The optimal measure of allelic association

N. E. Morton*†, W. Zhang*, P. Taillon-Miller‡, S. Ennis*, P.-Y. Kwok‡, and A. Collins*

*Human Genetics Research Division, Duthie Building (Mailpoint 808), Southampton General Hospital, Tremona Road, Southampton SO16 6YD, United Kingdom; and ‡Division of Dermatology, Washington University, St. Louis, MO 63110

Contributed by N. E. Morton, February 6, 2001

Allelic association between pairs of loci is derived in terms of the association probability ρ as a function of recombination θ, effective population size N, linear systematic pressure v, and time t, predicting both ρ_{rt}, the decrease of association from founders and ρ_{ct}, the increase by genetic drift, with $\rho_t = \rho_{rt} + \rho_{ct}$. These results conform to the Malecot equation, with time replaced by distance on the genetic map, or on the physical map if recombination in the region is uniform. Earlier evidence suggested that ρ is less sensitive to variations in marker allele frequencies than alternative metrics for which there is no probability theory. This robustness is confirmed for six alternatives in eight samples. In none of these 48 tests was the residual variance as small as for ρ. Overall, efficiency was less than 80% for all alternatives, and less than 30% for two of them. Efficiency of alternatives did not increase when information was estimated simultaneously. The swept radius within which substantial values of ρ are conserved lies between 385 and 893 kb, but deviation of parameters between measures is enormously significant. The large effort now being devoted to allelic association has little value unless the ρ metric with the strongest theoretical basis and least sensitivity to marker allele frequencies is used for mapping of marker association and localization of disease loci.

Dependence of alleles at two loci is called allelic association, gametic association, or linkage disequilibrium (LD). It has long been of interest for evolutionary genetics (1) and now has become a focus for genetic epidemiology. Its applications include localizing genes of unknown sequence (positional cloning), determining whether a particular allele is descended from a single founder (monophyletic), identifying regions of unusually high or low association that may reflect variations in recombination or a selective sweep, and recognizing effects of population structure and history. Many metrics have been used for LD, for most of which there is no genetic theory (2, 3). Some measure statistical significance and are therefore sensitive to sample size. All metrics are sensitive to variations in marker allele frequencies, and some are acutely sensitive in comparisons of expected and simulated LD. This variability has become increasingly important as many researchers explore LD in different chromosome regions and populations. Their observations are ineffectual unless they use the metric with the strongest theoretical basis and least sensitivity to marker allele frequencies. Fortunately, the two optima coincide. Here we derive genetic and statistical theory and examine the performance of alternative metrics on large samples of random haplotypes.

Genetic Theory

Let the four haplotype frequencies in a random sample for two diallelic loci, A and B, be arranged as in Table 1 with $\pi_{11}\ \pi_{22} \geq \pi_{12}\ \pi_{21}$, $\pi_{12} \leq \pi_{21}$, and $Q \leq R, 1 - Q$. These constraints can always be satisfied by exchanging rows and columns. As Q/R decreases, the probability increases that A^1 is the youngest allele that by chance arose in a gamete carrying B^1. Because π_{ij} is a linear function of $\rho \geq 0$, Table 1 may be written as a commingling of two tables

$$\rho\begin{bmatrix} Q & 0 \\ R-Q & 1-R \end{bmatrix} + (1-\rho)\begin{bmatrix} QR & Q(1-R) \\ (1-Q)R & (1-Q)(1-R) \end{bmatrix} = \begin{bmatrix} \pi_{11} & \pi_{12} \\ \pi_{21} & \pi_{22} \end{bmatrix} \quad [1]$$

with $0 \leq \rho \leq 1$ defined as the association probability. Given the observed haplotype frequencies, all other measures of association are functions of Q, R, and ρ. Under certain conditions to be examined, ρ is an estimate of the probability that a random haplotype has descended without recombination from a founder population in which π_{12} was zero, whereas the complementary class with frequency $1 - \rho$ has undergone at least one crossover between the loci and so the alleles are independent. When $\rho = 0$, there is linkage equilibrium. When $\rho = 1$, there is complete disequilibrium. However, this limit was not reached for the founder population if the $A^1\,B^2$ haplotype was polyphyletic because of gene conversion, population admixture, or recurrent mutation.

