25

Linkage Disequilibrium Mapping: The Role of Population History, Size, and Structure

N. H. Chapman
Department of Biostatistics
University of Washington
Seattle, Washington 98195

E. A. Thompson[1]
Departments of Biostatistics and Statistics
University of Washington
Seattle, Washington 98195

I. SUMMARY

Linkage disequilibrium mapping attempts to infer the location of a disease gene from observed associations between marker alleles and disease phenotype. This approach can be quite powerful when disease chromosomes are descended from a single founder mutation and the markers considered are tightly linked to the disease locus. The success of linkage disequilibrium map-

[1] To whom correspondence should be addressed.

ping in fine-scale localization has led to the suggestion that genome-wide association testing might be useful in the detection of susceptibility genes for complex traits. Such studies would likely be performed in small, relatively isolated founder populations, where heterogeneity of the disease is less likely. To interpret the patterns of association observed in such populations, we need to understand the effect of population size, history, and structure on linkage disequilibrium. In this chapter, we first review measures of allelic association at a single locus. Measures of association between two loci are described, and some theoretical results are reviewed. We then consider some methods for inferring linkage between a marker and a rare disease, focusing on those that model the ancestry of the disease chromosomes. Next we discuss factors whose effect on disequilibrium are understood, and finally we describe the characteristics of some human populations that may be useful for disequilibrium mapping of complex traits.

II. INTRODUCTION

The goal of genetic linkage analysis is to infer the location of a disease gene based on coinheritance of the disease phenotype with some genetic marker whose chromosomal location is known. Disequilibrium mapping attempts to do the same thing, only without benefit of the pedigree relating disease chromosomes to one another: It relies on allelic associations between marker alleles and disease phenotype and is based on the idea that strong associations will be due to linkage, rather than chance. Thus identity by descent (IBD) due to coancestry is inferred from identity by state (IBS) data, in the form of observed allelic associations.

In combination, linkage analysis and disequilibrium analysis have been quite successful in the localization of genes for a number of simple Mendelian disorders (e.g., Hästbacka *et al.*, 1992, 1994; Puffenberger *et al.*, 1994; Risch *et al.*, 1995; Goddard *et al.*, 1996). The two approaches are quite complementary in situations such as this. Linkage analysis using recombination events within a pedigree can map a disease locus to a region of approximately 1 cM (Boehnke, 1994). Disequilibrium analysis (also known as haplotype analysis) can then pinpoint the disease locus by assuming a common ancestor for the chromosomes carrying the disease allele, and using all recombinations on paths back to that ancestor.

The success of disequilibrium testing in this context has led a number of investigators (e.g., Risch and Merikangas, 1996; Brown and Hartwell, 1998) to consider the use of genome-wide disequilibrium testing as an approach to finding susceptibility loci for common complex diseases. In particular, small isolated populations are of interest, since disease observed in these populations may be due to one or two alleles present in founding individuals, potentially eliminating the problem of heterogeneity. Furthermore, recently founded small populations may exhibit more disequilibrium than larger outbred populations (Kruglyak, 1999a). The utility of a population for this kind of study depends on being able to distinguish disequilibrium maintained by linkage from background disequilibrium, which exists as a result of the population's size and structure. This problem motivates our consideration of the effect of population structure on disequilibrium.

In Section III, we review measures of association for a single locus and give an example of the effect of population structure on these measures. Measures of association between two loci are described in Section IV, and some theoretical results describing how their means and variances change over time are reviewed. In Section V, we consider some methods for inferring linkage between a marker and a rare disease, focusing on those that model the ancestry of the disease chromosomes. In Section VI we discuss factors (e.g., time, population size) whose effects on disequilibrium are understood, and finally, in Section VII, we describe the characteristics of some human populations that may be useful for disequilibrium mapping of complex traits.

III. IBD AND ALLELIC ASSOCIATIONS AT A SINGLE LOCUS

Wright (1922) introduced the single-locus measures of relationship in use today. As a measure of IBD, he defined the fixation index f. This is commonly called the *coefficient of inbreeding*, and as Malécot (1948, 1969) elaborated, it is the probability that the two genes at a single locus within an individual are IBD. A related quantity is the *coefficient of kinship* between two individuals, which is defined as the probability of IBD between two homologous genes, one segregating from each individual. The coefficient of inbreeding for an individual is equal to the coefficient of kinship between his or her parents. Wright also introduced a measure ρ of IBS, defined as the correlation between allelic states on uniting gametes. Consider a locus with alleles **A** and **a**, allele frequencies p_A and p_a, and genotype frequencies p_{AA}, p_{Aa}, and p_{aa}. For two uniting gametes, let $X = 1$ if the maternal gamete carries allele **A** ($X = 0$ otherwise) and similarly let $Y = 1$ if the paternal gamete carries allele **A** ($Y = 0$ otherwise). Then in an infinitely large population where matings happen between individuals whose coefficient of kinship is equal to f, Wright (1922) showed that

$$\rho = \mathrm{corr}(X,Y) = \frac{p_{AA} - p_A^2}{p_A p_a} = \frac{p_A f + p_A^2(1 - f) - p_A^2}{p_A p_a} = f. \qquad (25.1)$$

This equation, probably the first presentation of the relationship between IBD and IBS, demonstrates the special case where the two measures are equal.

Another concept related to allelic association at a single locus is Hardy–Weinberg Equilibrium, which was described independently by Hardy (1908) and Weinberg (1908). In an infinite random mating population, the genotype frequencies p_{AA}, p_{Aa}, and p_{aa} will be equal to p_A^2, $2p_Ap_a$ and p_a^2 respectively. This relationship between the genotype frequencies and the allele frequencies defines Hardy–Weinberg Equilibrium (HWE). Departures from HWE reflect associations between alleles at that locus. Equation (25.1) shows that when the population is in HWE, $\rho = 0$ and there is no association.

Population structure can give rise to allelic association. Consider in particular the example of population subdivision. Suppose there are k subpopulations of equal size, and p_i denotes p_A in the ith subpopulation. Suppose that each subpopulation is in HWE, and let

$$\bar{p} = \frac{\Sigma p_i}{k} \qquad \text{and} \qquad \sigma^2 = \frac{\Sigma p_i^2}{k} - \bar{p}^2$$

denote the mean frequency of allele **A** in the pooled populations, and the variance of the k allele frequencies, respectively. Then the genotype frequencies in the pooled population are given by:

$$\begin{aligned} p_{AA} &= \bar{p}^2 + \sigma^2 \\ p_{Aa} &= 2\bar{p}(1 - \bar{p}) - 2\sigma^2 \\ p_{aa} &= (1 - \bar{p})^2 + \sigma^2. \end{aligned}$$

The pooled population is not in HWE; there is an excess of homozygotes. Note that

$$\rho = \frac{p_{AA} - p_A^2}{p_A(1 - p_A)} = \frac{\bar{p}^2 + \sigma^2 - \bar{p}^2}{\bar{p}(1 - \bar{p})} = \frac{\sigma^2}{\bar{p}(1 - \bar{p})},$$

so that when σ^2 is not zero, an association exists in the pooled population, even though no associations exist in the subpopulations (Wahlund, 1928). This example demonstrates the effect that population structure can have on allelic association at a single locus.

