between unlinked loci, for which $c = 1/2$, $D$ decreases only by a factor of 1/2 each generation. That behavior is in marked contrast to what happens to genotype frequencies after one generation of random mating. The foundation of population genetics is that the Hardy–Weinberg (HW) genotype frequencies are established in one generation of random mating regardless of the initial genotype frequencies. That $D$ between unlinked loci does not go to zero in a single generation of random mating is surprising because both HW genotype frequencies and linkage equilibrium indicate statistical independence. At the HW frequencies, the presence of an allele on one homologue is independent of the presence of an allele on the other. At linkage equilibrium, the presence of an allele at one locus in a haplotype

is independent of the presence of an allele at the other locus. Yet the above results show that independence between homologues is established in one generation, but the independence of loci on different chromosomes is established more slowly.

These results can be related to human populations in a way that illustrates the reason that linkage disequilibrium has become such an important part of human population genetics. There are approximately 24,000 coding genes in the human genome, which has a total recombination length of 30 Morgans or 3000 cM (Lander 2001). Therefore, the average recombination distance between adjacent coding genes is 1/8 cM or $c = 0.00125$. If $D$ is initially nonzero between adjacent coding genes, then $D$ will decrease to 37% of its initial value in $1/0.00125 = 800$ generations. The generation time in humans is about 25 years, so 800 generations represent about 20,000 years. In other words, the extent of linkage disequilibrium between adjacent coding genes in the human genome is expected to decay on a timescale comparable to major events in the history of modern humans, i.e., the colonization of North America and the arrival of agriculture in Europe or the domestication of horses, sheep, and cattle.

Another implication of the preceding theory gives us an additional reason for being concerned with LD. We first recall that Eq. (2.5) tells us that $D' = 1$ when $D$ takes either its maximum positive value or its minimum negative value. In that case, the set of equations in (2.2) implies that at least one of the haplotypes has 0 frequency. The reverse is also true, namely, that if only three of four haplotypes are present in a population, then necessarily $D' = 1$. (Understanding this fact allows a few students in population genetics courses to quickly answer examination questions that occupy the rest of their classmates for a considerable time.)

Now consider what happens when a mutation occurs at a locus that was previously monomorphic and is linked to another locus that is polymorphic. To be specific, suppose the *A/a* locus is polymorphic and the *B/b* locus is initially fixed for *b* (i.e., all individuals in the population carry the *b* allele). In a particular generation, *B* appears in one copy as a new mutant. When *B* appears, it does so on a chromosome that initially carries an *A* or an *a*, but it cannot appear on both types of chromosomes. Suppose it appears on an *A*-bearing chromosome. In that generation, there are then only three haplotypes, *AB*, *Ab*, and *ab*. The fourth haplotype (*aB*) is not present. Therefore, when *B* appears as a new mutant allele, we know that $D' = 1$ regardless of the frequency of *A* and regardless of the recombination distance between *A* and *B*. That is, when a new mutant allele appears, $D' = 1$ between it and every polymorphic locus on the same chromosome.

What happens to $D'$ after *B* appears depends on $c$. For loci sufficiently far apart on the same chromosome that they are effectively unlinked ($c = 1/2$), $D$ and hence $D'$ decrease rapidly to zero. For more closely linked loci ($c < 1/2$), $D$ and $D'$ decrease more slowly, with the rate of decrease being slowest for very closely linked loci. Thus, each new mutant will be expected to be in strong LD with alleles at closely linked polymorphic loci for a long time, roughly $1/c$ generations. That prediction is valid for every new mutant. Because mutants are appearing every generation, the overall pattern of LD we expect is one in which $D'$ is large between closely linked loci and then decreases with increasing recombination distance between loci.

