Note

A Genealogical Interpretation of Linkage Disequilibrium

Gilean A. T. McVean

Department of Statistics, University of Oxford, Oxford OX1 3TG, United Kingdom

Manuscript received April 10, 2002
Accepted for publication July 15, 2002

ABSTRACT

The degree of association between alleles at different loci, or linkage disequilibrium, is widely used to infer details of evolutionary processes. Here I explore how associations between alleles relate to properties of the underlying genealogy of sequences. Under the neutral, infinite-sites assumption I show that there is a direct correspondence between the covariance in coalescence times at different parts of the genome and the degree of linkage disequilibrium. These covariances can be calculated exactly under the standard neutral model and by Monte Carlo simulation under different demographic models. I show that the effects of population growth, population bottlenecks, and population structure on linkage disequilibrium can be described through their effects on the covariance in coalescence times.

MEASURES of the nonrandom association between alleles at different loci, or linkage disequilibrium, are widely used to infer properties of population history, recombination, and the location of mutations contributing to disease susceptibility and adaptive evolution. Associations between alleles are generated by the stochastic nature of mutation and sampling in a finite population, as well as certain forms of geographical structure (*e.g.*, Ohta 1982), and natural selection (*e.g.*, Strobeck 1983). In contrast, recombination acts to break down such associations. Comparison of empirical patterns of linkage disequilibrium to those expected from population genetics theory, and across different genomic regions, can provide much information about the forces shaping genetic diversity.

The rise of coalescent theory (Kingman 1982) as a tool for interpreting patterns of genetic diversity in samples has led to a shift in focus in theoretical population genetics from mutations to genealogies. Most importantly, if mutations have no effect on organismal fitness, the genealogy of a sample can be separated entirely from the mutational process. Consequently, all information about important evolutionary parameters (such as demography and the action of selection at linked sites) is contained in the genealogy, which can be estimated only indirectly from the distribution of mutations among sampled chromosomes.

The question of how statistics of linkage disequilibrium relate to aspects of the underlying genealogy is therefore of considerable interest. Here I show that a quantity that approximates the expectation of a commonly used statistic of linkage disequilibrium, r^2, can be expressed in terms of covariances in coalescence times. The result provides an intuitive basis for understanding how linkage disequilibrium behaves under different demographic scenarios.

Address for correspondence: Department of Statistics, 1 S. Parks Rd., Oxford OX1 3TG, UK. E-mail: mcvean@stats.ox.ac.uk

GENEALOGICAL APPROACH

Linkage disequilibrium and identity coefficients: The r^2 statistic of linkage disequilibrium is equivalent to the square of the correlation coefficient between the alleles A at locus x and B at locus y,

$$r^2_{A(x)B(y)} = \frac{D^2_{A(x)B(y)}}{f_{A(x)}(1 - f_{A(x)})f_{B(y)}(1 - f_{B(y)})} \qquad (1)$$

(Hill and Robertson 1968), where $D_{A(x)B(y)} = f_{A(x)B(y)} - f_{A(x)}f_{B(y)}$ is the standard measure of linkage disequilibrium, with $f_{A(x)B(y)}$ indicating the frequency of chromosomes carrying the A and B alleles. Although it is impossible to derive a simple analytic expression for the expectation of (1), we can consider the related quantity of the ratio of expectations

$$\sigma^2_d = \frac{E[D^2_{A(x)B(y)}]}{E[f_{A(x)}(1 - f_{A(x)})f_{B(y)}(1 - f_{B(y)})]} \qquad (2)$$

(Ohta and Kimura 1971). The ratio of expectations (2) is a considerable overestimate of the expectation of r^2 if alleles at all frequencies are considered. However, if the expectation is conditioned on intermediate alleles frequencies (*e.g.*, >10%), the two are in close agreement (see Figure 3 and Hudson 1985).