Estimation of f or ρ in human populations is difficult because it involves the estimation of variances and covariances. Edwards (1971) used data from the ABO blood group to consider the joint estimation of ABO allele frequencies and the level of IBD (f) in a population. He demonstrated that because of a singularity in the likelihood, it is very difficult to get good estimates of small values of f, which are likely found in human populations. Morton *et al.* in "The bioassay of kinship" (1971) capitalized on an important point: population history is shared by all loci, and therefore better estimates of f or ρ can be obtained by pooling information from several unlinked loci. The bioassay of kinship used allelic associations resulting from isolation by distance in a structured population to estimate ρ, from which the authors hoped to infer coancestry (f).

In practice, estimates of ρ have often been used in place of estimates of f, as if the two processes are equivalent. This is not generally the case, because allele frequencies drift, particularly in small populations. Thompson (1976) considered a single small ($N < 500$) population with no hierarchic structure and examined the joint evolution of heterozygosity ($H = 1 - \rho$) and nonidentity ($1 - f$) at a single locus, over time scales of up to 50 generations. In populations of this sort, which are typical of human populations of interest, the variances of both processes were very high, and correlations between the two were generally low and occasionally negative.

IV. ALLELIC ASSOCIATIONS BETWEEN TWO LOCI

The gametic correlation ρ described in the preceding section is a locus-specific measure of allelic association between two gametes. A measure of allelic association of more interest in the context of mapping is that between two loci on the same gamete. In this section, we define two such measures and review theoretical properties of their means and variances in finite populations. Consider two loci, one with alleles **A** and **a**, the other with alleles **B** and **b**. Let p_A and p_B denote the frequencies of alleles **A** and **B**, respectively. If the four possible haplotypes **AB**, **Ab**, **aB**, and **ab** occur with frequencies p_{AB}, p_{Ab}, p_{aB}, and p_{ab}, respectively, then let

$$D = p_{AB} - p_Ap_B$$

or equivalently

$$D = p_{AB}p_{ab} - p_{aB}p_{Ab}.$$

If the alleles **A** and **B** are independently distributed on haplotypes according to their allele frequencies, $D = 0$. A population in which D is nonzero is said to exhibit disequilibrium.

Robbins (1918), who was the first to suggest the use of D, studied its properties in an infinite population. He showed that for loci separated by a recombination fraction of θ,

$$D_t = (1 - \theta)D_{t-1}.$$

Thus disequilibrium decays to zero over many generations (when $\theta > 0$), so that in the limit, the alleles at each locus are are independently distributed over haplotypes in the population.

To relate D to the measure ρ discussed earlier, consider a randomly selected haplotype, and let $X = 1$ if the haplotype carries allele **A** at locus 1 ($X = 0$ otherwise) and let $Y = 1$ if the haplotype carries allele **B** at locus 2 ($Y = 0$ otherwise). Then $D = \text{cov}\,(X,Y)$. This suggests the use of a measure of disequilibrium directly analogous to ρ, given by

$$r = \text{corr}(X,Y) = \frac{D}{\sqrt{p_A p_a p_B p_b}}. \qquad (25.2)$$

This measure is sometimes preferred because it is less sensitive to allele frequencies than D, but its evolution is much more difficult to study because allele frequencies change over time.

Studies of the properties of disequilibrium in finite populations are of more interest, since human and animal populations often are small or are descended from groups with small numbers of founders. Earlier work concentrated exclusively on random mating populations, presumbly because of the complexity of the calculations involved.

A. Expectation of *D* in a finite population

Karlin and McGregor (1968) considered a population consisting of N diploid individuals, corresponding to $2N$ haplotypes. These $2N$ haplotypes give rise to the next generation by donating gametes to a "gamete pool" from which the $2N$ haplotypes of the next generation are randomly selected, with replacement. This is known as the "random union of gametes" model. The gamete pool is produced by considering all diploid genotypes possible in the parent generation, each contributing in proportion to its probability, which is assumed to be equal to the product of the haplotype frequencies. For example, consider the diploid genotype **AB/ab**. This genotype is assumed to occur with probability $2p_{AB}p_{ab}$, and it contributes gametes of types **AB, Ab, aB,** and **ab** with probabilities $\frac{1}{2}(1 - \theta)$, $\frac{1}{2}\theta$, $\frac{1}{2}\theta$, and $\frac{1}{2}(1 - \theta)$, respectively. This model has also been called the **haploid** model, since it is equivalent to the case of each haplotype in the offspring generation being the result of a "mating" of two randomly selected haplotypes in the parent generation. Karlin and McGregor formulated the model as a Markov chain, with state-space described by the 4-vector of haplotype counts. They showed that

$$E(D_{t+1}) = \left(1 - \frac{1}{2N}\right)(1 - \theta)E(D_t)$$

and stated that

$$\text{Var}(D_t) \sim \gamma\mu^t,$$

where γ is a positive constant depending on the initial conditions, and $\mu > (1 - 1/(2N))(1 - \theta)$. While both $E(D_t) \rightarrow 0$ and $\text{Var}(D_t) \rightarrow 0$ the variance of D approaches zero at a slower rate than its expectation. This suggests that even when enough generations have passed for $E(D_t)$ to be close to zero, $\text{Var}(D_t)$ may be large, and therefore a particular population may have a value of D_t quite different from zero.

Watterson (1970) considered the same problem, using a slightly different model for random mating. The model, called the "random union of zygotes" model, assumes a constant population size of N diploids. Each member of the offspring generation is obtained by randomly selecting (with replacement) two individuals from the parent generation, and generating a gamete from each. This model is also referred to as the **diploid** model because it retains the diploid nature of the individuals in the parent generation. Watterson also used a Markov chain formulation, with state-space described by the 10-vector of diploid genotype counts. He showed that

$$E(D_{t+1}) = \left(1 - \frac{1}{2N} - \theta\right)E(D_t),$$

a result also found by Hill and Robertson (1966). The expected disequilibrium approaches zero somewhat more slowly for the haploid model, although the difference is small for small values of $\theta/(2N)$. In fact if $\theta = 0$, the two models are exactly equivalent.

B. Variance of *D* and ρ

The papers cited in Section IV.A confirmed that disequilibrium decays to zero in a finite randomly mating population, just as it does in an infinite population, but the rate of decay is slower for smaller populations. Karlin and McGregor's work suggests that the variability of disequilibrium could be quite important in small populations. This problem was first explored by Hill and Robertson (1968) for the haploid model, assuming a population initially in equilibrium (i.e., $D_0 = 0$). They defined a vector of moments

$$y = \begin{pmatrix} E[p_A p_a p_B p_b] \\ E[D(p_a - p_A)(p_b - p_B)] \\ E[D^2] \end{pmatrix}$$

and used the multinomial distribution of the haplotype frequencies to find a matrix $M(N,\theta)$ such that $y_{t+1} = M(N,\theta)y_t$. For the special case of $\theta = 0$, they derived an explicit formula for $E[D_t^2]$. For values of θ other than zero, the value of $E[D_t^2]$ can be found by iteratively applying $\mathbf{M}$ to y_0.

Weir and Hill (1980) considered the same problem for several types of random mating in a diploid population:

- MS: a monoecious population of size N in which selfing is allowed (equivalent to Karlin and McGregor's diploid model).
- ME: a monoecious population of size N in which selfing is not allowed.
- DR: a dioecious population of M males, F females, and effective population size $N = 4MF/(M + F)$, where each child is the offspring of a random pairing.
- DH: a dioecious population with lifetime pairing, M males, each mated to s females. Thus $F = sM$ and $N = 4MF/(M + F)$. This model describes monogamy when $s = 1$.