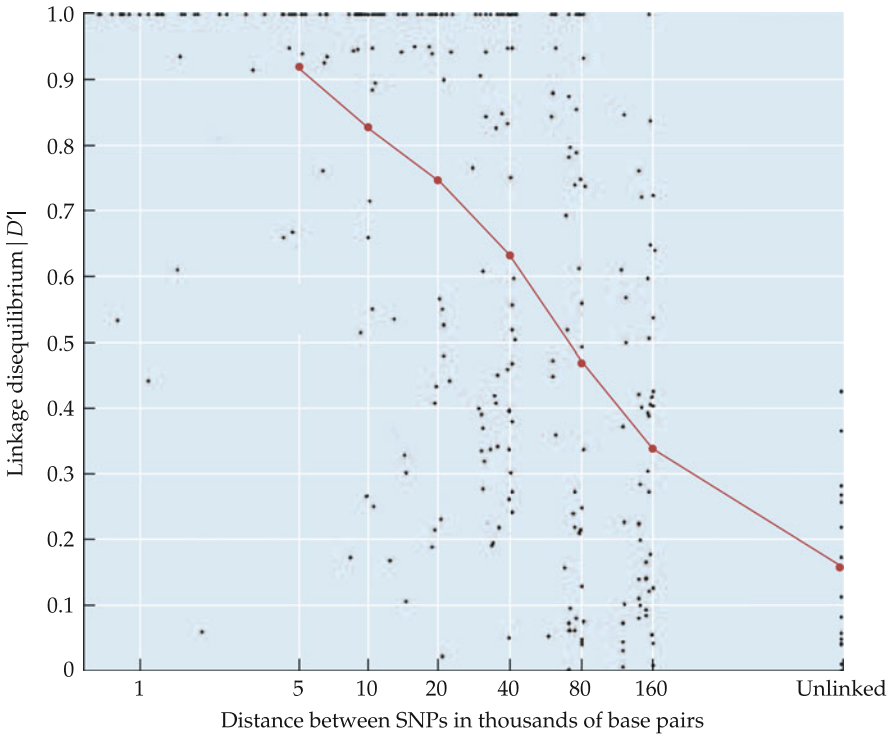


**Fig. 2.2** D′ plotted against distance separating SNPs in the human genome. The line indicates the average values of $D'$. (Reproduced with permission from Nielsen and Slatkin, 2013, *An Introduction to Population Genetics: Theory and Applications,* p. 121, Oxford University Press)

This prediction ignores other population genetic forces, particularly genetic drift and natural selection, which also affect LD, but it reflects the combined effects of recombination and random mating which affect the whole genome.

This prediction is consistent with many observations of LD in the human and other genomes. In humans, significant LD is usually found between polymorphic nucleotide positions that are separated by 50 kb or less but usually less so between sites separated by 100 kb or more (Reich et al. 2001). There is, however, considerable variation in $D'$ values even between sites separated by the same distance (Fig. 2.2), something that is not predicted by the simple theory presented so far.

## 2.6 Genetic Drift and LD

The preceding theory assumes a population of effectively infinite size. That is what allowed us to assume that allele frequencies do not change from generation to generation. Real populations are of finite size, and that implies that allele

frequencies change from generation to generation because of genetic drift (Wright 1931). Genetic drift also affects the extent of LD. The mathematical theory that predicts the effect of drift on $D$ is too complicated to be presented here, but the main conclusions from the theory are relatively simple (Hill and Robertson 1968; Ohta and Kimura 1969). At an equilibrium under drift, random mating, and recombination, genetic drift will maintain some LD between closely linked sites. Although the expected value of $D$ is 0, the expectation of $D^2$ is nonzero and decreases roughly with $1/c$ as $c$ increases.

## 2.7 Genealogical Interpretation of LD

There is a close relationship between the gene genealogies of two loci and the extent of LD between them. Refer to Chap. 1 for a discussion of gene genealogies for a single locus. First, consider the case in which there is no recombination between the $A$ and $B$ loci. Because there is no recombination, the gene genealogies of the two loci are the same, as shown in Fig. 2.3. If there is only one mutation at each of the two loci, as shown, no more than three of the four possible haplotypes will be present in the sample. Which haplotype is missing depends on where on the genealogy the two mutations occur. As shown in Fig. 2.3, $AB$ is missing, but if $B$ instead arose on one of the descendent branches carrying $A$, then $aB$ would be missing. And if $A$ and $B$ happened to have arisen on the same branch, then only $ab$ and $AB$ haplotypes would be present. Therefore, in the absence of recombination and recurrent mutation, $D' = 1$ necessarily, as noted above. It follows that, if $D' < 1$, either recombination or recurrent mutation occurred. At the level of individual nucleotides, recurrent mutation is unlikely, which implies that observing $D' < 1$ for two loci indicates

