How Population Growth Affects Linkage Disequilibrium

Alan R. Rogers[1]
Department of Anthropology, University of Utah, Salt Lake City, Utah 84112
ORCID ID: 0000-0003-3987-3346 (A.R.)

ABSTRACT The "LD curve" relates the linkage disequilibrium (LD) between pairs of nucleotide sites to the distance that separates them along the chromosome. The shape of this curve reflects natural selection, admixture between populations, and the history of population size. This article derives new results about the last of these effects. When a population expands in size, the LD curve grows steeper, and this effect is especially pronounced following a bottleneck in population size. When a population shrinks, the LD curve rises but remains relatively flat. As LD converges toward a new equilibrium, its time path may not be monotonic. Following an episode of growth, for example, it declines to a low value before rising toward the new equilibrium. These changes happen at different rates for different LD statistics. They are especially slow for estimates of σ_{d}^2, which therefore allow inferences about ancient population history. For the human population of Europe, these results suggest a history of population growth.

LINKAGE disequilibrium (LD) refers to the statistical association between pairs of genetic loci. It is used routinely in localizing disease genes, in detecting natural selection, and in studying population history. In all of these contexts, it is necessary to account for effects of changes in population size.

These effects arise because inhabitants of small populations tend to be close relatives. The genealogical paths that separate them are short, and this reduces the opportunity for recombination. For this reason, LD rises after a fall in population size and falls after a rise.

These effects are understood in a general way and are often studied by computer simulation (Kruglyak 1999; Pritchard and Przeworski 2001). Although this approach has led to important insights, our understanding is still rudimentary. This article uses a deterministic algorithm to explore the effects of growth, of decline, and of temporary reductions (bottlenecks) in population size. It shows that each type of history leaves a distinctive signature in the "LD curve," which relates the LD between pairs of sites to the distance that separates them along the chromosome.

The paper uses σ_{d}^2 (defined below) as a measure of LD. This choice is unusual, because σ_{d}^2 has always been of secondary importance. As we shall see, however, σ_{d}^2 has dynamical properties that give it deeper time depth than alternative measures of LD. It is readily estimated from data and can be predicted by a deterministic theory, which makes it easy to study the response to changes in population size. This article shows that σ_{d}^2 is of more than secondary importance. It is useful in its own right as a measure of LD.

Materials and Methods

Software

The various methods were evaluated against simulated data generated by MACS (Chen *et al.* 2009). Simulation results in Figure 1 are based on my own coalescent program, hetage, written in C. All other analyses were done using ldpsiz, a package of C programs available at github.com/alanrogers/ldpsiz. This package includes several programs: eld, which estimates σ_{d}^2 and r^2 from genetic data; preld, which calculates σ_{d}^2 from the history of population size; and sald, which fits history parameters to an observed LD curve, using the method of simplex-simulated annealing (Press *et al.* 1992, pp. 451–455), with modifications described by Gao and Han (2012).

Measuring LD

Consider a pair of loci (nucleotide sites), A and B. At locus A, alleles A_1 and A_0 have frequencies a and $1-a$. At locus B, alleles B_1 and B_0 have frequencies b and $1-b$. The

Copyright © 2014 by the Genetics Society of America
doi: 10.1534/genetics.114.166454
Manuscript received November 10, 2013; accepted for publication May 28, 2014; published Early Online June 6, 2014.

[1] Address for correspondence: University of Utah, 270 S 1400 E Rm 102, Salt Lake City, UT 84112. E-mail: rogers@anthro.utah.edu

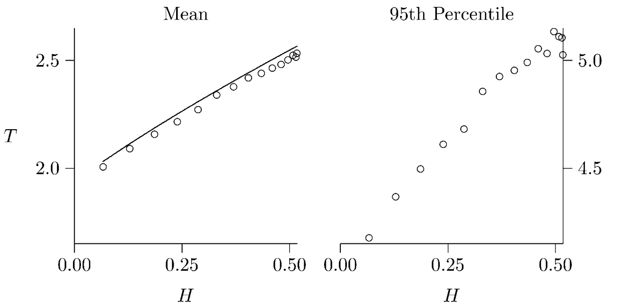

Figure 1 The depth, *T*, of gene genealogy given heterozygosity, *H*. The left and right panels show the mean and the 95th percentile in units of 2*N* generations. Solid line shows expected values, based on the model of Griffiths and Tavaré (1998, equation 1.5, p. 276). Open circles show results from coalescent simulations. These results assume a sample of 30 diploid individuals and a mutation rate of 0.02 per 2*N* generations. The slope is greater in the right panel than in the left, indicating that heterozygosity has a stronger effect on the upper tail of the distribution than on the mean.

disequilibrium coefficient, D, is defined such that $ab + D$ is the frequency of gamete type A_1B_1. The sign of D is arbitrary, depending on how one labels the alleles, so the magnitude of LD is often measured by D^2.

These measures are sensitive to heterozygosity at the two loci, so many authors prefer the squared correlation coefficient (Hill and Robertson 1968),

$$r^2 = \frac{D^2}{a(1-a)b(1-b)}. \tag{1}$$

Unfortunately, there is no consensus regarding the expected value of this statistic, even in the simplest case of neutral loci in a randomly mating population of constant size (compare Sved 2009 with Song and Song 2007 and Durrett 2008, p. 98).

Ohta and Kimura (1969, p. 233) proposed a related measure of LD, the "squared standard linkage deviation," which was also motivated by a desire to minimize the effect of heterozygosity:

$$\sigma_\mathrm{d}^2 = \frac{E[D^2]}{E[a(1-a)b(1-b)]}. \tag{2}$$

This measure is usually viewed as an approximation to $E[r^2]$. It is most useful in this role when the population size is large and constant and both loci have appreciable heterozygosity (Hudson 1985). In other situations, the two parameters can differ greatly. But even when σ_d^2 fails as an approximation, it is still useful as a measure of LD.

This is not to say that σ_d^2 provides a complete description of variation at a pair of loci. That is a problem with multiple dimensions, which cannot be solved by any scalar-valued measure of LD (Weir 1996, pp. 125–127). Such measures are simplifications and are necessarily incomplete (Schaper *et al.* 2012). However, as we shall see, σ_d^2 captures enough to provide an interesting window into the history of population size.

Estimation of σ_d^2

One can estimate σ_d^2 from data by replacing the expected values in Equation 2 with averages across the genome. For this purpose, I treat each polymorphic site [single-nucleotide polymorphism (SNP)] as "focal" and compare the focal SNP with each other SNP within some given range. From each comparison, I calculate D^2 and $a(1-a)b(1-b)$. These values are tabulated separately within bins based on the distance in centimorgans between the SNPs. After all comparisons have been tabulated, the estimate ($\hat{\sigma}_\mathrm{d}^2$) is calculated as the sum of D^2 within a bin divided by the corresponding sum of $a(1-a)b(1-b)$.

In this article, I ignore pairs of SNPs separated by >0.3 cM, because there is little LD at these distances in most parts of the human genome.

To estimate uncertainties, I use a moving-blocks bootstrap (Lieu and Singh 1992), with 300 SNPs per block. With diploid data, some genotypes may be unphased. In such cases, I use the EM algorithm to estimate D^2, as explained in the *Appendix*.

Deterministic evolution of σ_d^2

Under neutral evolution with constant population size, σ_d^2 converges to an equilibrium value (Ohta and Kimura 1971; McVean 2002). In addition, several authors have introduced recurrence equations, which make it possible to study the transient behavior of σ_d^2 after changes in population size (Weir and Cockerham 1974; Hill 1975; Strobeck and Morgan 1978). The model of Strobeck and Morgan (1978) allows faster calculations but is less numerically stable than that of Hill (1975). I present some results using the former model but focus primarily on the latter. In the *Appendix*, I summarize Hill's model and show that it holds not only under the mutational model that he studied, but also under the model of infinite sites (Kimura 1969). It is thus appropriate for use with DNA sequence data.