Their method is based on two-locus descent measures, which are defined as the joint probability of non-IBD of $\boldsymbol{a}$ with $\boldsymbol{a'}$ and $\boldsymbol{b}$ with $\boldsymbol{b'}$, where $\boldsymbol{a}$ and $\boldsymbol{a'}$ denote alleles at one locus on different chromosomes, and $\boldsymbol{b}$, $\boldsymbol{b'}$ denote alleles at a second locus. There are three classes of descent measures, described as di-, tri-, and quadrigametic, according to whether the alleles being compared are on two, three, or four chromosomes. For example, a trigametic descent measure would be Pr($\boldsymbol{a}$ is not IBD to $\boldsymbol{a'}$, and $\boldsymbol{b}$ is not IBD to $\boldsymbol{b'}$), with $\boldsymbol{a}$, $\boldsymbol{a'}$, $\boldsymbol{b}$, and $\boldsymbol{b'}$ as depicted in Figure 25.1. Weir and Hill (1980) show that for a population starting in equilibrium, $\mathrm{Var}(D_t) = E(D_t^2)$ is a simple transformation of the descent measures in generation t and the initial allele frequencies. In a contemporaneous paper, Weir *et al.* (1980) developed transition matrices to describe the two-locus descent measures at time $t + 1$ as a function of the two-locus descent measures at time t, for each of the four types of random mating just listed. The transition matrices depend on the mating system, the size N of the population, and the recombination fraction θ between the two loci. Thus Weir and Hill's method allows the exact calculation of $\mathrm{Var}(D_t)$ for a particular N and θ: first the values of the relevant descent measures are calculated, and then they are transformed appropriately.

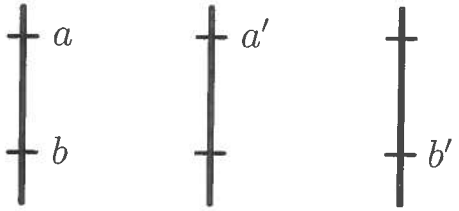

Figure 25.1 Three chromosomes.

Figure 25.2 shows the variance of D_t across monogamous random mating populations of size $N = 20$ or $N = 50$, where the founding population was in equilibrium, and the initial allele frequencies were $p_A = p_B = 0.5$. In this situation, although $E(D_t)$ over all populations is zero, the high variance of D_t, particularly for small recombination fractions, suggests that the value of D_t in any particular population could be quite different from zero. The plots suggest that for larger populations, the peak variance is smaller and happens later in time. There are a number of examples of relatively young human isolates who were founded by small numbers of individuals, so it seems likely that some of these populations will exhibit substantial disequilibrium between linked loci.

Hill and Robertson (1968) and Weir and Hill (1980) were also interested in the behavior of the measure r [see Equation (25.2)] in finite populations. Hill and Robertson approached the problem by simulation, using a haploid population of size 16 chromosomes, starting with allele frequencies $p_A = p_B = 0.5$, and $D_0 = 0$. The correlation r has the property of being undefined when either of the loci is fixed, so they based their estimates of $E(r_t^2)$ on populations in which both loci were still segregating. These simulations apparently show $E(r_t^2)$ approaching a limiting value depending on θ, but the authors note that $E(r_t^2)$ is necessarily not well estimated for large numbers of generations, because so many lines have fixed. Weir and Hill (1980) chose to approximate $E(r_t^2)$ by $E(D_t^2)/E(p_A^t p_a^t p_B^t p_b^t)$, since both these quantities can be obtained exactly from their methods. This approximation does not appear to have been tested in the small populations that are of interest, so we do not discuss their findings further here.

Sved (1971) studied the joint evolution of IBD at linked loci in a finite random mating population of size N and found an exact expression for $E(r_t^2)$. He showed that $E(r_t^2) = Q_t$, where Q_t is defined as the probability, for two haplotypes IBD at locus A in generation t, that there have been no crossover events between loci A and B on either of the two pathways from the common ancestor at locus A. Sved shows that for the haploid model, it is possible to obtain an iterative equation for Q_t as a function of Q_{t-1}; thereby showing that for a

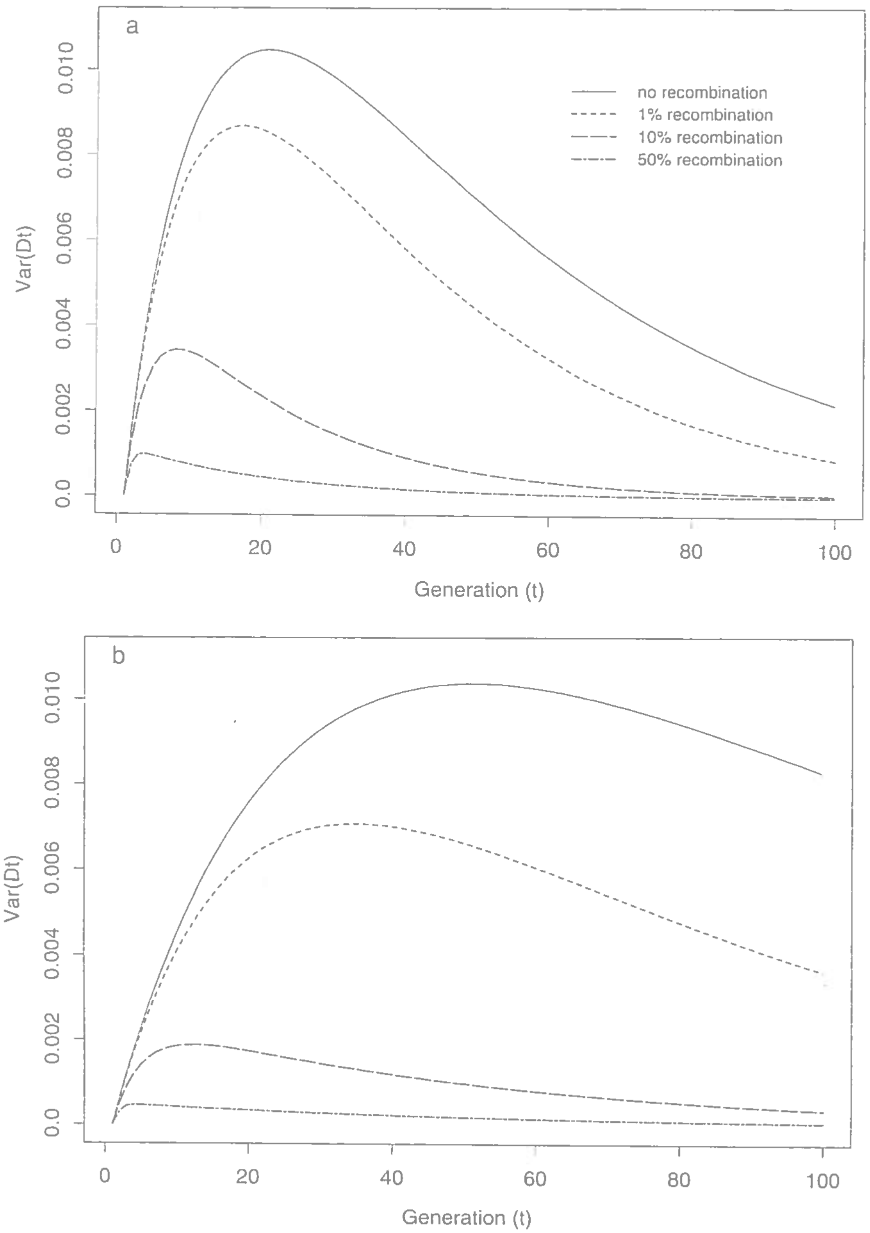

Figure 25.2 Var (D_t) as a function of time, across monogamous random mating populations: (a) $N = 20$ and (b) $N = 50$.

population that starts in equilibrium (i.e., $r_0^2 = 0$),

$$E(r_t^2) = \frac{1}{1 + 4N\theta}\left\{1 - \left[\left(1 - \frac{1}{2N}\right)(1 - \theta)^2\right]^t\right\}, \tag{25.3}$$

for small values of θ. As $t \to \infty$ in Equation (25.3),

$$E(r_t^2) \to \frac{1}{1 + 4N\theta}. \tag{25.4}$$

This equation gives an expression for $E(r^2)$ in a population that has been evolving long enough to be close to drift–recombination equilibrium. For small populations, this approximation may be valid only for very large t. This two-locus result is similar in spirit to Wright's demonstration of the conditions under which $\rho = f$ at a single locus [see Equation (25.1)] in that it relates the square of the correlation, a measure based on IBS, to a quantity that is a probability of IBD.