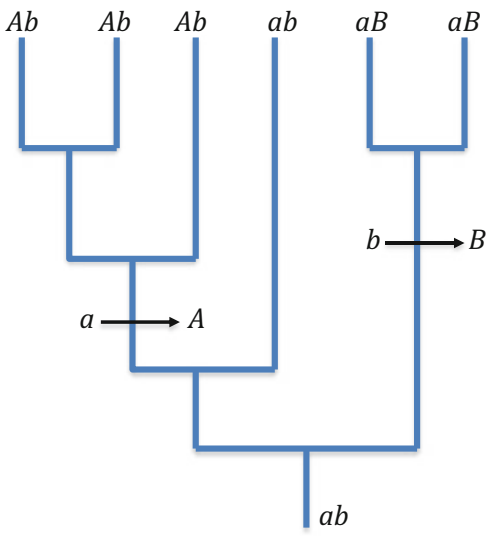


**Fig. 2.3** Illustration of the gene genealogy of two completely linked loci, showing the generation of haplotypes by mutation

that recombination occurred between them. When there is recombination, the gene genealogies are no longer the same. Recombination has the effect of breaking the gene genealogy of one of the loci and attaching somewhere else on the genealogy of the other locus. When that occurs, the relationship between $D$ or $D'$ and the recombination rate is no longer simple, and the genealogical approach does not in general lead to tractable analytic results. The similarity of the genealogies at the two loci will be determined, in part, by the recombination rate between the two loci. Loci separated by smaller genetic distances will have genealogies that are more correlated with one another than loci which are farther apart.

## 2.8 Natural Selection and LD

If the genotypes at two linked loci affect survival and reproduction, the resulting natural selection can increase the extent of linkage disequilibrium under some conditions and even maintain permanent disequilibrium in the face of recombination (Lewontin and Kojima 1960; Felsenstein 1965; Karlin and Feldman 1970). The effect of selection on LD is weak, however, because it depends not on the selection coefficients themselves but on the degree of epistasis, which is necessarily smaller. To illustrate in a simple context, assume there is haploid selection on two linked biallelic loci (*A/a* and *B/b*). Let the relative fitnesses of the four haplotypes be $w_{AB}$, $w_{Ab}$, $w_{aB}$, and $w_{ab}$. Selection of this type will tend to increase $D$ if $R = (w_{AB}w_{ab})/(w_{Ab}w_{aB}) > 1$ (Felsenstein 1965). In other words, $R$ is greater than 1 when the increase in fitness from having $A$ and $B$ together ($w_{AB}/w_{ab}$) exceeds the product of the gains from having $A$ or $B$ separately ($w_{Ab}/w_{ab}$ and $w_{aB}/w_{ab}$).

For diploid populations, more complicated but similar conditions have been derived that show when selection can overcome the effects of recombination and random mating and maintain permanent LD. Roughly speaking, permanent LD can be maintained under restrictive conditions, namely, that there has to be overdominance in fitness at each locus and $c$ has to be less than a quantity that summarizes the extent of epistasis in fitness (Karlin and Feldman 1970). It is currently unclear whether epistasis in fitness among closely linked loci contributes to observable patterns of LD.

## 2.9 Genetic Hitchhiking

Natural selection at a locus affects the frequencies of neutral alleles closely linked to it, a process termed "genetic hitchhiking." (Maynard Smith and Haigh 1974) The idea is simple. As described above, when a new mutant arises, it is in complete LD ($D' = 1$) with alleles at linked polymorphic loci. If that mutant increases rapidly in frequency because it confers a selective advantage to carriers, then neutral alleles on the same chromosome will increase in frequency also. For example, if $B$ arises on an $A$-bearing chromosome and subsequently increases in frequency, $A$ also will increase in frequency. The result will be an excess of $AB$ chromosomes.

Recombination between the $A$ and $B$ loci will reduce that excess. Simple theory shows that, if the selection coefficient in favor of $A$ is $s$, then substantial LD will be created by hitchhiking at neutral loci for which $c < s$. That is, neutral loci that are very closely linked to the selected locus will remain in substantial LD with the advantageous allele, while more distant neutral loci will not.