For bootstrap confidence intervals, I accelerate these calculations by using Equation A2 of the *Appendix* to approximate a system of ordinary differential equations, which is then solved using standard software.

Sampling bias

Drawing a sample is equivalent to one generation of evolution with a very small population—one whose size equals that of the sample. Because drift introduces LD, there is much more LD in a sample than in the population from which it was drawn.

Hudson (1985, Equation 6) showed that sampling bias in r^2 equals $1/n$, when the sample consists of n gametes. Sampling bias also inflates the value of $\hat{\sigma}_\mathrm{d}^2$. The program preld includes this bias in calculations of the expected value, σ_d^2. These calculations use the model of Hill (1975) or Strobeck and Morgan (1978) to project the state vector forward one generation, with the population size set equal to the sample size.

Fitting population history to an observed LD curve

The program eld (a part of ldpsiz) estimates the LD curve from genetic data. A second program, sald, can then be used to search for the population history parameters that minimize the sum of squared differences between observed and predicted LD curves. This minimization problem searches a complex surface, with many local minima. To search for the global minimum, sald uses the method of simplex-simulated annealing. For each data set, sald starts 10 simulated annealing jobs on parallel threads. Typically, several of these converge to the same global minimum. sald returns the best of the 10 solutions as fitted parameters.

Results

$\hat{\sigma}_d^2$ is average r^2, weighted by heterozygosities

Equation 2 is equivalent to

$$\sigma_{\mathrm{d}}^2 = \frac{E\left[H_A H_B r^2\right]}{E[H_A H_B]}, \tag{3}$$

where $H_A = 2a(1 - a)$ is the expected heterozygosity at locus A and H_B is that at locus B. This implies that σ_{d}^2 is the expectation of r^2 weighted by the product of the two heterozygosities.

This weighting also carries over to the estimate, $\hat{\sigma}_{\mathrm{d}}^2$, which is obtained by replacing expectations with averages. Such estimates will be insensitive to loci with low heterozygosity and, for this reason, also insensitive to sequencing error. They should be useful with low-coverage sequence data.

Loci with high heterozygosity have deep gene trees

Weighting by heterozygosity exaggerates the influence of unusually deep gene trees. At some given nucleotide position, let T represent the age of the last common ancestor of the sample. I call this the "depth" of the gene tree for that nucleotide position. Griffiths and Tavaré (1998, Equation 1.5, p. 276) derive the conditionally expected depth, $E[T \mid x, n]$, given that x copies of the derived allele were observed in a sample of haploid size n. Given the heterozygosity, we can solve for x, but we cannot tell the derived from the ancestral allele. It may be present in x copies or in $n - x$. At mutation–drift equilibrium, however, these alternatives have probabilities proportional to $1/x$ and $1/(n - x)$ (Fu 1995, equation 1). Using these values as weights, I average $E[T \mid x, n]$ and $E[T \mid n - x, n]$ to calculate expected tree depth given heterozygosity. The results are shown as a solid line in Figure 1. As Figure 1 shows, tree depth increases with heterozygosity. Simulated values—shown as open circles—agree closely with the theory.

Simulated values also allow us to examine the upper tail of the distribution, as seen in Figure 1, right. The 95th percentile of tree depth increases with heterozygosity even more steeply than does the mean. Thus, heterozygosity has an exaggerated effect on the upper tail of the distribution.

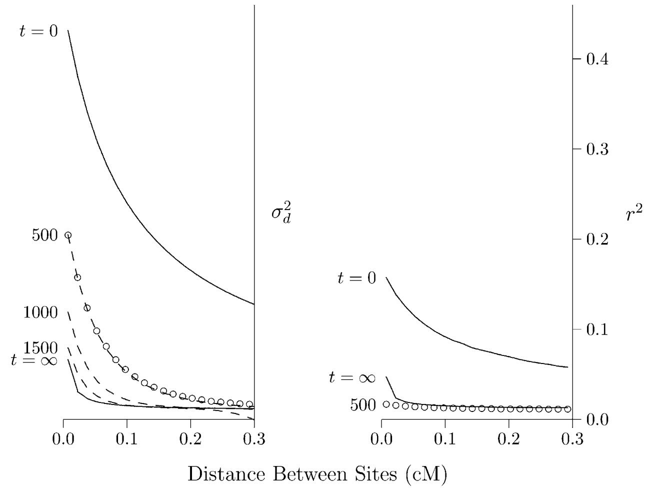

Figure 2 Effect of population expansion on the LD curve. The population grew suddenly at time t = 0 from $2N$ = 10^3 to 10^5. Left panel: LD is measured by σ_{d}^2. Solid lines show the predicted values at the initial equilibrium (t = 0) and at the eventual equilibrium (t = ∞). Dashed lines show a series of transient states that occur at various values of t, the number of generations since the expansion. These lines are all calculated using the method of Hill (1975). Open circles show values simulated using MACS for t = 500. Right panel: LD is measured by r^2, and points and lines are based on computer simulation. Calculations assume that $u = 1.48 \times 10^{-8}$ per site per generation, and the haploid sample size is 100.

Geneticists often study loci selected for their high heterozygosity (Lewontin 1967; Rogers and Jorde 1996; Clark *et al.* 2005). A sample of loci selected in this fashion will have gene trees that are unusually deep, especially in the upper tail of the distribution.

Because σ_{d}^2 is weighted by heterozygosity, it exaggerates the influence of these deep gene trees. For this reason, it should be sensitive to earlier portions of a population's history. Following a change in population size, it should converge more slowly to the new equilibrium.

Effect of population growth

LD will decline following an expansion of population size, because genetic drift is weaker in large populations and produces less LD. This process is illustrated in Figure 2, which shows the effect of an expansion from size $2N = 10^3$ to $2N = 10^5$. The LD curve of the initial population (labeled $t = 0$) represents an equilibrium between mutation, drift, and recombination. After the expansion, the LD curve will eventually converge to a new equilibrium, which is labeled $t = \infty$. This new equilibrium, however, is reached only gradually.

LD is measured by σ_{d}^2 in Figure 2, left, and by r^2 in Figure 2, right. The dashed lines were calculated using the method of Hill (1975), and the open circles were estimated from data simulated using MACS (Chen *et al.* 2009). Note that σ_{d}^2 is still far from equilibrium at generation 500 and does not approach equilibrium until about generation 1500.

The situation is very different in Figure 2, right, where LD is measured by r^2. By generation 500, r^2 is close to equilibrium. At the left edge of the graph, it is slightly *below* the equilibrium. We return to this point later, but for the moment, note simply that r^2 converges much faster than σ_d^2. For this reason, the two measures are useful for studying different timescales. We learn from r^2 about the recent past and from σ_d^2 about more ancient events. Presumably, this difference arises because σ_d^2 is weighted toward loci with high heterozygosity. As shown in Figure 1, this weighting makes it more sensitive to the distant past.

Returning to Figure 2, left, note that the right portion of the curve converges faster than the left. This is because the postexpansion population is large, and genetic drift is weak. The dynamics are therefore dominated by recombination, which is stronger on the right side of the graph. The result is that midway through the process—say at generation 500—the LD curve declines very steeply. Thus, a steeply declining LD curve is the signature of recent population growth (Kruglyak 1999; Pritchard and Przeworski 2001).

The method of Hayes et al. (2003)

Figure 2 suggests that σ_d^2 is likely to be of greater use than r^2 in inferring the ancient history of population size. Yet the latter statistic is often used instead (Hayes *et al.* 2003; McEvoy *et al.* 2011; Tenesa *et al.* 2007), using a method that is based on the formula $E[r^2] \approx \rho$, where

$$\rho = \frac{1}{1 + 4Nc}. \tag{4}$$

Here, N is diploid population size, and c is the recombination rate (Sved 1971; Sved and Feldman 1973). Although this formula has been criticized (Littler 1973, p. 272; McVean 2002, pp. 987–988; McVean 2007, p. 923; Durrett 2008, p. 98), it is widely used.