V. ESTIMATION OF θ FROM OBSERVED ASSOCIATIONS

The possibility of strong disequilibrium between tightly linked loci has prompted attempts to use the observed disequilibrium between two loci to infer the recombination fraction θ between them.

Chakravarti *et al.* (1984) considered the use of haplotype data from African-American, Italian, Greek, and Indian populations to estimate the recombination rate in a region of the human-β-globin gene cluster. From Sved's result in Equation (25.4), they wrote

$$\frac{1}{r^2} - 1 = 4Nkd,$$

where k is the rate of recombination per kilobase (kb) and d is the physical distance between two loci, measured in kilobases ($\theta = kd$). This inspired a regression of $1/r^2 - 1$ on d for a selection of pairs of loci known distances apart, yielding an estimate of k in terms of $4N$. By assuming a value of N that represents the average size of the population over its evolutionary history, an estimate of k is obtained. Weir and Hill (1986) point out some flaws in this approach. First, Equation (25.4) is for a population with two segregating sites whose haplotypes were initially in equilibrium. In the situation considered by Chakravarti *et al.* (1984), it seems more likely that the variable sites arose by mutation, and so there was initial disequilibrium. Weir and Hill (1986) show that this is an important factor, which renders Equation (25.4) inappropriate in this context. Second, Chakravarti *et al.* (1984) ignore the fact that their estimates are based on a relatively small sample of individuals, and therefore there is substantial sampling error in the estimation of r^2. Thompson *et al.* (1988) considered this issue further and demonstrated that power to detect disequilibrium can be very low, particularly if the rare alleles are in repulsion phase. While the work of Chakravarti *et al.* is an important early attempt to glean useful information from observed disequilibria, it is limited by the assumptions of the model used and by the inherent variability of the process being sampled.

Other approaches to drawing inferences about θ from observed disequilibria have attempted to take into account the history of the population in which the disequilibria are observed. When a new variant first arises, it necessarily exists on one and only one chromosomal background. There is initially complete association between the new variant and the alleles at other loci on that ancestral haplotype. This initially strong association is eroded away over subsequent generations by recombination between the locus where the variant arose and neighboring loci. Edwards (1981) elaborated on this concept by describing the "half-life" of a haplotype. For a chromosome segment whose ends are separated by recombination fraction θ, the half-life is the time (in generations) until the probability that the segment remains intact is 50%. For example, a segment for which $\theta = 0.005$ has a half-life of approximately 3500 years (140 generations), whereas a segment in which recombination happens much more frequently has a much shorter half-life (e.g., $\theta = 0.02$ corresponds to a half-life of 900 years). This demonstrates the importance of the time depth of the variant of interest to the genetic scale over which one might hope to see a conserved ancestral haplotype.

Segments of conserved ancestral haplotype are the cause of association between the variant and alleles at nearby loci. Since associations are observed by sampling multiple haplotypes, it is not recombination events on a single lineage back to the ancestral haplotype that are important, but rather, recombination events on all sampled lineages tracing back to that ancestral haplotype. Therefore the important parameter in determining the likely length of the conserved segment from a sample of haplotypes bearing the variant is the total number of meioses on the entire tree relating the sample to the original ancestral haplotype (Aranson *et al.*, 1977).

Arnason *et al.* (1977) were interested in estimating an upper bound for the recombination fraction between the **HLA-B** locus, and the **Bf** locus. Their data consisted of 100 apparently unrelated haplotypes, sampled from Iceland, northwest Newfoundland and Labrador, central England, and Norway. The haplotypes were chosen because they all bore allele **B8** at **HLA-B.** All haplotypes also carried the **S** allele at locus **Bf,** providing evidence of strong disequilibrium. To estimate an upper bound for θ, Aranson *et al.* assumed that the observed allelic association in these apparently diverse populations was due to a common ancestral haplotype, which existed in approximately 3000 B.C. They then reasoned that the probability of none of c haplotypes experiencing visible recombination over t generations is given by:

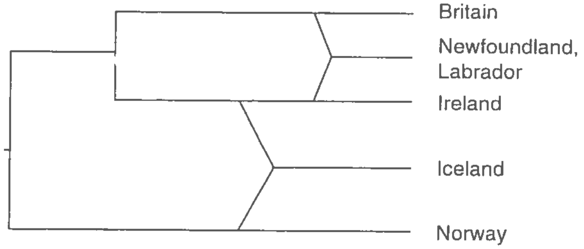

Figure 25.3 Evolutionary relationships of populations sampled by Arnason *et al.* (1977).

$$\Pr(\text{no visible recombinations}) = \left(1 - \frac{\theta}{3}\right)^{ct}, \qquad (25.5)$$

since the frequency of the **S** allele at **Bf** is approximately two-thirds. To apply this reasoning to their data set, they used the tree shown in Figure 25.3, to describe the ancestry of their sample of haplotypes. Using the demographic histories of the populations as a guide, they obtained a minimal number of ancestral haplotypes (c) in each branch, and the number of generations (t) over which they existed, to estimate $\Sigma\ ct \approx 2000$ over all links in the tree. Only very small values of θ produce small enough values of $(1 - \theta/3)^{ct}$ to be consistent with the observed data, indicating very tight linkage between **HLA-B** and **Bf.** This paper appears to be the first to use the ancestral relationships of the sampled haplotypes along with the observed associations to infer the recombination fraction between two loci. Thompson (1978) estimated more precisely the number of ancestors of the 100 sampled haplotypes at particular times in history, and obtained a result whose practical implications were very similar to those of Arnason *et al.* (1977).

More recently, efforts to infer θ from disequilibria have focused on rare recessive diseases, often in isolated populations, where all chromosomes carrying the disease variant are assumed to be descended from a single founder haplotype. We restrict our attention here to methods that explicitly model the ancestry of the disease chromosomes. Thompson and Neel (1997) modeled the ancestry of a sample of chromosomes carrying a rare monophyletic variant. Conditional on the expected total population of variant-bearing haplotypes at all times in history, they employ a continuous time Moran model approximation to obtain the distribution of the coalescence times. The distribution is intractable, but they describe a simulation approach which allows the sampling of ancestries from the correct distribution.