Population genetics theory provides the expected association ρ_t in the tth generation after founders. Association between loci plays the same role as identity by descent in kinship theory, except that it deals with one gamete instead of two and the initial value ρ_0 need not be zero. For loci A and B, a random haplotype in generation t is identical by descent from a specified haplotype in $t - 1$ with probability $1/2N_{t-1}$, where N_i is the effective population size in the ith generation. In the complementary event, the probability that a random haplotype in $t - 1$ has undergone no recombination since the founder generation is ρ_{t-1}. Therefore the association probability is

$$\rho_t = (1-v)(1-\theta)\left[\frac{1}{2N_{t-1}} + (1 - 1/2N_{t-1})\rho_{t-1}\right], \quad [2]$$

where v is the linear pressure toward linkage equilibrium from migration and mutation and θ is the recombination frequency. The role of N is to describe stochastic variation, and the role of v is to ensure that over many generations the allele frequencies remain realistically close to their present values. This recurrence satisfies

$$\rho_t - L = (\rho_0 - L)(1-v)^t(1-\theta)^t \prod_{i=0}^{t-1}(1 - 1/2N_i)$$

where L is the association as $t \to \infty$.

As implied by Crow and Kimura (4), we may equate $\Pi_{i=0}^{t-1}(1 - 1/2N_i)$ to $e^{-t/2N}$, where N is the unknown effective size over the unknown sequence $N_0, N_1, \ldots, N_{t-1}$. This is a definition of N, not an approximation, and so the N_i may be constant, randomly varying, exponentially increasing, or an

Abbreviations: LD, linkage disequilibrium (= allelic association); SNP, single-nucleotide polymorphism; cM, centimorgan.

†To whom reprint requests should be addressed. E-mail: nem@soton.ac.uk.

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Table 1. Haplotype frequencies in a random sample

	Locus B		
Locus A	Allele 1	Allele 2	Allele frequency
Allele 1	$\pi_{11} = Q\rho + QR(1-\rho)$	$\pi_{12} = (1-\rho)Q(1-R)$	$Q = \pi_{11} + \pi_{12}$
Allele 2	$\pi_{21} = (R-Q)\rho + R(1-Q)(1-\rho)$	$\pi_{22} = (1-R)\rho + (1-Q)(1-R)(1-\rho)$	$1-Q = \pi_{21} + \pi_{22}$
Allele frequency	$R = \pi_{11} + \pi_{21}$	$1-R = \pi_{12} + \pi_{22}$	1

arbitrary sequence, for a change in the order of the N_i does not alter ρ_t. Because t is large relative to v and θ, we may take $(1-v)^t\,(1-\theta)^t = e^{-(v+\theta)t}$. Rearranging, we finally obtain

$$\begin{aligned}\rho_t &= \rho_{rt} + \rho_{ct}\\ \rho_{rt} &= \rho_0 e^{-(1/2N+v+\theta)t}\\ \rho_{ct} &= L[1-e^{-(1/2N+v+\theta)t}]\end{aligned} \qquad [3]$$

where ρ_{rt} is residual kinship that declines from ρ_0 in founders, and ρ_{ct} is the increasing kinship caused by drift from founders. As t increases, $\rho_{rt} \to 0$ and $\rho_{ct} \to L$. If N is constant,

$$\begin{aligned}L &= \frac{(1-v)(1-\theta)}{2N-(2N-1)(1-v)(1-\theta)}\\ &\cong 1/[1+2N(v+\theta)] \text{ for } \theta \to 0\\ &\cong 1/(1+2N) \text{ for } \theta = 1/2.\end{aligned} \qquad [4]$$

Whether N is constant, ρ_t may be expressed as

$$\rho_t = (1-L)Me^{-\theta t} + L \qquad [5]$$

where $M = (\rho_0 - L)e^{-(v+1/2N)t}/(1-L)$.
The exponential term in M is nearly zero if $v + 1/2N \ll 1/t$, and then $M = 1$ if $\rho_0 = 1$. A value of M not significantly less than 1 is consistent with monophyletic inheritance and a relatively short history. A value of M significantly less than 1 is evidence for $\rho_0 < 1$, and in that sense for polyphyletic origin. Mutation and gene conversion are not negligible over large t, and so polyphyletic origin is likely for a common marker in a large population.