Hayes *et al.* (2003) study a generalization of r^2 and argue from simulations that the expectation of this statistic is $\approx\rho$. Furthermore, this approximation even works for populations that have increased in size at a constant linear rate, provided that one interprets N as the population size that obtained $1/2c$ generations in the past. By inverting Equation 4, they are able to estimate N over a wide range of values of c, which correspond under their model to different points in the past.

Tenesa *et al.* (2007) work directly with r^2, but estimate the history of population size in the same way. They also employ a modified predictor,

$$\tilde{\rho} = \frac{1}{2 + 4Nc}, \tag{5}$$

which in their view accounts better for the effect of mutation. This method is also used by McEvoy *et al.* (2011). There are two difficulties with this formula. First, Tenesa *et al.* (2007, p. 521) derive it from a formula of Hill (1975, p. 124), which refers not to $E[r^2]$ but to σ_d^2. Thus, it approximates $E[r^2]$ only to the extent that σ_d^2 does. In addition, it uses a fairly crude standard of approximation, taking 10 as $\approx$11.

The arguments justifying both of these methods assume that population size has increased linearly. Of course, no population can increase linearly forever, so the period of linear increase must end at some point in the past. Furthermore, the method's results are often interpreted in terms of nonlinear growth trajectories, such as bottlenecks (Hayes *et al.* 2003, pp. 639–670; Tenesa *et al.* 2007, p. 525; McEvoy *et al.* 2011, p. 822). Let us consider how the method behaves in this broader context.

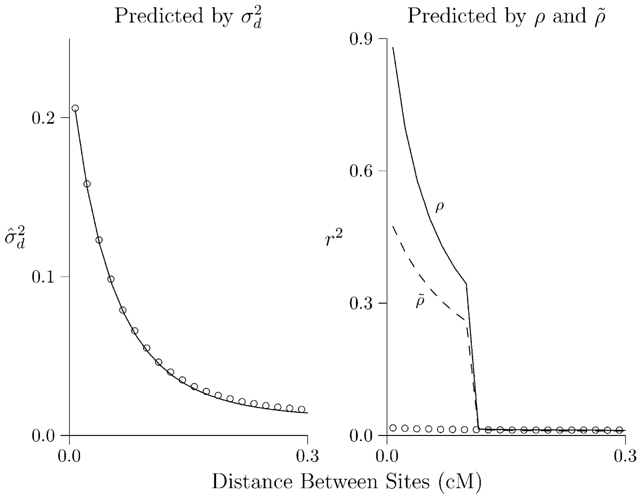

Figure 3 Predicted and simulated LD. Each panel shows the same population 500 generations after expansion from $2N = 10^3$ to 10^5. Lines show values predicted by σ_d^2 [left panel (Hill 1975)] and by ρ (right panel, solid line) and $\tilde{\rho}$ (right panel, dashed line). Open circles show results estimated from simulations. Calculations assume that $u = 1.48 \times 10^{-8}$ per site per generation, and the haploid sample size is 100. Simulations involve 10^9 bp DNA.

Figure 3 considers a population that expands suddenly in size and is observed 500 generations later. In both panels of Figure 3, the solid lines show predicted LD. In Figure 3, left, LD is measured by $\hat{\sigma}_d^2$ and predicted by σ_d^2 (Hill 1975). In Figure 3, right, it is measured by r^2 and predicted by ρ and $\tilde{\rho}$. The open circles show values simulated using MACS (Chen *et al.* 2009). In the context of this population history, σ_d^2 provides excellent predictions of $\hat{\sigma}_d^2$, but neither ρ nor $\tilde{\rho}$ predicts r^2.

Presumably, this reflects the nonlinearity of the growth trajectory assumed in Figure 3. We expect better accuracy in Figure 4, which evaluates ρ and $\tilde{\rho}$ against a history involving 300 generations of linear population growth, approximated as a series of 20 steps. The period of linear growth corresponds to the region to the right of the dashed line in Figure 4, right. At least in this region, predicted and simulated values should match. Yet as before, σ_d^2 predicts well, but ρ and $\tilde{\rho}$ do not.

Finally, Figure 5 evaluates ρ, $\tilde{\rho}$, and σ_d^2 against a history of constant population size. In this case, we get an accurate prediction of $\hat{\sigma}_d^2$ from σ_d^2 and a fairly accurate prediction from $\tilde{\rho}$. This makes sense, because (as mentioned above) $\tilde{\rho}$ is an approximation to the equilibrium formula for σ_d^2. On the other hand, neither ρ nor $\tilde{\rho}$ provides a useful prediction of r^2.

In these analyses, the observed and predicted values of r^2 are distressingly far apart. Perhaps this discrepancy reflects an error in the simulation software (MACS) or in my own

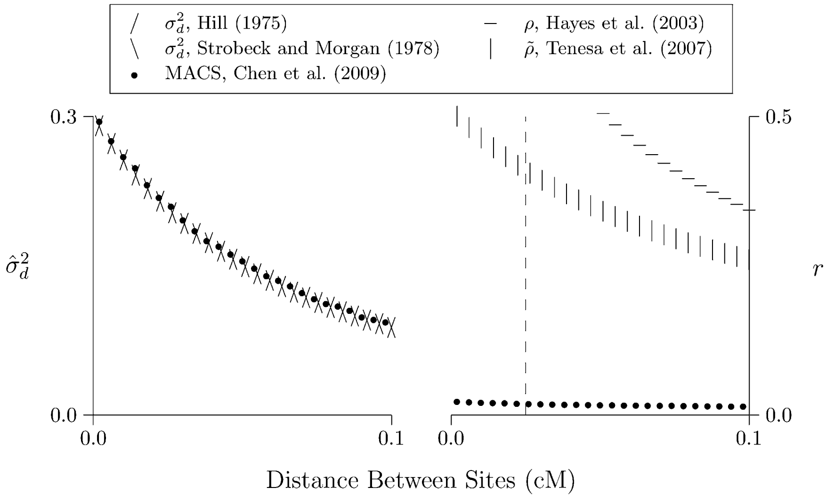

Figure 4 Linkage disequilibrium ($\hat{\sigma}_{\rm d}^2$ or r^2) after an episode of linear growth. The population begins at mutation-drift equilibrium with $2N = 10^3$ and then grows linearly to $2N = 10^6$ over a period of 300 generations. Linear growth is approximated by 20 epochs of 15 generations, within each of which $2N$ is constant. In the right panel, the region to the right of the vertical line corresponds to the period of linear growth, according to the model of Hayes *et al.* (2003). All calculations assume that $u = 1.48 \times 10^{-8}$ and $c = 10^{-8}$ per nucleotide. Simulations involve 10^9 bp of DNA, which are sequenced in a sample of 100 homologous chromosomes.

code for estimating r^2 from simulated data. To find out, I used this software to replicate the published results of Hudson (1985). The results, shown in Figure 6, are reassuring. Given the same parameter values, the code used here produces results indistinguishable from those of Hudson.

It seems fair, therefore, to conclude that the problem lies in the predictors, ρ and $\tilde{\rho}$, neither of which is useful in the cases studied here. It seems unlikely that they will be useful as a basis for inference in real populations.

Effect of population collapse

A collapse in population size produces an effect on the LD curve very different from that of population growth. This is shown in Figure 7, which illustrates two important points. First, the whole process is over very quickly. Even with $\sigma_{\rm d}^2$, we cannot look very far into the past. Second, the transient curves (the dashed lines) are quite flat. As time passes, the initial curve (labeled $t = 0$) is elevated without much change in shape. Presumably, this is because the effect of drift is dominant in a small population, so differences in recombination rate do not matter much. The result is that the LD curve becomes high but flat after a collapse in population size.

This pattern is superficially similar to that generated by gene conversion, which reduces LD between closely linked sites (Ardlie *et al.* 2001). It seems unlikely, however, that these effects will be confused. The effect of gene conversion is probably minor for sites separated by >1 kb (Frisse *et al.* 2001, p. 838; Chen *et al.* 2007, p. 764), whereas that of a population collapse extends much farther.