Hästbacka *et al.* (1992), who considered a sample of Finnish diastrophic dysplasia (DTD) chromosomes, estimated θ between the disease locus and a marker locus where association was observed. Using the same reasoning as Arnason *et al.* (1977), they observed that

$$\pi = (1 - \theta)^t \approx e^{-t\theta}, \tag{25.6}$$

where π = Pr (randomly sampled disease chromosome still carries ancestral marker allele). Assuming that the ancestral marker allele is that which was most common in their sample of disease chromosomes, they then equated the observed proportion of disease chromosomes carrying that allele to $e^{-t\theta}$ and solved for θ using $t = 100$ (based on the population history). They obtained an estimate of 64 kb for the distance between the disease locus and the marker of interest, which proved to be amazingly accurate—a gene for DTD 70 kb from the marker in question was later cloned (Hästbacka *et al.*, 1994). Despite the accuracy of the estimate in this example, questions remain about the usefulness of this approach. Equation (25.6) describes the probability over all possible evolutionary histories of the population from founding to the present, while the data available are necessarily from a single observation of the evolutionary process. The estimate obtained is therefore a moment estimator based on a single observation. The confidence bounds suggested by the authors may not be appropriate: Kaplan *et al.* (1995) showed by simulation that the upper bound was less than the true value of θ over 40% of the time. The approach of Hästabacka *et al.* is an advance over the work of Chakravarti *et al.* (1984) because it does not assume that the population is in drift–recombination equilibrium. However, neither does it explicitly model the history of the population.

Kaplan *et al.* (1995) presented the first likelihood approach to inference of θ from disequilibria, using simulation to take into account the population history. Assuming a single disease mutation occurring on a particular marker background at time zero, the number of disease chromosomes on that background at generation $t + 1$ is modeled as a function of the number at generation t. The disease chromosome population can then be simulated up to a time representing the time of sampling, and the probability of the observed data is calculated by means of either a binomial (for a diallelic marker) or a multinomial (for a multiallelic marker) distribution. Realizations that produce numbers of disease chromosomes inconsistent with that observed in the sampled population are excluded from the calculation, effectively conditioning on the observed disease allele frequency. Simulation at many values of θ gives a likelihood curve.

Rannala and Slatkin (1998) present an approach similar to that of Kaplan *et al.* (1995), but they use a coalescent to model the ancestry of the disease variant. Conditional on the time of the initial mutation, and the number of disease alleles sampled, Rannala and Slatkin obtain a realization from the joint distribution of coalescence times for the sample. For a simple of i disease chromosomes, there are i coalescence times, counting the time at which the mutation first occurred. The number of disease chromosomes carrying a particular marker allele immediately after coalescence event $i + 1$ depends on the number carrying that allele immediately after event i, the mutation rates, and the recombination fraction between the disease and the marker. Forward simulation of the disease haplotype population continues until the number of disease chromosomes carrying that marker allele immediately after the ith coalescent event is realized. The probability of the observed data at the time of sampling is then calculated conditional on this realization. Repeated simulation at different values of θ gives a likelihood curve. Like the method of Kaplan *et al.* (1995), this method can produce realizations of the disease haplotype population that are almost incompatible with the data and therefore make a very small contribution to the likelihood.

Graham and Thompson (1998) also consider disequilibrium likelihoods in the situation of a rare monophyletic disease mutation. They assume that the pattern of population growth is known (although it does not need to be constant), and that the time of the initial mutation is known. The ancestry relating the sampled chromosomes is realized by means of a two-stage process similar to that of Thompson and Neel (1997): (1) the size of the ancestral population is realized for each generation, from the present back to the time of the initial mutation, and (2) the coalescent relating the sampled haplotypes is realized, conditional on the ancestral population sizes. Once the coalescent has been realized, recombination events between the disease locus and the marker locus are placed on the tree, with probability depending on the branch length and the recombination fraction θ. These recombination events define recombinant classes, where a recombinant class is defined as the subset of the current sample that is descended from a given recombination event. Graham and Thompson describe an analytical expression for the probability of the observed data conditional on the recombinant classes. The simulation of recombinant classes rather than allelic classes eliminates the problem of obtaining realizations that are incompatible with the data, and as a result, the method of Graham and Thompson is likely to be more computationally efficient.

All the likelihood methods discussed earlier were developed in the context of a rare monophyletic disease mutation. As our attention is turning to more common complex diseases, it is informative to consider the effect of heterogeneity on the strength of allelic associations. If heterogeneity is in the form of multiple disease mutations at a single locus, the picture is encouraging. Disease haplotypes carrying different disease mutations at the same locus will each exhibit association with the marker allele ancestral to that disease mutation. As long as the marker locus is reasonably polymorphic, these associations will still be apparent, as in the cases of cystic fibrosis (Tsui, 1995) and Werner's syndrome (Matsuomo *et al.*, 1997), where the mutation of highest frequency makes up 70 and 51% (respectively) of the disease haplotype population. Multiple unliked disease loci (rather than multiple disease mutations at a single locus) are more problematic. In such cases, the association between a marker allele and a disease allele at a closely linked disease locus will be swamped by the absence of an association between that marker allele and the other disease loci, to which the marker is not linked. The difficulty introduced by heterogeneity has led a number of authors (e.g., Chapman and Wijsman, 1998; Kruglyak, 1999b) to suggest the use of isolated founder populations for disequilibrium mapping, in the hope that because of small numbers of founders, or population bottlenecks, diseases observed in such populations will be homogeneous. Kruglyak (1999b) showed that for the disease to be homogeneous, either the population must be descended from a very small number of founders or the disease variant must be quite rare. To use founder populations effectively for gene mapping, it is important that we first understand the factors other than linkage that can influence disequilibrium in these populations.

VI. EFFECTS OF SIZE AND STRUCTURE ON ALLELIC ASSOCIATIONS

The problem of the effect of population size on observed disequilibrium can be compared to the effect of degree of relationship on observed IBD sharing, which is important in affected relative pair methods. Closely related pairs are expected to share large portions of their genome IBD, and therefore areas of sharing are easy to locate but do not localize genes well. Conversely, distantly related relative pairs are expected to share very little of their genome IBD, and as a result, the identification of shared regions (which is difficult) can localize a disease gene quite precisely (Thompson, 1997). Analogously, disequilibrium is expected to stretch over larger distances in small populations, and over smaller distances in large populations. Large regions detectable in small populations may be useful for initial mapping, whereas the small, more difficult-to-find regions of disequilibrium in larger populations are useful for fine-scale localization.

Graham (1998) has considered the effect of population growth rate on the shape of the coalescent relating disease haplotypes. Visualization of the coalescent relating disease chromosomes is helpful in thinking about the effects of growth rate on disequilibrium, since it is recombination events on this coalescent that reduce allelic association over time. For illustration, we consider two extreme examples of coalescent shape. Classical coalescent theory (Felsenstein, 1971; Kingman, 1982) was developed for a population of constant size, where the shape of the coalescent is similar to the one shown in Figure 25.4a. In this case, sampling additional haplotypes simply adds more tips to the tree. Most of the meioses in the tree are in the earliest branches, which means that most recombinations can be expected to occur in these branches. As a result, sampled disease haplotypes are likely to share recombination events, which means that allelic associations should be well preserved. Figure 25.4b shows an example of a so-called star phylogeny, which might exist when the population of disease chromosomes is growing extremely rapidly (Slatkin, 1996). In this case, coalescent events between haplotypes occurred long ago, and as a result, sampled haplotypes appear to be independent. Recombination events are not shared by many haplotypes, so allelic associations are greatly reduced. Graham (1998) showed that for a growing population, the coalescent is often similar in shape to that in Figure 25.4a, with most meioses being in the early branches of the tree. Even in a fast growing population (10% per generation), coalescent events are fairly evenly spaced, and therefore a substantial fraction of recombinations will happen early enough in time to be shared by many disease haplotypes.