Consider first the numerator in (2). The expectation of D is zero (irrespective of the recombination rate and demographic model), hence $E[D^2] = \mathrm{Var}(D)$. Strobeck and Morgan (1978) and Hudson (1985) showed that the expected square of disequilibrium can be written in terms of identity coefficients for sets of sequences,

$$\begin{aligned}\mathrm{Var}(D) &= E[f^2_{A(x)B(y)}] - 2E[f_{A(x)B(y)}f_{A(x)}f_{B(y)}] \\ &\quad + E[f^2_{A(x)}f^2_{B(y)}] \\ &= F_{x(ij)y(ij)} - 2F_{x(ij)y(ik)} + F_{x(ij)y(kl)}. \end{aligned} \tag{3}$$

The three terms are, respectively, the probability that two sequences i and j are identical in state at both sites x and y; the probability that sequences i and j are identical at site x and that i and k are identical at site y; and finally, the probability that sequences i and j are identical at site x and sequences k and l are identical at site y. Note that for finite sample sizes, the possibility that i, j, k, and l are not all distinct has to be taken into account (Hudson 1985). Also note that the identity-coefficient approach of Sved (1971) is quite different from that presented here, because he implicitly assumes that allele frequencies remain constant over time.

Identity coefficients in a genealogical context: We consider the identity coefficients in (3) for the case where both sites are polymorphic and each polymorphism is the result of a single mutation [single-nucleotide polymorphisms (SNPs)]. When there are just two alleles at both loci, the square of the disequilibrium coefficient is independent of how alleles are defined; hence we consider the identity coefficients between the derived mutations (denoted by an asterisk). The identity coefficient $F^*_{x(ij)y(kl)}$ can be expressed as the expectation of the probability that the mutations occur in the portion of the genealogy ancestral to sequences i and j at site x and ancestral to k and l at site y, divided by the probability that one mutation occurs at each site. Assuming the mutation rate per base pair per generation, μ, is the same at both sites,

$$F^*_{x(ij)y(kl)} = \frac{E[I^m_{x(ij)}I^m_{y(kl)}e^{-\mu(T_x+T_y)}]}{E[T_xT_ye^{-\mu(T_x+T_y)}]}, \tag{4}$$

where $I^m_{x(ij)}$ is the branch length (in generations) leading from the most recent common ancestor (MRCA) of sequences i and j to the MRCA of the entire sample and $E[T_xT_y]$ is the expected product of the total tree length at sites x and y. The mutation rate is a nuisance parameter that can be eliminated by taking the limit as $\mu \to 0$ (Nielsen 2000).

$$F^*_{x(ij)y(kl)} = \frac{E[I^m_{x(ij)}I^m_{y(kl)}]}{E[T_xT_y]}. \tag{5}$$

By writing

$$I^m_{x(ij)} = T^m_x - t_{x(ij)} \tag{6}$$

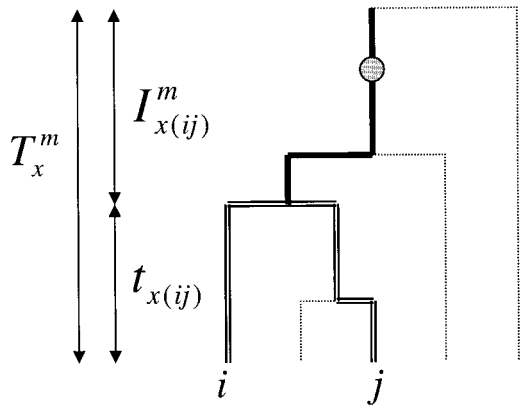

Figure 1.—Statistics of the genealogy.

(Figure 1), where $t_{x(ij)}$ is the coalescence time for sequences i and j at site x, and T^m_x is the time until the MRCA for the entire sample at site x, it can be shown that

$$\mathrm{Var}(D) = \frac{\mathrm{Cov}[t_{x(ij)}, t_{y(ij)}] - 2\,\mathrm{Cov}[t_{x(ij)}, t_{y(ik)}] + \mathrm{Cov}[t_{x(ij)}, t_{y(kl)}]}{E[T_xT_y]}. \tag{7}$$

We can use a similar procedure to find the denominator in Equation 2. The expectation $E[f_{A(x)}(1 - f_{A(x)})f_{B(y)}(1 - f_{B(y)})]$ for the case of SNPs can be expressed as the expected probability that two alleles drawn with replacement will be different at the x locus and another two drawn with replacement will be different at the y locus. Taking the limit as $\mu \to 0$,

$$\begin{aligned}E[f_{A(x)}(1 - f_{A(x)})f_{B(y)}(1 - f_{B(y)})] &= \lim_{\mu\to 0}\frac{E[t_{x(ij)}t_{y(kl)}e^{-\mu(T_x+T_y)}]}{E[T_xT_ye^{-\mu(T_x+T_y)}]} \\ &= \frac{E[t]^2 + \mathrm{Cov}[t_{x(ij)}, t_{y(kl)}]}{E[T_xT_y]}, \end{aligned} \tag{8}$$

where $E[t]$ is the expected coalescence time for a pair of chromosomes. Combining Equations 7 and 8 gives an expression for σ^2_d:

$$\sigma^2_d = \frac{\mathrm{Cov}[t_{x(ij)}, t_{y(ij)}] - 2\,\mathrm{Cov}[t_{x(ij)}, t_{y(ik)}] + \mathrm{Cov}[t_{x(ij)}, t_{y(kl)}]}{E[t]^2 + \mathrm{Cov}[t_{x(ij)}, t_{y(kl)}]}. \tag{9}$$