Effect of a bottleneck

Figure 8 shows the effect of a 100-generation bottleneck in population size. Population size was $2N = 10^3$ during the bottleneck but 10^5 before and after. The bottleneck ended at time 0, and we observe it in various subsequent generations. In the graph, the curve labeled $t = 0$ is at the end of the bottleneck and exhibits an LD curve that is high but flat, for reasons just discussed. As time passes, the right portion of the curve (the portion with high rates of recombination) falls much faster than the left, so that by generation 1000, the curve is almost L shaped. This is the signature of a bottleneck in population size.

Note the dramatic difference between Figure 2 and Figure 8—between expansion from equilibrium and from a bottleneck. Both curves are declining toward the same equilibrium, but the left portion of the curve declines much more slowly after a bottleneck than after expansion from equilibrium. This can only reflect the state of the population just before the increase in size. The initial population of Figure 2 had been small much longer than that of Figure 8. Consequently, it had less heterozygosity. As we see in the next section, this accelerates the rate of decline in LD between closely linked sites.

It is sometimes suggested that bottlenecks inflate long-range LD, just the opposite of the pattern seen above (Slatkin 2008, pp. 481–482; Tenaillon *et al.* 2008). This discrepancy, however, evaporates on close inspection. When

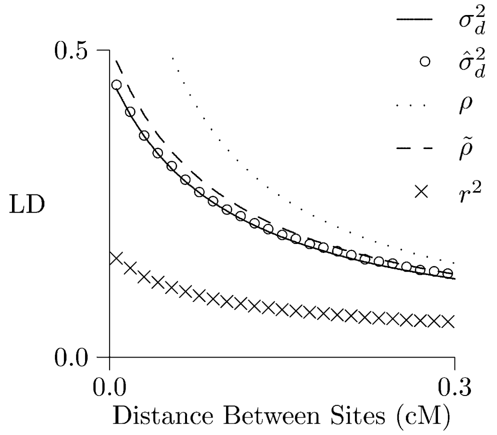

Figure 5 LD curve under constant population size. Simulated values of r^2 and $\hat{\sigma}_{\rm d}^2$ were generated using MACS (Chen *et al.* 2009). $\sigma_{\rm d}^2$ was calculated using the method of Hill (1975), ρ was calculated from Equation 4, and $\tilde{\rho}$ from Equation 5. All calculations assume that $2N = 1000$, $u = 1.48 \times 10^{-8}$, and $c = 10^{-8}$ per nucleotide. Simulations involve 10^9 bp of DNA, which are sequenced in a sample of 100 homologous chromosomes.

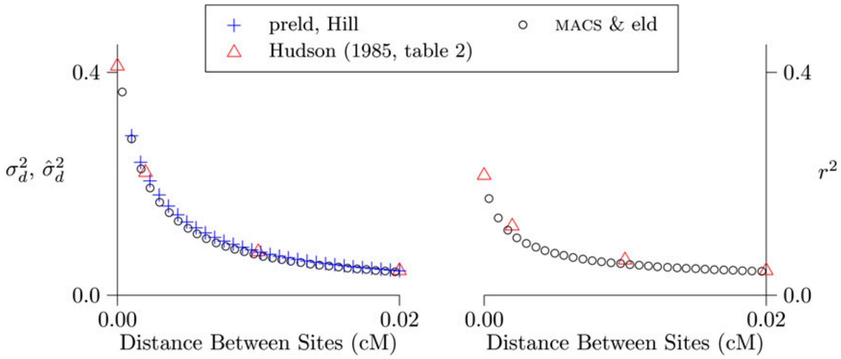

Figure 6 Replication of results of Hudson (1985, table 2). The left panel compares methods for calculating σ_d^2; the right panel compares methods for calculating r^2. The programs MACS, eld, and preld are described in *Materials and Methods*. Parameter values: $\theta = 0.1$, $2N = 50{,}000$, $n = 50$. MACS simulated chromosomes 2 Mb in length. Following Hudson, I correct $\hat{\sigma}_\mathrm{d}^2$ but not r^2 for sampling bias.

Tenaillon *et al.* (2008, figure 6) discuss “long-range” LD, they are referring to pairs of nucleotide sites separated by ~0.01 cM. In my own analysis, such pairs would be called “tightly linked” and would fall at the left edge of the LD curve. In this region of Figure 8, LD is indeed elevated, in agreement with Tenaillon *et al.* (2008). In another study, Thornton and Andolfatto (2006, p. 1611) report elevated LD following a bottleneck. But as they do not discuss the shape of the curve, there is no conflict between their results and mine.

The time paths of individual points on the LD curve

Figure 9 shows the time path of σ_d^2 after an expansion in population size. Because the new population is larger, the new equilibrium will have less LD. But σ_d^2 does not decline monotonically toward this new equilibrium. Instead, it falls rapidly to a smaller value before climbing back slowly toward to the new equilibrium.

The initial decline in σ_d^2 does not result from any decline in its numerator, $E[D^2]$. Indeed, Figure 10 shows that the numerator actually grows. The decline in σ_d^2 happens because growth in its numerator is outstripped by that in its denominator, $E[a(1-a)b(1-b)]$, which increases under the influence of mutation. The proportional increase is large, because our initial population was small and therefore had little heterozygosity. For this reason, the denominator was initially so small that increments caused by mutation had a large proportional effect.

Two factors account for the postexpansion growth in the numerator, $E[D^2]$, of σ_d^2. First, D is proportional to heterozygosity (Kaplan and Weir 1992, p. 334), which increases in response to mutation. Ordinarily, the positive effect on D^2 would be offset by the negative effect of recombination. But in the early generations following our expansion, this negative effect is very weak. This is because the initial population had low heterozygosity. Few recombinant gametes are produced in such a population, because such gametes are produced only by double heterozygotes.

The nonmonotone time path in Figure 9 refers to σ_d^2, but it seems plausible that r^2 might obey similar dynamics. This may explain why, in Figure 2, right, the left end of the curve for $t = 500$ is below the equilibrium.

This also explains why the left portion of the LD curve declines faster after expansion from equilibrium than after a bottleneck. The left portion of the curve refers to tightly linked sites, with weak recombination. In the postexpansion population, genetic drift is also weak because the population is large. Because recombination and drift are both weak, mutation dominates the dynamics. The proportional effect of mutation is large when heterozygosity is low. In the case of expansion from equilibrium, the population has been small a long time, so heterozygosity is low, the proportional effect of mutation is large, and σ_d^2 declines rapidly. Heterozygosity is not so low at the end of a bottleneck, because the population has been small only briefly. Consequently, the proportional effect of mutation is smaller, and σ_d^2 declines more slowly.

LD in the human population of Europe

The methods discussed above generate predictions about LD from assumptions about the history of population size. It is also possible, by fitting predicted to observed curves, to work in the other direction—from data to inferences about

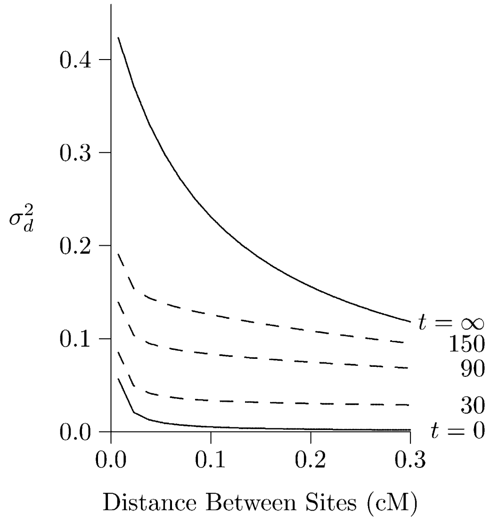

Figure 7 Effect of population collapse on the LD curve. At time $t = 0$, this population collapsed in size from $2N = 10^5$ to 10^3. Calculations use the method of Hill (1975), with $u = 1.48 \times 10^{-8}$. No correction for sampling bias was needed, as these calculations are not compared with simulations.