Kruglyak (1999b) used a simulation approach to consider the effects of two different population growth scenarios on observed disequilibrium. For the case of a population growing at a constant rate since founding, smaller growth rates result in higher levels of disequilibrium after a given number of generations, as the foregoing coalescent argument would suggest. The second scenario considered a population founded 20 generations ago by 100 individuals, with a current size of 10,000 individuals. The population was assumed to have a period of constant size, either early or late in its history. Disequilibrium in the current population is larger when the period of constant size happened earlier in the population's history. This is because genetic drift in the constant growth period effectively reduces the size of the founding population, thereby increasing disequilibrium.

Unfortunately, the effect of population structure is much less well understood. In large populations, subdivision, isolation by distance, and admixture can all give rise to associations that could be misinterpreted as reflecting shared ancestry. Beerli (1999) has studied the effect of migration on the shape of the coalescent, and this work may give insight into the effect of migration on disequilibrium. In smaller populations, and within subgroups of larger ones, patterns of nonrandom mating, together with genetic drift, can also affect

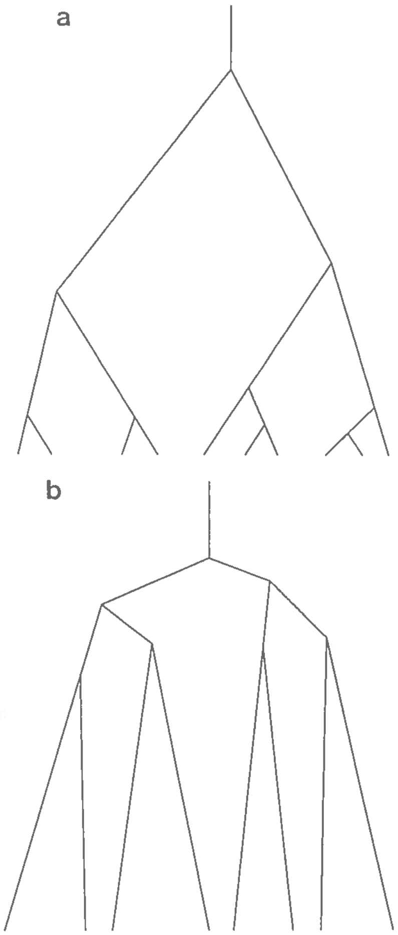

Figure 25.4 Examples of typical coalescent shapes for different growth rates: (a) constant size and (b) fast growth.

disequilibrium. To interpret the patterns of disequilibrium we can observe in the wide variety of human populations available for study, we must quantify the effects of each of these aspects of population structure. Then we must learn how to differentiate these associations from those due to common ancestry and linkage. In the next section, we describe some of the populations that have been suggested for disequilibrium mapping, with attention to known aspects of their size, history, and structure.

VII. STRUCTURE IN HUMAN POPULATIONS: SOME EXAMPLES

Table 25.1 shows examples of the wide variety of ages, sizes, and structures observed in human populations. Of the large isolates, the Amerindians are by far the oldest. About 500 of the earliest Americans are thought to have crossed the Bering land bridge, during a period about 30,000 years ago. The population grew quite slowly until about 8000 years ago, when agricultural practices were adopted, and the population grew rapidly to approximately 20 million by the time of Columbus (Fiedel, 1987; Denevan, 1992). This population must have been highly structured, both because of their enormous range (North and South America) which results in isolation by distance, and because of the tribal structure within smaller geographic areas. Because of the time depth of this population, disequilibrium between loci with respect to the initial founder population is unlikely to be detectable. Because of the recent rapid expansion of the population, however, most rare variants will have occurred recently (Thompson and Neel, 1997) and will be localized to a particular tribe or linguistic group. Disequilibrium around such a variant should be detectable in modern Amerinds.

The modern Japanese population was founded approximately 94 generations ago, by approximately 1000 rice-growing immigrants who emigrated from the mainland (Benedict, 1989). Relatively little is known about the growth of the population until the period 1603–1867, when the population size remained remarkably constant, at about 30,000 individuals. Political change in 1867 resulted in rapid growth, and the population of Japan today is about 120 million. This population is probably much less structured than the Amerindians,

Table 25.1 The Scope of Human Population Structure

Populations	Time since founding (years)	Number of founders	Modern size	Structure
Large				
Amerindians	30,000	500	20 million[a]	Tribal
Japanese	2,400	1,000	120 million	Homogeneous
Finns	2,000	1,000	5 million	Homogeneous
Small				
Icelanders	1,200	20,000	250,000	Homogeneous
Ch-SLSJ[b]	350	8,000	300,000	Unknown
Hutterites	350	80	36,000	Leut and colonies
Tiny				
Newfoundland	180	400	1,600	3 villages
Tristan da Cunha	180	19	300	Homogeneous

[a] At the time of Columbus.
[b] Charlevoix–Saguenay Lac Saint Jean.

although some historical subdivision due to geography seems likely. The time depth of the population (2400 years) proved ideal for the fine-scale localization of the Werner's syndrome gene, which apparently dates to the founding of the population (Graham and Thompson, 1998).

Another large population in which disequilibrium mapping of a rare allele has been successful is that of Finland. Finland was founded by about 1000 settlers, approximately 2000 years ago. The population has remained quite isolated since then, and numbers about 5 million (Nevanlinna, 1972). Like the Japanese, the Finnish population is appropriate for fine-scale localization because of its time depth, as Hästbacka *et al.* (1994) demonstrated when they cloned a gene for diastrophic dysplasia in Finns. Kruglyak (1999b) studied the properties of such a population by simulation and showed that even with this relatively small number of founders, a disease variant would have to be quite rare (< 1%) to be monophyletic in such a population.

The potential of smaller isolated populations has been less well explored. Iceland is a particularly interesting example. Iceland was first settled circa 900 by about 20,000 people who came primarily from Norway, but also from Ireland and Scotland (Bjarnason *et al.*, 1973). The population grew very rapidly both by births and continued immigration to about 70,000 by the eleventh century, and remained of roughly that size until the early 1900s, whereupon it grew dramatically to its current size of 250,000 people. While relatively isolated from other populations, there was mobility within the country, and as a result the population is probably quite homogeneous. There was likely considerable disequilibrium within the founding population because it was a Norse–Celtic mixture. Because of the shorter time depth (relative to Finns and Japanese), disequilibrium may be detectable over longer genetic distances. However, the unusually large size of the founding population makes it likely that even rare alleles would have been represented in multiple copies.

Of similar size, but much younger, is the population of the Charlevoix–Saguenay Lac Saint Jean (Ch-SLSJ) region of Quebec. Many of the residents of this area are the descendants of about 8000 French colonists who immigrated to the Charlevoix region during 1608–1760, while the area was under French control (Heyer, 1995). The population expanded into the Saguenay region in the mid-1800s and continued to grow both internally and through considerable immigration, to its size of about 300,000 today (Heyer and Tremblay, 1995). Although disequilibrium may persist over moderate distances in such a young population, the relatively high level of immigration could result in heterogeneity for all but the rarest diseases.

The Hutterites (Morgan and Holmes, 1982) are an isolate of similar age but much smaller (36,000 people). Descended from a remarkably small founder group (80 people), the Hutterites are a religious group who originated in Europe in the 1500s and immigrated to North America circa 1880. Upon arrival in North America, they split into three almost completely separate subdivisions (known as leut). Within leut, the Hutterite population is divided into colonies, and when growth necessitates division, a colony will split into two colonies. Immigration into Hutterite colonies is almost nonexistent. Disequilibrium may extend over large genetic distances in the Hutterites, since they are a very young population and are descended from so few founders. In addition, the population is unusually rich in structure, and the structure is very well documented. Study of the Hutterite population may therefore yield important insight into the effects of both large and small scale structure on disequilibrium.