In other words, the expected linkage disequilibrium as measured by the r^2 statistic can be approximated in terms of the covariance in coalescence times for pairs of sequences. For example, the middle term in the numerator of (9) is the covariance in coalescence time at site x for sequences i and j and at site y for sequences i and k; see Figure 2. More generally, the kth moment of the distribution of D will depend on the covariances in coalescence times for sets of up to k chromosomes ancestral at each site. Because no assumptions are made about the underlying demographic model in the derivation of (9), it provides a general way of describing the

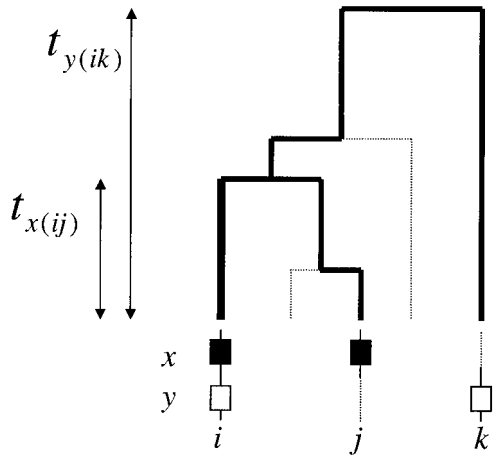

FIGURE 2.—Cov[$t_{x(ij)}$, $t_{y(ik)}$] measures the covariance in coalescence time at site x for chromosomes i and j and site y for chromosomes i and k.

relationship between linkage disequilibrium and aspects of the underlying genealogy.

For finite sample size, a modification is required to include the possibility that i, j, k, and l are not all distinct,

$$\sigma_d^2 = \frac{[n^2 - 2(n-1)]C_{ij,ij} - 2(n-2)^2 C_{ij,ik} + (n-2)(n-3)C_{ij,kl} + nE[t]^2}{n(n-1)E[t]^2 + 2C_{ij,ij} + 4(n-2)C_{ij,ik} + (n-2)(n-3)C_{ij,kl}} \quad (10)$$

(following HUDSON 1985), where the C's are abbreviations for the covariances in the previous equations. It is also worth noting that by dividing both the numerator and denominator of (9) by the variance in coalescence time for pairs of sequences, σ_d^2 can be written in terms of correlations in coalescence times,

$$\sigma_d^2 = \frac{\rho_{ij,ij} - 2\rho_{ij,ik} + \rho_{ij,kl}}{E[t]^2/\mathrm{Var}(t) + \rho_{ij,kl}}, \quad (11)$$

where the subscripts refer to the three configurations of sample chromosomes. One advantage of writing the expression in terms of correlations rather than covariances is that correlations will be influenced largely by recombination, whereas demographic factors can strongly influence the mean and variance of coalescence times.

Conditional linkage disequilibrium: If the expectation is conditioned on the exclusion of rare mutations (those represented fewer than a times in the sample), the covariances in coalescence times in (9) have to be augmented by the covariances in times between coalescing and the first point that the lineage ancestral to the MRCA has at least a descendants in the sample. However, the magnitude of the extra terms is small, and a good approximation is obtained with a slight modification to the denominator, replacing $E[t]$ with $E[t] - E[\delta^a]$, where $E[\delta^a]$ is the expected time until an ancestral lineage has at least a descendants. In the standard coalescent $E[\delta^a] = 2(a-1)/n$ for $a < n$ (SAUNDERS *et al.* 1984).

In practice, linkage disequilibrium is typically conditioned on the exclusion of rare alleles, rather than rare mutations. However, because rare alleles typically represent rare mutations, the error introduced by conditioning on rare mutations rather than rare alleles is small. For example, among loci for which the rare allele is represented only once, the rare allele represents the rare mutation with probability $1 - 1/n$ under the standard neutral model.