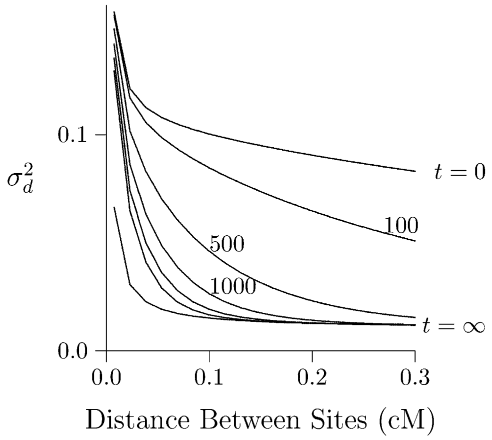

Figure 8 Effect of a 100-generation bottleneck in population size. The curves show σ_d^2 at various points after recovery from a 100-generation bottleneck, during which the population had size $2N = 10^3$. Before and after the bottleneck, its size was 10^5. Other details are as in Figure 7.

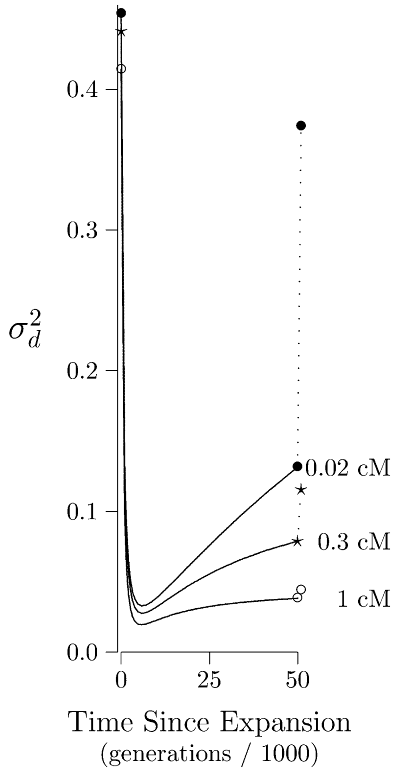

Figure 9 The time path of σ_d^2 after a population expansion. The initial population was small ($2N = 10^3$) and at mutation–drift equilibrium. At time 0 it expands suddenly to a much larger size ($2N = 10^5$). Each curve shows the time path of σ_d^2 for a different rate of recombination, measured in centimorgans. There is a curve for tight linkage (•, 0.02 cM), one for somewhat weaker linkage (⋆, 0.3 cM), and one for linkage that it weaker still (°, 1 cM). Dotted lines connect the final value of each curve to its ultimate equilibrium, which would be reached if the population stayed large forever.

population history. This is a problem in statistical inference, and methods for this purpose will be described elsewhere.

The present methods, however, are useful in exploratory data analysis, as shown in Figure 11. The open circles show $\hat{\sigma}_\mathrm{d}^2$, as estimated from chromosome 1 in European data (1000 Genomes Project Consortium 2012). These are surrounded by dashed lines, which indicate a 95% confidence region generated by moving-blocks bootstrap. These show that chromosome 1 provides accurate estimates of σ_d^2. Hill (1981) expressed skepticism about the possibility of estimating population size from data on LD. At that time only limited data were available, and it was not possible to estimate LD statistics with great accuracy. This inaccuracy bled into estimates of population size. The narrow confidence region in Figure 11 shows that things have changed.

The solid line in Figure 11 shows the equilibrium curve that fits these data best—the one that minimizes squared errors between observed and predicted values of $\hat{\sigma}_\mathrm{d}^2$. This is the curve for a population of constant size, $2N = 8621$. In Figure 11, right, we see the differences between observed and fitted values. Because these differences are positive on the left but negative on the right, it is clear that the observed LD curve declines more steeply than any equilibrium curve. As we have seen, this suggests a history of population expansion in Europe, in agreement with many other analyses of European data.

Discussion

σ_d^2 has never been valued for its own sake. It is seen instead as an approximation to the quantity of real interest—the expected value of r^2. Yet it is often a poor approximation, even in populations of constant size (Maruyama 1982; Hudson 1985, pp. 616–617). It is even worse when population size varies.

Following a change in population size, the mean of r^2 converges toward its new equilibrium far faster than does σ_d^2. Presumably, this is because σ_d^2 is sensitive to loci with high heterozygosity, and the gene trees of such loci are deep. Whatever the cause, this difference in rates of convergence has two effects. First, it makes σ_d^2 useless as an approximation to r^2 in populations that have varied in size. On the other hand, it also means that σ_d^2 itself provides a deep record of demographic history.

To take advantage of this extended record, one must estimate σ_d^2 directly from data. This is easy to do, by using averages in place of the expectations in Equation 2. With genome-scale data, such estimates are quite accurate. This provides a measure of LD that is unique in that one can easily calculate expected LD from the history of population size. With other measures, such inferences would require extensive computer simulations.

Following a population expansion, drift is weak and the dynamics of LD are dominated by recombination. The LD curve begins to decline, and this decline is fastest in the right-hand portion of the curve, where recombination rates are highest. Consequently, the curve will be unusually steep for hundreds of generations following a population expansion.

Following a population collapse, drift becomes strong and dominates the dynamics. It pushes LD upward rapidly

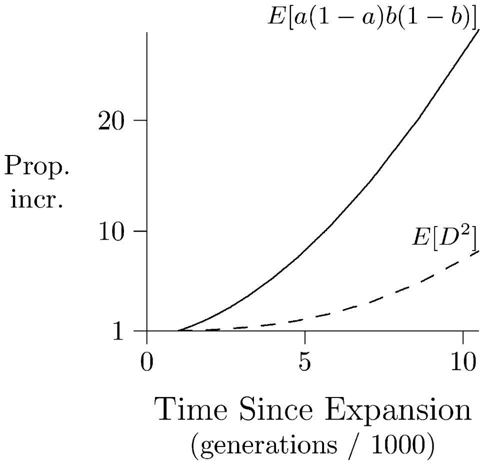

Figure 10 The time path of the numerator and denominator of σ_d^2 following a population expansion. Dashed line shows numerator of σ_d^2, relative to its value in generation 1000. Solid line shows the denominator in the same way. Loci are separated by 0.02 cM. Population history is as in Figure 9.

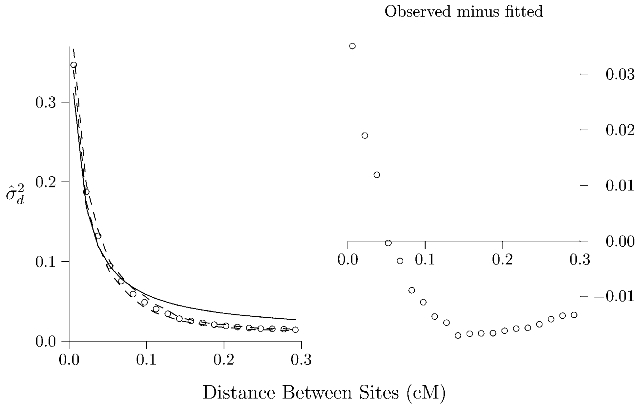

Figure 11 LD curve for chromosome 1 in Europe. In the left panel, open circles show the estimated values of σ_d^2, and dashed lines show a 95% bootstrap confidence region, generated by moving-blocks bootstrap. Solid line shows the equilibrium curve that minimizes squared errors between observed and expected values. In the right panel, differences between observed and fitted values are shown. All analyses are based on a sample of 120 chromosomes. Source: CEU data set from the 1000-Genomes project (1000 Genomes Project Consortium 2012). Genetic map data, downloaded from the 1000-Genomes website, were estimated from phased haplotypes in HapMap Release 22 (NCBI 36) (HapMap 2007).

and relatively uniformly throughout the curve. Thus, the LD curve becomes high and flat. This pattern has been reported for several human populations, including high-latitude foragers (Kaessmann *et al.* 2002) and Ashkenazi Jews (Shifman and Darvasi 2001). Yet as Pritchard and Przeworski (2001, p. 7) observe, it has been difficult to explain:

> We have several examples in which large regions exhibit more LD than would be expected under either a model of constant population size or a model with rapid population growth. Yet, at the same time studies of polymorphism at a small scale reveal less LD than would be expected. These observations at different scales are hard to accommodate in a single explanation since factors that increase long-distance LD will tend to have an even larger effect on closely linked sites.