Table 25.1 also shows two extremely small and young populations. The Newfoundland population includes the inhabitants of what are now three small villages, on the west coast of the island. Many of the inhabitants are descendants of a single founder couple and their children and grandchildren, who founded the population in 1820 (Marshall *et al.*, 1979). Others married into the population, and the current population traces back to about 400 founders. In 1975 the population numbered 1627 people, with 1521 descended from the original founding couple. The complete genealogy at that time contained just over 4000 people, 2600 of whom were descendants of the original founding couple. Because of the youth of the population, disequilibrium around a variant originating in the founder couple is likely to stretch over relatively long distances.

The other extremely small population is that of Tristan da Cunha, a very remote island in the south Atlantic. The population, numbering about 275 in 1961, is descended from a total of 19 ancestors, 10 of whom were on the island prior to 1855 (Roberts, 1971). These founders are of extremely diverse origin—some being immigrants from the United States or Britain and other being survivors of shipwrecks in the region. The population has exhibited steady but slow growth, with two severe bottlenecks: the population was reduced from 103 to 33 in 1856, and from about 100 in 1880 to about 60 in 1892. The population has grown steadily since then, but emigration has kept the population small. A population of this size is necessarily quite homogeneous, and therefore not likely to tell us much about the effects of structure on disequilibrium. However, the youth of the population, the diversity of its founders, and the two bottlenecks in its history suggest that considerable disequilibrium likely exists in this population.

Lonjou *et al.* (1999) compared observed disequilibria in two regions of the genome for a wide variety of populations, ranging from outbred large geographic areas such as Europe, the Near East, and the Americas, to small isolates, such as Basques, the Ainu of Japan, and the population on Tristan da Cunha. The Jewish populations of Europe, Africa, and the Middle East comprise another ethnic group considered by Lonjou *et al.* (1999) that shows a variety of histories and structures. Disequilibrium mapping in these populations has been considered by Risch *et al.* (1995) and Levy *et al.* (1996). Lonjou found that generally, levels of disequilibrium in isolated populations were only slightly higher than in outbred populations. The loci they considered were diallelic and had moderate allele frequencies, and so their results suggest that isolates may not be as useful for complex disease mapping as had been hoped. Since, however, all locus pairs considered were extremely tightly linked (<0.2 cM apart), the similarity between observed disequilibria in small isolates and large outbred populations may simply demonstrate the very large time scale required to break down such associations.

Kruglyak (1999a), in a discussion of the results of Lonjou *et al.* (1999), notes that these authors' conclusions are based on data for only two regions of the genome and therefore are subject to the unknown evolutionary histories of the two regions. He also notes that disequilibrium in the Ainu was significantly higher than in the large populations, suggesting that disequilibrium mapping might be feasible in this isolate and in others not considered in the Lonjou study. Each isolate has its own unique evolutionary history, and therefore some may prove more useful than others for disequilibrium mapping of complex disease.

Most notably, Kruglyak (1999a) points out the need for systematic empirical study of disequilibrium across the entire human genome, both within isolates and within larger outbred populations, to identify populations useful for disease mapping. We believe that empirical studies of disequilibrium in human populations should be complemented by theoretical research undertaken with the goal of understanding the effects of population history and structure on disequilibrium.

Acknowledgments

This work was supported in part by funding from the Burroughs Wellcome Fund for the Program in Mathematics and Molecular Biology and by National Institutes of Health grant GM-46255.

References

Arnason, A., Larsen, B., Marshall, W. H., Edwards, J. H., MacKintosh, P., Olaisen, B., and Teisberg, P. (1977). Very close linkage between *HLA-B* and *Bf* inferred from allelic association. *Nature* **268,** 527–528.

Beerli, P. (1999). Estimation of migration rates and population sizes in geographically structured populations. *In* "Advances in Molecular Ecology," (G. R. Carvalho, ed.), pp. 39–53. NATO Science Series A: Life Sciences. IOS Press, Amsterdam.

Benedict, R. (1989). "The Chrysanthemum and the Sword." Houghton Mifflin, Boston.

Bjarnason, O., Bjarnason, V., Edwards, J. H., Fredriksson, S., Magnusson, M., Mourant, A. E., and Tills, D. (1973). The blood groups of the Icelanders. *Ann. Hum. Genet.* **36,** 425–455.

Boehnke, M. (1994). Limits of resolution of genetic linkage studies: Implications for the positional cloning of human disease genes. *Am. J. Hum. Genet.* **55,** 379–390.

Brown, P. O., and Hartwell, L. (1998). Genomics and human disease: Variations on variation. *Nat. Genet.* **18,** 91–93.

Chakravarti, A., Buetow, K., Antonarakis, S., Waber, P., Boehm, C., and Kazazian, H. (1984). Nonuniform recombination within the human β-globin gene cluster. *Am. J. Hum. Genet.* **36,** 1239–1258.

Chapman, N. H., and Wijsman, E. M. (1998). Genome screens using linkage disequilibrium tests: Optional marker characteristics and feasibility. *Am. J. Hum. Genet.* **63,** 1872–1885.

Denevan, W. M. (1992). "The Native Population of the Americas in 1492." University of Wisconsin Press, Madison.

Edwards, A. W. F. (1971). Estimation of the inbreeding coefficient from ABO blood-group phenotype frequencies. *Am. J. Hum. Genet.* **23,** 97–98.

Edwards, J. H. (1981). Allelic association in man. *In* "Population Structure and Genetic Disorders, Proceedings of the Seventh Sigfred Juselius Foundation Symposia, New York" (A. W. Eriksson, ed.), pp. 239–256. Academic Press, New York.

Felsenstein, J. (1971). The rate of loss of multiple alleles in finite haploid populations. *Theor. Popu. Bio.* **2,** 391–403.

Fiedel, S. J. (1987). "Prehistory of the Americas." Cambridge University Press, New York.

Goddard, K. A. B., Yu, C. E., Oshima, J., Miki, T., Nakura, J., Piussan, C., Martin, G. M., Schellenberg, G. D., Wijsman, E. M., and members of the International Werner's Syndrome Collaborative Group (1996). Toward localization of the Werner syndrome gene by linkage disequilibrium and ancestral haplotyping: Lessons learned from analysis of 35 chromosome 8p11.1–21.1 markers. *Am. J. Hum. Genet.* **58,** 1286–1302.

Graham, J. (1998). Disequilibrium fine-mapping of a rare allele via coalescent models of gene ancestry. Ph. D. thesis, University of Washington, Seattle.

Graham, J., and Thompson, E. A. (1998). Disequilibrium likelihoods for fine-scale mapping of a rare allele. *Am. J. Hum. Genet.* **63,** 1517–1530.

Hardy, G. H. (1908). Mendelian proportions in a mixed population. *Science* **28,** 49–50.

Hästbacka, J., de la Chapelle, A., kaitila, I., Sistonen, P., Weaver, A., and Lander, E. (1992). Linkage disequilibrium mapping in isolated founder populations: Diastrophic dysplasia in Finland. *Nat. Genet.* **2,** 204–211.