LINKAGE DISEQUILIBRIUM IN THE STANDARD NEUTRAL MODEL

The expectation of (9) can be derived under the standard coalescent using the results of GRIFFITHS (1981, 1991; see also PLUZHNIKOV and DONNELLY 1996). If the sample size is sufficiently large such that all sequences picked at random from the sample are distinct, the covariances in coalescence times (in units of $2N_e$ generations) are

$$\mathrm{Cov}[t_{x(ij)}t_{y(ij)}] = \frac{18 + \rho}{18 + 13\rho + \rho^2}$$

$$\mathrm{Cov}[t_{x(ij)}t_{y(ik)}] = \frac{6}{18 + 13\rho + \rho^2}$$

$$\mathrm{Cov}[t_{x(ij)}t_{y(kl)}] = \frac{4}{18 + 13\rho + \rho^2}$$

(GRIFFITHS 1981, 1991; KAPLAN and HUDSON 1985; PLUZHNIKOV and DONNELLY 1996), where $\rho = 4N_e r$. Note these differ from the results of PLUZHNIKOV and DONNELLY (1996) by a factor of 4 as we consider just the time to the MRCA, not the total branch length leading to the MRCA. Under the standard coalescent, $E[t] = 1$, hence the ratio of the expectations (2) is

$$\sigma_d^2 = \frac{10 + \rho}{22 + 13\rho + \rho^2}. \quad (12)$$

This is the same result as given by OHTA and KIMURA (1971) and WEIR and HILL (1986) and is the expected linkage disequilibrium as the sample size tends to infinity. The modification for finite sample size (10) has a negligible effect for large n (see Equation 3 of WEIR and HILL 1986). Figure 3 shows how the value of (12) varies with the recombination rate for $n = 50$ and also how it compares to the expectation of r^2. When rare alleles are excluded, Equation 12 provides a close approximation to the expectation of r^2 (after correcting the denominator).

DISCUSSION

Interpreting linkage disequilibrium in terms of the underlying genealogy can help in understanding the behavior of linkage disequilibrium under different demographic scenarios, such a population growth (SLATKIN

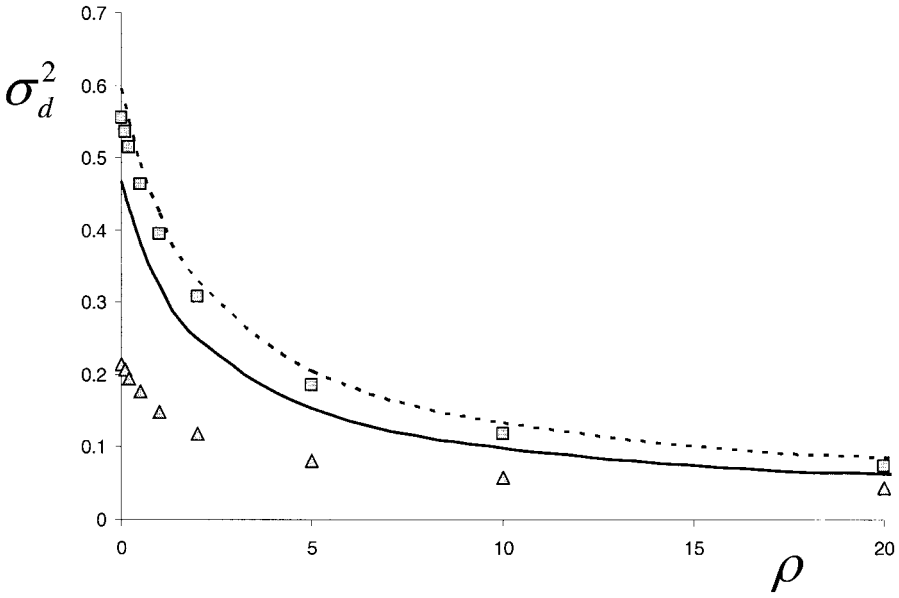

FIGURE 3.—The relationship between the scaled recombination rate, $\rho = 4N_e r$, σ_d^2 (lines), and the average value of r^2 (points) for all segregating sites (solid line and triangles) and those for which the derived mutation is present in at least 10% of samples (dotted line and squares). Values of r^2 were obtained by coalescent simulation under the standard neutral model with $n = 50$.

1994; KRUGYLAK 1999), bottlenecks (REICH *et al.* 2001), and geographical subdivision (WAKELEY *et al.* 2001).