As we have seen, this pattern arises naturally from a recent reduction in population size.

A bottleneck in population size begins with an episode of population collapse. At the end of the bottleneck, the LD curve is therefore high but flat. As population size rises at the end of the bottleneck, genetic drift becomes weak and recombination grows in importance. The right portion of the LD curve falls faster than the left, so the curve becomes steeper.

In the left portion of the curve, recombination is weak. Genetic drift is also weak, because of the increase in population size. This allows mutation to play an important role, and its effect distinguishes the two forms of expansion: from equilibrium and from a bottleneck. In the former case, the preexpansion population had little heterozygosity, so each mutational increment has a large proportional effect on the denominator of σ_d^2. After a bottleneck, the opposite is true, so the left portion of the LD curve declines more slowly. The curve becomes even steeper after a bottleneck than after an expansion from equilibrium.

In summary, (1) when recombination is strong relative to drift, LD declines and the curve becomes steeper; (2) when drift is strong relative to recombination, LD rises but the curve stays flat; and (3) where drift and recombination are both weak, the rate of decline in LD decreases with heterogyzosity.

In the human population of Europe, the LD curve is steep, suggesting a history of population expansion. This might reflect the spread of modern humans into Europe, the spread of farmers during the Neolithic, or the spread of Indo-European speakers.

Acknowledgments

I am grateful for comments from R. Bohlender, E. Cashdan, R. Gutenkunst, H. Harpending, W. G. Hill, and J. Seger and for access to computer facilities in the laboratories of Lynn Jorde and Henry Harpending.

Literature Cited

Ardlie, K., S. Liu-Cordero, M. Eberle, M. Daly, J. Barrett *et al.*, 2001 Lower-than-expected linkage disequilibrium between tightly linked markers in humans suggests a role for gene conversion. Am. J. Hum. Genet. 69: 582–589.

Chen, G. K., P. Marjoram, and J. D. Wall, 2009 Fast and flexible simulation of DNA sequence data. Genome Res. 19: 136–142.

Chen, J.-M., D. N. Cooper, N. Chuzhanova, C. Férec, and G. P. Patrinos, 2007 Gene conversion: mechanisms, evolution and human disease. Nat. Rev. Genet. 8: 762–775.

Clark, A. G., M. J. Hubisz, C. D. Bustamante, S. H. Williamson, and R. Nielsen, 2005 Ascertainment bias in studies of human genome-wide polymorphism. Genome Res. 15: 1496–1502.

Dempster, A. P., N. M. Laird, and D. B. Rubin, 1977 Maximum likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc. B 39: 1–38.

Durrett, R., 2008 *Probability Models for DNA Sequence Evolution*, Ed. 2. Springer-Verlag, New York.

Frisse, L., R. Hudson, A. Bartoszewicz, J. Wall, J. Donfack *et al.*, 2001 Gene conversion and different population histories may explain the contrast between polymorphism and linkage disequilibrium levels. Am. J. Hum. Genet. 69: 831–843.

Fu, Y., 1995 Statistical properties of segregating sites. Theor. Popul. Biol. 48: 172–197.

Gao, F., and L. Han, 2012 Implementing the Nelder-Mead simplex algorithm with adaptive parameters. Comput. Optim. Appl. 51: 259–277.

Griffiths, R., and S. Tavaré, 1998 The age of a mutation in a general coalescent tree. Stoch. Models 14: 273–295.

Hamming, R. W., 1973 *Numerical Methods for Scientists and Engineers*, Ed. 2. Dover, New York.

HapMap, 2007 A second generation human haplotype map of over 3.1 million SNPs. Nature 449: 851–862.

Hayes, B., P. Visscher, H. McPartlan, and M. Goddard, 2003 Novel multilocus measure of linkage disequilibrium to estimate past effective population size. Genome Res. 13: 635–643.

Hill, W. G., 1975 Linkage disequilibrium among multiple neutral alleles produced by mutation in finite population. Theor. Popul. Biol. 8: 117–126.

Hill, W. G., 1981 Estimation of effective population size from data on linkage disequilibrium. Genet. Res. 38: 209–216.

Hill, W. G., and A. 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.

Kaessmann, H., S. Zöllner, A. C. Gustafsson, V. Wiebe, M. Laan *et al.*, 2002 Extensive linkage disequilibrium in small human populations in Eurasia. Am. J. Hum. Genet. 70: 673–685.

Kaplan, N., and B. Weir, 1992 Expected behavior of conditional linkage disequilibrium. Am. J. Hum. Genet. 51: 333.

Kimura, M., 1969 The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutation. Genetics 61: 893–903.

Kruglyak, 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.

Lewontin, R. C., 1967 An estimate of average heterozygosity in man. Am. J. Hum. Genet. 19: 681–685.

Lieu, R. Y., and K. Singh, 1992 Moving blocks jacknife and bootstrap capture weak dependence, pp. 225–248 in *Exploring the "Limits" of the Bootstrap*, edited by R. LePage, and L. Billard. John Wiley & Sons, New York.

Littler, R., 1973 Linkage disequilibrium in two-locus, finite, random mating models without selection or mutation. Theor. Popul. Biol. 4: 259–275.

Maruyama, T., 1982 Stochastic integrals and their application to population genetics, pp. 151–166 in *Molecular Evolution*, *Protein Polymorphism*, *and the Neutral Theory*, edited by M. Kimura. Springer-Verlag, Berlin.

McEvoy, B., J. Powell, M. Goddard, and P. Visscher, 2011 Human population dispersal "out of Africa" estimated from linkage disequilibrium and allele frequencies of SNPs. Genome Res. 21: 821–829.

McVean, G., 2002 A genealogical interpretation of linkage disequilibrium. Genetics 162: 987–991.

McVean, G., 2007 Linkage disequilibrium, recombination and selection, pp. 909–944 in *Handbook of Statistical Genetics*, Ed. 3, edited by D. Balding, M. Bishop, and C. Cannings. John Wiley and Sons, Chichester, UK.

Ohta, T., and M. Kimura, 1969 Linkage disequilibrium at steady state determined by random genetic drift and recurrent mutation. Genetics 63: 229.

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.

1000 Genomes Project Consortium, 2012 An integrated map of genetic variation from 1,092 human genomes. Nature 491: 56–65.

Press, W. H., S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, 1992 *Numerical Recipes in C: The Art of Scientific Computing*, Ed. 2. Cambridge University Press, New York.

Pritchard, J., and M. Przeworski, 2001 Linkage disequilibrium in humans: models and data. Am. J. Hum. Genet. 69: 1–14.

Rogers, A. R., and C. Huff, 2009 Linkage disequilibrium between loci of unknown phase. Genetics 182: 839–844.

Rogers, A. R., and L. B. Jorde, 1996 Ascertainment bias in estimates of average heterozygosity. Am. J. Hum. Genet. 58: 1033–1041.

Schaper, E., A. Eriksson, M. Rafajlovic, S. Sagitov, and B. Mehlig, 2012 Linkage disequilibrium under recurrent bottlenecks. Genetics 190: 217–229.

Shifman, S., and A. Darvasi, 2001 The value of isolated populations. Nat. Genet. 28: 309–310.

Slatkin, M., 2008 Linkage disequilibrium—understanding the evolutionary past and mapping the medical future. Nat. Rev. Genet. 9: 477–485.