These theoretical results have useful practical applications. If there is substantial LD in a region surrounding a functionally important allele, it is likely that the allele has increased in frequency recently because of positive selection. For example, the *A*– allele of the *G6PD* gene in a west African population has a frequency of 11%. Loci as far away as 700 kb are in significant LD with the *A*– allele, a distance much larger than the normal scale of LD in the human genome (Saunders et al. 2005). Data of this type not only indicate that the *A*– allele increased because of positive selection but also make it possible to infer that the selection coefficient in favor of *A*– was at least 0.05 and that it arose by mutation between 3000 and 6000 years ago (Slatkin 2008).

## 2.10 Population Subdivision

Population subdivision creates LD when there are local differences in allele frequencies. We can see why by considering a simple example. Suppose that two populations are fixed for different alleles at each of two loci, population 1 is fixed for $A$ and $B$, while population 2 is fixed for $a$ and $b$. In this case, every individual in both populations is doubly homozygous, either *AABB* or *aabb.* Next, suppose that a researcher who is concerned with LD at these two loci samples individuals from both populations. If the researcher does not realize that there are in fact two distinct populations, individuals from both would be combined into a single sample. The resulting sample would be a mixture of *AABB* or *aabb* individuals. In this sample, only *AB* and *ab* haplotypes would be present, so there is apparently perfect LD between these two loci ($D' = 1$). Yet, that conclusion is obviously an artifact of mixing individuals from two populations with quite different allele frequencies.

Although this example is an extreme case that can be understood without doing any calculations, the conclusion is quite general. If allele frequencies at two loci differ at all between two or more populations and if samples from those two populations are combined, there will in general more LD in the mixture than in the separate populations (Mitton et al. 1973; Nei and Li 1973). This effect is called the "two-locus Wahlund effect" because of its similarity to the classic Wahlund effect, which is the decrease in heterozygosity in a mixture of two or more populations. In the simple example, there are no heterozygous individuals at either locus, which is an extreme case of the Wahlund effect.

It is usually possible to distinguish the two-locus Wahlund effect from selection as a cause of LD because the Wahlund effect affects all pairs of loci at which allele frequencies differ among subpopulations, while selection will probably affect only one genomic region. Still, the two-locus Wahlund effect is important for the design and interpretations of genome-wide association studies (GWAS). GWAS

are discussed in greater detail in Chap. 5. Because a GWAS is designed to detect significant LD between alleles that cause a complex disease and SNP markers, the two-locus Wahlund effect can create a spurious signal of association if individuals from different subpopulations are mixed together in the cases and controls. The term "population stratification" is used in this context. It is difficult to completely eliminate the effects of population stratification even if care is taken not to combine individuals from different ethnic groups. The problem is that the actual extent of variation among subpopulations in the frequencies of causative alleles is unknown, and hence, it is not clear how narrowly defined a subpopulation has to be in order to eliminate the effect of population stratification. For example, in carrying out a GWAS in people of European ancestry, is it appropriate to include people of both northern and southern European ancestry in the same study or not? Including both would increase the sample size and hence increase the statistical power to detect significant associations but at the risk of inducing spurious false-positive associations. This trade-off is especially problematic for rarer complex diseases for which the total number of affected individuals might be small. One resolution of the problem is to allow for some population stratification by using overall genomic averages of LD, called "genomic controls," to infer the overall extent of LD created by subtle population stratification (Devlin et al. 2001).

Gene flow among populations that have diverged can maintain LD in each subpopulation separately. When there is gene flow, the organisms themselves do the mixing and create LD between all pairs of loci that differ in allele frequency among the subpopulations. Mathematical analysis shows that substantial LD between closely linked loci can be maintained by this mechanism (Mitton et al. 1973; Nei and Li 1973).

## 2.11 Conclusion

When the term linkage disequilibrium was introduced by Lewontin and Kojima (Lewontin and Kojima 1960) in 1960, it was in the context of an abstract mathematical model developed to understand the combined effects of natural selection and recombination in an infinitely large population at equilibrium. Extensive further mathematical studies of LD were carried out in the 1960s and 1970s, but there was almost no attempt to relate the theory to data because almost no information about closely linked loci was available. This was the era during which genetic variation was studied by detecting differences in electrophoretic mobility of proteins (Hubby and Lewontin 1966). Polymorphic protein-coding loci that could be studied with electrophoresis were not usually closely enough linked for LD to be detectable. Linkage disequilibrium remained a somewhat arcane and mathematically difficult part of population genetics, known and appreciated by only a few specialists.