Growing populations: The effect of population growth is to distort genealogies such that the mean coalescence is reduced relative to the case of no growth and, more importantly, the variance in coalescence times is even more reduced. The effects of population growth on the correlations in coalescence times are more subtle. Consider two genealogies, which have experienced the same number of recombination events, but one generated under a standard neutral model and one generated by a growing population model. Under high rates of growth, gene genealogies assume a star-like shape, such that the vast proportion of the total tree length is composed of external branches. So if a recombination event is thrown onto the genealogy, the probability that it occurs in the history of a randomly chosen pair of sequences from the sample approaches $2/n$. In contrast, in a constant population size, the probability that the recombination event affects the ancestry of the chosen pair is $\sim(\sum_{i=1}^{n-1} 1/i)^{-1}$, which is $>2/n$ for $n > 3$. Consequently, in growing populations fewer recombination events will influence the history of a randomly chosen pair of chromosomes from the sample, leading to higher correlations in coalescence times; see Table 1. Overall, the reduction in variance of coalescence times caused by population growth has a greater effect on linkage disequilibrium (LD) than the increase in correlations, leading to a decrease in LD.

Population bottlenecks: Recent population bottlenecks can increase linkage disequilibrium considerably, because in contrast to the case of population growth, bottlenecks affect the mean coalescence time more than the variance. If the probability that a pair of chromosomes coalesces during a recent bottleneck is ϕ, and we assume the bottleneck is instantaneous (hence chromosomes coalescing during the bottleneck have coalescence time equal to zero), the mean coalescence time is $1 - \phi$ and the variance is $1 - \phi^2$ (in units of $2N_e$ generations). So the ratio $E[t]^2/\mathrm{Var}(t)$ is reduced relative to the case of no bottleneck.

The effect of population bottlenecks on correlations in coalescence time is more complex. Bottlenecks distort gene genealogies such that the majority of the tree length occurs when there are few ancestral lineages (those that survived the bottleneck); consequently most recombination events will influence these ancestral lineages. The correlations in coalescence time are therefore increased by the probability of coalescing during the bottleneck and decreased by the effects of prebottleneck recombination. For weak bottlenecks, ancestral recombination is more important, whereas for strong bottlenecks, correlations are increased by coalescence events during the bottleneck. Table 2 shows the effects

TABLE 1

The effect of exponential population growth (rate λ) on genealogical correlations for a sample of $n = 50$ chromosomes (from 10^6 coalescent simulations)

λ	$\bar{R}$	$\bar{t}^2/\mathrm{Var}(t)$	$\rho_{ij,ij}$	$\rho_{ij,ik}$	$\rho_{ij,kl}$	$r^2_{0.1}$
0.0	10.0	1.00	0.319	0.085	0.058	0.25
1.0	10.0	1.91	0.460	0.110	0.067	0.22
5.0	10.0	3.37	0.561	0.107	0.058	0.20
10.0	10.0	4.37	0.604	0.106	0.060	0.19

The per generation recombination rate is adjusted such that the average number of recombination events in the history of the genealogy $\bar{R}$ is constant. The last column indicates the average value of r^2 between mutations for which the rare allele at both loci has a frequency of at least 0.1.

TABLE 2

The effect of recent bottlenecks of severity ϕ (the probability of coalescence during the bottleneck for a pair of chromosomes) on genealogical correlations for a sample of $n = 50$ chromosomes (from 10^6 coalescent simulations)

ϕ	$\bar{R}$	$\bar{t}^2/\mathrm{Var}(t)$	$\rho_{ij,ij}$	$\rho_{ij,ik}$	$\rho_{ij,kl}$	$r^2_{0.1}$
0.0	10.0	1.0	0.319	0.085	0.058	0.25
0.1	10.0	0.79	0.265	0.053	0.030	0.26
0.2	10.0	0.64	0.260	0.046	0.023	0.32
0.5	10.0	0.33	0.316	0.086	0.045	0.62

The last column indicates the average value of r^2 between sites at which the rarer allele has a frequency of at least 0.1.

of recent bottlenecks on the correlations in coalescence time for the same average number of recombination events in the history of the sample. Overall, the effect on the variance in coalescence times is more important than the effect on correlations, such that LD is increased by bottlenecks.