Song, Y., and J. Song, 2007 Analytic computation of the expectation of the linkage disequilibrium coefficient r^2. Theor. Popul. Biol. 71: 49–60.

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.

Sved, J. A., 2009 Correlation measures for linkage disequilibrium within and between populations. Genet. Res. 91: 183–192.

Sved, J. A., and M. W. Feldman, 1973 Correlation and probability methods for one and two loci. Theor. Popul. Biol. 4: 129–132.

Tenaillon, M., F. Austerlitz, and O. Tenaillon, 2008 Apparent mutational hotspots and long distance linkage disequilibrium resulting from a bottleneck. J. Evol. Biol. 21: 541–550.

Tenesa, A., P. Navarro, B. Hayes, D. Duffy, G. Clarke *et al.*, 2007 Recent human effective population size estimated from linkage disequilibrium. Genome Res. 17: 520–526.

Thornton, K., and P. Andolfatto, 2006 Approximate Bayesian inference reveals evidence for a recent, severe bottleneck in a Netherlands population of *Drosophila melanogaster*. Genetics 172: 1607–1619.

Weir, B., and C. Cockerham, 1974 Behavior of pairs of loci in finite monoecious populations. Theor. Popul. Biol. 6: 323–354.

Weir, B. S., 1996 *Genetic Data Analysis II: Methods for Discrete Population Genetic Data*. Sinauer Associates, Sunderland, MA.

Communicating editor: J. Wall

Appendix

Hill's model of LD evolution

Hill (1975) incorporates mutation, using the model of infinite alleles. However, most modern work involves either DNA sequence data or SNPs. In these data, alleles are nucleotide states: A, T, G, or C. Nearly all loci have just two alleles, so it is not appropriate to assume an infinity of alleles. Yet as we shall see, Hill's (1975) results also apply to a more appropriate mutational model—that of "infinite sites" (Kimura 1969), which assumes that mutation never strikes the same site twice.

Hill begins with a vector of moments:

$$y_{hi,jk} = \begin{pmatrix} E\left[a_h a_i b_j b_k\right] \\ E\left[a_h b_j D_{ik} + a_h b_k D_{ij} \right. \\ \left. \quad + \; a_i b_j D_{hk} + a_i b_k D_{hj}\right] \\ E\left[D_{hj} D_{ik} + D_{hk} D_{ij}\right] \end{pmatrix}.$$

Here a_i and b_j are the frequencies of allele A_i at locus A and allele B_j at locus B. The disequilibrium coefficients, D_{ij}, are defined such that $a_i b_j + D_{ij}$ is the frequency of gamete type $A_i B_j$. The dynamics of these moments are approximately linear, after dropping terms in u^2, $1/N^2$, and u/N, where u is the mutation rate and N the diploid population size. Thus, Hill (1975, equation 1) shows that

$$y_{hi,jk}(t+1) \approx \mathbf{DRM} y_{hi,jk}(t),$$

where **D**, **R**, and **M** are matrices describing the linear effects of drift, recombination, and mutation. So far, the model describes only the changes in frequency of existing alleles. Thus it applies not only to Hill's model of infinite alleles, but also to the model of infinite sites.

To incorporate mutation, Hill defines a new vector,

$$x = \sum_{h \neq i} \sum_{j \neq k} y_{hi,jk}, \tag{A1}$$

where the sums run across all pairs of alleles at each locus. The dynamics of this new vector include an additive contribution, Δx_{mut}, which represents the effect of mutation to new alleles:

$$x(t+1) \approx \mathbf{DRM} x(t) + \Delta x_{\text{mut}}(t). \tag{A2}$$

This equation can be iterated across many generations, and it is easy to incorporate changes in population size. At the end of this process, σ_{d}^2 is calculated as $x_3/2x_1$ (Hill 1975, p. 124).

The first term on the right side of Equation A1 applies equally to both mutational models. Some reinterpretation is required, however, to relate the second term to the model of infinite sites. In this new context, Δx_{mut} becomes the contribution of mutations at other sites in the genome, rather than that from new alleles at the same pair of sites. The mutational increments are, however, identical in the two models.

To see that this is so, let us take a close look at the mutational contributions to x, the vector defined above in Equation A1. Under the model of infinite sites, mutation never strikes the same site twice, so each polymorphic site has exactly two alleles. In this biallelic context, we can simplify Hill's vector of moments as

$$h = \begin{pmatrix} E[a(1-a)b(1-b)] \\ E[(1-2a)(1-2b)D] \\ E\left[D^2\right] \end{pmatrix}. \tag{A3}$$

Here, a and b are the frequencies of alleles A_1 and B_1, and D is the coefficient of linkage disequilibrium. It is defined such that $ab + D$ is the frequency of gamete type $A_1 B_1$. When each locus is biallelic, Hill's multiallelic moments reduce to $x_1 = 4h_1$, $x_2 = 4h_2$, and $x_3 = 8h_3$ (Hill 1975, p. 123).

The mutational increments, Δx_{mut} and Δh_{mut}, both depend on the heterozygosity, $H = E[2a(1-a)]$. The dynamics of heterozygosity under infinite sites are the same as those under infinite alleles (Hill 1975, p. 120):

$$H_{t+1} \approx 2u + \left(1 - \frac{1}{2N} - 2u\right) H_t.$$

Table A1 Effect of a single mutation at *B*, assuming that *A* is initially polymorphic

	Gamete frequencies				
Case	A_1B_1	A_1B_0	A_0B_1	A_0B_0	D
General	w	x	y	z	$wz - xy$
Before mutation	0	a	0	$1 - a$	0
A_1-linked mutation	$1/2N$	$a - 1/2N$	0	$1 - a$	$(1 - a)/2N$
A_0-linked mutation	0	a	$1/2N$	$1 - a - 1/2N$	$- a/2N$

The h_i are nonzero only if both loci are polymorphic. On the other hand, our model assumes that mutation affects only monomorphic sites. This implies that mutational increments to h occur only at pairs of sites in which one site is polymorphic and the other monomorphic. Table A1 summarizes the case in which A is initially polymorphic, and a mutation strikes an initially monomorphic locus, B.

Consider first the increment to $h_1 = E[a(1 - a)b(1 - b)]$. Before mutation, $h_1 = 0$ because $b = 0$. After mutation, $b = 1/2N$, so h_1 becomes $E[a(1 - a)(1/2N - 1/4N^2)] \approx H/4N$. This accounts for only half the effect of mutation, because there are also pairs at which A is monomorphic and B polymorphic. The expected effect of a single mutation is thus $\sim H/2N$. The expected number of such mutations per generation is $2Nu$. In aggregate, therefore, the increment from mutation is $\Delta_{\text{mut}}h_1 \approx uH$.

To derive the mutational increment to $h_2 = E[(1 - 2a)(1 - 2b)D]$, we assume as before that A is polymorphic but B is not. Because D is zero when either site is monomorphic, $h_2 = 0$ before mutation. After mutation, there are two cases to consider. With probability a, the mutation falls on an A_1-bearing gamete, and h_2 becomes $(1 - 2a)(1 - 2/2N)(1 - a)/2N$, as shown in Table A1. On the other hand, with probability $1 - a$ the mutation falls on an A_0-bearing gamete, and h_2 becomes $-(1 - 2a)(1 - 2/2N)a/2N$. In expectation, the new value of h_2 is 0. This result is conditional on a mutation at locus B, but an identical argument applies for mutation at locus A. Thus, $\Delta_{\text{mut}}h_2 = 0$.

The third moment is $h_3 = E[D^2]$. Because B is initially monomorphic, $D_2 = 0$ before the mutation. When a mutation occurs at B, it strikes an A_1-bearing gamete with probability a and an A_0-bearing gamete with probability $1 - a$. The resulting values of D are shown in Table A1. Squaring these and weighting by a and $1 - a$ gives

$$E\left[a\left\{\frac{(1-a)^2}{4N^2}\right\} + (1-a)\left\{\frac{a^2}{4N^2}\right\}\right] = \frac{H}{8N^2}.$$

This is the effect on D^2 of a single mutation at B. There are $2Nu$ such mutations per generation, so the expected increment from mutation at locus B is $2Nu \times H/8N^2 = uH/4N$. Finally, multiply by 2 to account for cases in which A is initially monomorphic and B polymorphic. This gives $\Delta_{\text{mut}}h_3 = uH/2N \approx 0$, ignoring terms of order u/N.