That situation changed with the development of direct means of assessing polymorphisms at the DNA sequence level—first restriction fragment length polymorphisms (RFLPs), then microsatellite loci, and finally SNPs. Instead of being obscure, linkage disequilibrium became well-known, then fashionable, and finally

essential. By 1999, that situation changed drastically. In the program for the 1999 annual meeting of the American Society of Human Genetics, 17% of the paper titles or abstracts contained the term "linkage disequilibrium." Since then, LD has only increased in importance in human genetics and is the foundation of GWAS, discussed in Chap. 5. As genomic tools become more widely used in plant and animal populations, LD will assume equal prominence in evolutionary biology.

## References

Consortium H (2003) The international HapMap project. Nature 426:789–796
Consortium IH (2005) A haplotype map of the human genome. Nature 437:1299–1320
Consortium IH (2007) A second generation human haplotype map of over 3.1 million SNPs. Nature 449:851–861
Daly MJ, Rioux JD, Schaffner SF, Hudson TJ, Lander ES (2001) High-resolution haplotype structure in the human genome. Nature 29:229–232
Devlin B, Roeder K, Wasserman L (2001) Genomic control, a new approach to genetic-based association studies. Theo Popul Biol 60:155–166
Felsenstein J (1965) The effect of linkage on directional selection. Genetics 52:349–363
Geiringer H (1944) On the probability theory of linkage in Mendelian heredity. Ann Math Stat 15:25–57
HapMap3 (2010) Integrating common and rare genetic variation in diverse human populations. Nature 467:52–58
Hedrick PW, Thomson G, Klitz W (1986) Evolutionary genetics: HLA as an exemplary system. Academic, New York
Hill WG, Robertson A (1968) Linkage disequilibrium in finite populations. Theor Appl Genet 38:226–231
Hubby JL, Lewontin RC (1966) A molecular approach to study of genic heterozygosity in natural populations. I. Number of alleles at different loci in Drosophila Pseudoobscura. Genetics 54:577–594
Karlin S, Feldman MW (1970) Linkage and selection: two locus symmetric viability model. Theo Popul Biol 1:39–71
Lander ES (2001) Initial sequencing and analysis of the human genome. Nature 409:860–921
Lewontin RC (1964) The interaction of selection and linkage. I. General considerations; heterotic models. Genetics 49:49–67
Lewontin RC, Kojima K (1960) The evolutionary dynamics of complex polymorphisms. Evolution 14:458–472
Maynard Smith J, Haigh J (1974) The hitch-hiking effect of a favourable gene. Genet Res 23:23–35
Mitton JB, Koehn RK, Prout T (1973) Population genetics of marine pelecypods. 3. Epistasis between functionally related isoenzymes of Mytilus-Edulis. Genetics 73:487–496
Nei M, Li W (1973) Linkage disequilibrium in subdivided populations. Genetics 75:213–219
Ohta T, Kimura M (1969) Linkage disequilibrium at steady state determined by random genetic drift and recurrent mutation. Genetics 63:229–238
Reich DE, Cargill M, Bolk S, Ireland J, Sabeti PC, Richter DJ, Lavery T, Kouyoumjian R, Farhadian SF, Ward R et al (2001) Linkage disequilibrium in the human genome. Nature 411:199–204
Robinson WP, Cambon-Thomsen A, Borot N, Klitz W, Thomson G (1991) Selection, hitchhiking and disequilibrium analysis at three linked loci with application to HLA data. Genetics 129:931–948
Saunders MA, Slatkin M, Garner C, Hammer MF, Nachman MW (2005) The span of linkage disequilibrium caused by selection on G6PD in humans. Genetics 171:1219–1229

Slatkin M (2008) A Bayesian method for jointly estimating allele age and selection intensity. Genet Res 90:129–137
Weir BS (1996) Genetic data analysis II, Sunderland, Sinauer
Wright S (1931) Evolution in Mendelian populations. Genetics 16:97–159