Hästbacka, J., de la Chapelle, A., Mahtani, M., Clines, G., Reeve-Daly, M. P., Daly, M., Hamilton, B. A., Kusumi, K., Trivedi, B., Weaver, A., Coloma, A., Lovett, M., Buckler, A., Kaitila, I., and Lander, E. S. (1994). The diastrophic dysplasia gene encodes a novel sulfate transporter: Positional cloning by fine-structure linkage disequilibrium mapping. *Cell* **78,** 1073–1087.

Heyer, E. (1995). Mitochondrial and nuclear genetic contribution of female founders to a contemporary population in North East Quebec. *Am. J. Hum. Genet.* **56,** 1450–1445.

Heyer, E., and Tremblay, M. (1995). Variability of the genetic contribution of Quebec population founders associated with some deleterious genes. *Am. J. Hum. Genet.* **56,** 970–978.

Hill, W. G., and Robertson, A. (1966). The effect of linkage on limits to artificial selection. *Genet. Res.* **8,** 269–294.

Hill, W. G., and Robertson, A. (1968). Linkage disequilibrium in finite populations. *Theor, Appl. Genet.* **38,** 226–231.

Kaplan, N. L., Hill, W. G., Weir, B. S., (1995). Likelihood methods for locating disease genes in nonequilibrium populations. *Am. J. Hum. Genet.* **56,** 18–32.

Karlin, S., and McGregor, J. (1968). Rates and probabilities of fixation for two locus random mating finite populations without selection. *Genetics* **58,** 141–159.

Kingman, J. F. C. (1982). The coalescent. *Stochastic Process Appl.* **13,** 235–248.

Kruglyak, L. (1999a). Genetic isolates: Separate but equal? *Proc. Natl. Acad. Sci. USA* **96,** 1170–1172.

Kruglyak, L. (1999b). Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nat. Genet.* 22, 139–144.

Levy, E. N., Shen, Y., Kupelian, A., Kruglyak, L., Aksentijevich, L., Pras, E., Balow, J. E. B. Jr., Linzer, B., Chen, X., Shelton, D. A., Gumucio, D., Pras, M., Shohat, M., Rotter, J. I., Fischel- Ghodsian, N., Richards, R. I., and Kastner, D. L. (1996). Linkage disequilibrium mapping places the gene causing familial Mediterranean fever close to D16S246. *Am. J. Hum. Genet.* **58,** 523–534.

Lonjou, C., Collins, A., and Morton, N. E. (1999). Allelic association between marker loci. *Proc. Natl. Acad. Sci. USA* **96,** 1621–1626.

Malécot, G. (1948). "Les Mathématiques de l'Hérédite." Masson, Paris.

Malécot, G. (1969). "The Mathematics of Heredity." Freeman, San Francisco.

Marshall, W. H., Buehler, S. K., Crumley, J., Salmon, D., Landre, M.-F., and Fraser, G. R. (1979). A familial aggregate of Hodgkin's disease, common variable immunodeficiency, and other malignancy cases in Newfoundland. I. Clinical features. *Clin. Invest. Med.* **2,** 153–159.

Matsuomo, T., Imamura, O., Yamabe, Y., Kuromitsu, J., Tokutake, Y., Shimamoto, A., Suzuki, N., *et al.* (1997). Mutation and haplotype analyses of Werner's syndrome gene based on its genomic structure: Genetic epidemiology in the Japanese population. *Hum. Genet.* **100,** 123–130.

Morgon, K., and Holmes, T. M. (1982). Population structure of a religious isolate: The Dariusleut Hutterites of Alberta. *In* "Current Developments in Anthropological Genetics," Vol. 2, pp. 429–448. Plenum, New York.

Morton, N. E., Yee, S., Harris, D. E., and Lew, R. (1971). The bioassay kinship. *Theor. Popul. Bio.* **2,** 507–524.

Nevanlinna, H. R. (1972). The Finnish population structure—A genetic and genealogical study. *Hereditas* **71,** 195–236.

Puffenberger, E. G., Kauffman, E. R., Bolk, S., Matise, T. C., Washington, S. S., Angrist, M., Weissenbach, J., Garver, K. L., Mascari, M., Ladda, R., *et al.* (1994). Identity-by-descent and association mapping of a recessive gene for Hirschsprung disease on human chromosome 13q22. *Hum. Mol. Genet.* **3,** 1217–1225.

Rannala, B., and Slatkin, M. (1998). Likelihood analysis of disequilibrium mapping, and related problems, *Am. J. Hum. Genet.* **62,** 459–473.

Risch, N., and Merikangas, K. (1996). The future of genetic studies of complex human diseases. *Science* **273,** 1516–1517.

Risch, N., de Leon, D., Ozelius, L., Kramer, P., Almasy, L., Singer, B., Fahn, S., Breakefield X., and Bressman, S. (1995). Genetic analysis of idiopathic torsion dystonia in Ashkenazi Jews and their recent descent from a small founder population. *Nat. Genet.* **9,** 152–159.

Robbins, R. B. (1918). Applications of mathematics to inbreeding problems. II. *Genetics* **3,** 73–92.

Roberts, D. F. (1971). The demography of Tristan da Cunha. *Popul. Stud.* **25,** 465–469.

Slatkin, M. (1996). Gene genealogies within mutant allelic classes. *Genetics* **143,** 579–587.

Sved, J. A. (1971). Linkage disequilibrium and homozygosity of chromosome segments in finite populations. *Theor. Popul. Biol.* **2,** 125–141.

Thompson, E. A. (1976). Population correlation and population kinship. *Theor. Popul. Biol.* **10,** 205–226.

Thompson, E. A. (1978). The number of ancestral genes contributing to a sample of B8 alleles. *Nature* **272,** 288.

Thompson, E. A. (1997). Conditional gene identify in affected individuals. *In* "Genetic Mapping of Disease Genes," pp. 137–146. Academic Press, London.

Thompson, E. A., and Neel, J. V. (1997). Allelic disequilibrium and allele frequency distribution as a function of social and demographic history *Am. J. Hum. Genet.* **60,** 197–204.

Thompson, E. A., Deeb, S., Walker, D., and Motulsky, A. G. (1988). The detection of linkage disequilibrium between closely linked markers: RFLPs at the *AI-CIII* apolipoprotein genes. *Am. J. Hum. Genet.* **42,** 113–124.

Tsui, L. C. (1995). The cystic fibrosis transmembrane conductance regulator gene. *Am. J. Resp. Crit. Care Med.* **151,** S47–S53.

Wahlund, S. (1928). Zusammensetzung von Populationen and Korrelationsercheinungen vom Standpunkt der Vererbungslehre aus betrachtet. *Hereditas* **11,** 65–106.

Watterson, G. A. (1970). The effect of linkage in a finite random-mating population. *Theor. Popul. Biol.* **1,** 72–87.

Weinberg, W. (1908). Über den Nach weis der Vererbung beim Menschen. *Jahreshe. Verein Vaterl. Naturk. Wuttemberg* **64,** 368–382.

Weir, B. S., and Hill, W. G. (1980). Effect of mating structure on variation in linkage disequilibrium. *Genetics* **95,** 477–478.

Weir, B. S., and Hill, W. G. (1986). Nonuniform recombination within the human β-globin gene cluster. *Am. J. Hum. Genet.* **38,** 776–778.

Weir, B. S., Avery, P. J., and Hill, W. G. (1980). Effect of mating structure on variation in inbreeding. *Theor. Popul. Biol.* **18,** 396–429.

Wright, S. (1922). Coefficients of inbreeding and relationship. *Am. Nat.* **56,** 330–338.