Population structure: Population structure increases linkage disequilibrium because of the correlations in coalescence times induced by coalescent events within subpopulations. This is true even for unlinked sites. Consider a two-deme model with symmetric migration between them at rate m per chromosome, per generation. Under such conditions, and assuming large n (sampled evenly from the two demes), the expected covariances in coalescence times at unlinked sites x and y are

$$\begin{aligned}\mathrm{Cov}[t_{x(ij)}, t_{y(ij)}] &= \frac{1}{4M^2}\\ \mathrm{Cov}[t_{x(ij)}, t_{y(ik)}] &= 0\\ \mathrm{Cov}[t_{x(ij)}, t_{y(kl)}] &= 0,\end{aligned}$$

where $M = 4N_e m$ (N_e is the sum of the effective population sizes for the two demes). So the ratio of expectations (2) is

$$\sigma_d^2 = \frac{1}{1 + 4M(1 + M)}. \qquad (13)$$

The implication of the result is that significant LD, even between unlinked markers, is expected in subdivided populations when the population migration rate is low ($M < 1$).

Other measures of LD: It is worth noting that another widely used statistic of linkage disequilibrium, $|D'|$ (Lewontin 1964), behaves in a very different manner to σ_d^2. This is because $|D'|$ can be less than one only if all four possible haplotypes are present for a pair of segregating sites. The expectation of $|D'|$ therefore depends on higher moments of coalescence times than the expectation of r^2.

Many thanks to Molly Przeworski, David Reich, Paul Fearnhead, Carsten Wiuf, Mikkel Schierup, Simon Myers, and two anonymous reviewers. G.M. is funded by the Royal Society.

LITERATURE CITED

Griffiths, R. C., 1981 Neutral two-locus multiple allele models with recombination. Theor. Popul. Biol. **19:** 169–186.

Griffiths, R. C., 1991 The two-locus ancestral graph, pp. 100–117 in *Selected Proceedings on the Symposium on Applied Probability* (IMS Lecture Notes, Monograph Series, Vol. 18), edited by I. V. Basawa and R. L. Taylor. Institute of Mathematical Statistics, Hayward, CA.

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

Hudson, R. R., 1985 The sampling distribution of linkage disequilibrium under an infinite allele model without selection. Genetics **109:** 611–631.

Kaplan, N., and R. R. Hudson, 1985 The use of sample genealogies for studying a selectively neutral *m*-loci model with recombination. Theor. Popul. Biol. **28:** 382–396.

Kingman, J. F. C., 1982 The coalescent. Stoch. Proc. Appl. **13:** 235–248.

Krugylak, L., 1999 Prospects for whole-genome linkage disequilibrium mapping of common disease genes. Nat. Genet. **22:** 139–144.

Lewontin, R. C., 1964 The interaction of selection and linkage. I. General considerations; heterotic models. Genetics **49:** 49–67.

Nielsen, R., 2000 Estimation of population parameters and recombination rates from single nucleotide polymorphisms. Genetics **154:** 931–942.

Ohta, T., 1982 Linkage disequilibrium with the island model. Genetics **101:** 139–155.

Ohta, T., and M. Kimura, 1971 Linkage disequilibrium between two segregating nucleotide sites under the steady flux of mutations in a finite population. Genetics **68:** 571–580.

Pluzhnikov, A., and P. Donnelly, 1996 Optimal sequencing strategies for surveying molecular genetic diversity. Genetics **144:** 1247–1262.

Reich, D. E., M. Cargill, S. Bolk, J. Ireland, P. C. Sabeti *et al.*, 2001 Linkage disequilibrium in the human genome. Nature **411:** 199–204.

Saunders, I., S. Tavaré and G. Watterson, 1984 On the genealogy of nested subsamples from a haploid population. Adv. Appl. Probab. **16:** 471–491.

Slatkin, M., 1994 Linkage disequilibrium in growing and stable populations. Genetics **137:** 331–336.

Strobeck, C., 1983 Expected linkage disequilibrium for a neutral locus linked to a chromosomal arrangement. Genetics **103:** 545–555.

Strobeck, C., and K. Morgan, 1978 The effect of intragenic recombination on the number of alleles in a finite population. Genetics **88:** 829–844.

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

Wakeley, J., R. Nielsen, S. N. Liu-Cordero and K. Ardlie, 2001 The discovery of single-nucleotide polymorphisms—and inferences about human demographic history. Am. J. Hum. Genet. **69:** 1332–1347.

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

Communicating editor: W. Stephan