In summary,

$$\Delta_{\text{mut}}h \approx [uH, 0, 0], \tag{A4}$$

in agreement with Hill (1975, p. 121). This shows that the mutational increment is the same under the models of infinite sites and infinite alleles. Because of this equivalence, Hill's results apply equally to both models of mutation.

Estimating Linkage Disequilibrium with Partially Phased Diploid Data

In 1000-genotypes data, not all genotypes are phased. At different loci, the unphased genotypes may correspond to different individuals. Thus, we need a method that can deal with unphased genotypes scattered throughout the data matrix.

The symbols j, k, l, and m represent alleles and will always equal either 0 or 1. I write phased two-locus genotypes in form $\genfrac{}{}{0pt}{}{jk}{lm}$, which says that a diploid individual has genotype jk at locus A and genotype lm at locus B. This represents the union of gametes $\genfrac{}{}{0pt}{}{j}{l}$ and $\genfrac{}{}{0pt}{}{k}{m}$. This genotype is unordered, in that we cannot distinguish the maternal gamete from the paternal one. Consequently, there is no distinction between $\genfrac{}{}{0pt}{}{jk}{lm}$ and $\genfrac{}{}{0pt}{}{kj}{ml}$. When I write genotypes in this form, I imply that linkage phase is known. In other words, genotypes $\genfrac{}{}{0pt}{}{jk}{lm}$ and $\genfrac{}{}{0pt}{}{jk}{ml}$ are not equivalent.

Consider a tiny two-locus data set, consisting of three diploid individuals:

$$G_1, G_1, G_3 = \begin{matrix} j_1k_1 \\ l_1m_1 \end{matrix}, \begin{matrix} j_2k_2 \\ l_2m_2 \end{matrix}, \begin{matrix} j_3k_3 \\ l_3m_3 \end{matrix}.$$

On the right, each row corresponds to a locus and each column to a gamete. To estimate linkage disequilibrium (D), we would need to calculate $S = \sum_{i=1}^{6} x_i y_i$, where x_i is the ith value in the upper row and y_i is the corresponding value in the lower one. This sum can be written as

$$S = s_1 + s_2 + s_3 = (j_1l_1 + k_1m_1) + (j_2l_2 + k_2m_2) + (j_3l_3 + k_3m_3).$$

Here, $s_i = j_il_i + k_im_i$ is the contribution of the ith diploid genotype. This calculation requires phased data, and it is the only step where such data are required in estimating D. To cope with unphased data, we need a method to estimate s_i.

Consider the function $s\begin{pmatrix} jk \\ lm \end{pmatrix} = jl + km$. If at least one genotype is homozygous, then phasing does not matter. For example, $s\begin{pmatrix} jk \\ ll \end{pmatrix} = s\begin{pmatrix} kj \\ ll \end{pmatrix}$. We can calculate s regardless of phasing. The only genotypes that need concern us are those in which both loci are heterozygous.

As mentioned above, all genic values (j, k, l, and m) are either 0 or 1. Double heterozygotes will look like either $\begin{matrix} 01 \\ 01 \end{matrix}$ or $\begin{matrix} 01 \\ 10 \end{matrix}$. (I ignore the equivalent representations $\begin{matrix} 10 \\ 10 \end{matrix}$ and $\begin{matrix} 10 \\ 01 \end{matrix}$, because genotypes are unordered.) With unphased data, we cannot distinguish between these two cases. Yet they imply different values: $s\begin{pmatrix} 01 \\ 01 \end{pmatrix} = 1$ but $s\begin{pmatrix} 01 \\ 10 \end{pmatrix} = 0$. For double heterozygotes, I replace s with its expected value, which equals the probability, w, that the genotype is of form $\begin{matrix} 01 \\ 01 \end{matrix}$ rather than $\begin{matrix} 01 \\ 10 \end{matrix}$.

To calculate this probability, I begin with standard results for the frequencies of the four gamete types, as shown in Table A2. Following Rogers and Huff (2009), I ignore recombination in the most recent generation. Under random mating, these gametes form at random to produce two-locus genotypes. The frequency of genotype $\begin{matrix} 01 \\ 01 \end{matrix}$ among double heterozygotes is thus

$$\begin{aligned} w &= \frac{2p_0p_3}{2p_0p_3 + 2p_1p_2} \\ &= \frac{D + Z}{D + 2Z}, \end{aligned} \tag{A5}$$

where

$$\begin{aligned} Z &= \alpha - \beta D + D^2, \\ \alpha &= a(1-a)b(1-b), \end{aligned}$$

and

$$\beta = a + b - 2ab.$$

These results can be used to estimate D using the EM algorithm (Dempster *et al.* 1977). Let K represent the number of unphased double heterozygotes. Each of these is of type $\begin{matrix} 01 \\ 01 \end{matrix}$ with probability w and of type $\begin{matrix} 01 \\ 10 \end{matrix}$ with probability $1 - w$. The expected log likelihood is

$$\begin{aligned} E \ln L &= \sum_{i=0}^{3} n_i \ln p_i + K[w(\ln p_0 + \ln p_3) + (1-w)(\ln p_1 + \ln p_2)] \\ &= (n_0 + Kw)\ln p_0 + (n_1 + K(1-w))\ln p_1 + (n_2 + K(1-w))\ln p_2 + (n_3 + Kw)\ln p_3. \end{aligned}$$

Note that $E \ln L$ depends on p_0, p_1, p_2, and p_3, which are themselves functions of D. The maximum-likelihood estimate of D is the value that maximizes $E \ln L$.

This is a unidimensional maximization problem, because D is the only unknown. D cannot fall outside the range $[\underline{D}, \overline{D}]$, where $\underline{D}$ is the maximum of $-a(1 - b)$ and $-b(1 - a)$, and $\overline{D}$ is the minimum of $a(1 - b)$ and $b(1 - a)$ (Lewontin 1964, p. 55). The initial value of D is set assuming that $w = 1/2$. In each iteration, if $d^2E \ln L/dD^2 < 0$, then the algorithm tries a Newton step

Table A2 Gamete types and frequencies

Gamete	Sample count	Population frequency
0 0	n_0	$p_0 = (1 - a)(1 - b) + D$
0 1	n_1	$p_1 = (1 - a)b - D$
1 0	n_2	$p_2 = a(1 - b) - D$
1 1	n_3	$P_3 = ab + D$

n_i is the number of copies of gamete i in the sample, excluding unphased double heterozygotes. p_i is the frequency of that gamete type within the population. a is the frequency of allele 1 at locus A, b is the frequency of allele 1 at locus B, and D is the conventional coefficient of linkage disequilibrium.

(Hamming 1973, p. 68). If the result is within $[\underline{D}, \overline{D}]$, then the Newton step is accepted. Otherwise, the algorithm uses a modified version of the bisect algorithm (Hamming 1973, p. 62) to move in the uphill direction, as indicated by the sign of $dE \ln L/dD$.

The modification to bisect allows the algorithm to make large steps when it seems likely that the optimum is at a boundary. For example, if $dE \ln L/dD > 0$ at the current value of D, then motion is to the right, toward $\overline{D}$. Before deciding how far to move, the algorithm checks the sign of the derivative at $\overline{D}$. If both derivatives are positive, then the algorithm takes a big step, moving 80% of the distance from D to $\overline{D}$. But if the two derivatives have opposite sign, then the algorithm takes a small step, moving only 50% of the way to $\overline{D}$. When $dE \ln L/dD < 0$, the algorithm is similar, except that motion is to the left, toward $\underline{D}$.