Most population genetics theory deals not with a single path between generation t and the founders but with a double path from founders to two gametes in t (e.g., ref. 4, equation 6.6.2). Sved (5) derived a theory for kinship φ, the probability that locus B is identical by descent (ibd), conditional on ibd for locus A. This double path replaces $1-\theta$ by $(1-\theta)^2$, the term $e^{-t/2N}$ remaining the same. For constant N, the limit of φ_t as t approaches infinity is $1/(1+4N\theta)$ for small θ, and $1/(1+6N)$ for $\theta = 1/2$ (ref. 5, equations 6 and 7). Because kinship does not have a simple relation to LD, we shall not pursue it.

Association as a function of time is not observable except in replicate populations simulated over large t. The theory becomes much more useful when time is transformed into distance along the chromosome, replacing θt by εd where ε is assumed constant for a specified region and d is the distance between loci on the genetic scale as centimorgans (cM) or on the physical scale as kilobases (kb). The association at distance d is predicted as

$$\rho_d = (1-L)Me^{-\varepsilon d} + L, \qquad [6]$$

which is the Malecot equation for isolation by distance (ref. 6, p. 84; ref. 7, p. 75; and ref. 8, equation 3). Let z be the number of distance units per morgan, or 100 if d is expressed in cM and the genetic map is accurate. If the physical scale is used, distance is more precise but proportionality of the genetic and physical scales is assumed and the value of z is uncertain. The rule of thumb that equates megabases (Mb) with cM predicts $z = (100$ cM/morgan) $(1{,}000$ kb/Mb$) = 10^5$ if z is expressed in kb, but there is as much variability in z as in mutation rates. The estimated duration is $z\varepsilon$ generations, and the swept radius over which M is reduced by e^{-1} and LD is useful for positional cloning is $1/\varepsilon$. Whether the Malecot equation is expressed in terms of t, cM, or kb, the L parameter predicts association between unlinked loci, and so provides an efficient alternative to the transmission disequilibrium test that requires family data and makes no use of homozygous parents and their children (9).

The Malecot equation or other realistic model for LD has several advantages. The parameters are meaningful and they illuminate differences among populations and chromosome regions. Oligogenes with effects on a particular phenotype large enough to be detectable by current methods have a density less than 1 per morgan. If the model is fitted to n markers in such a large region to test the null hypothesis that $\varepsilon = 0$, there is no need for a Bonferroni correction that would increase the critical logarithm of odds (lod) from 3 to at least $3 + \log n$ if the markers were tested individually (10). Therefore, testing each marker individually has low efficiency by comparison with the Malecot model.

Statistical Theory

An estimate $\hat{\psi}$ of an association metric with expected value ψ has an amount of information, K_ψ, that allows for simultaneous estimation of Q and R but not for the evolutionary variance that accumulated over time. If n independent samples are tested and m parameters are estimated, the composite likelihood is exp $[-\Sigma_i(\hat{\psi}_i - \psi_i)^2 K_{\psi i}/2]$, where the quadratic form has a χ^2_{n-m} distribution under a true hypothesis. Several of the alternatives to ρ are of the form $\psi = D/C$, where C is a function of Q and R only and so $C = \partial D/\partial\psi$ and $K_\psi = C^2 K_D$. They include the covariance D, the correlation r, the frequency difference f, and the regression b (Table 2). Other alternatives to ρ are of a more complicated form in which $\partial D/\partial\psi$ is a function of D. Examples of this class are D^2/C, δ, and the Yule (12) metric y (two or more other symbols have been used for all these metrics). If $\psi = D^2/C$, then $\partial D/\partial\psi = C/2D$, a function of association that is indeterminate under the null hypothesis. Therefore metrics like ρ^2, r^2, and f^2 should be avoided. The δ metric approaches ρ as $Q \to 0$ and is invariant under case-control sampling. It is defined as $\delta = |\pi_{11}\pi_{22} - \pi_{12}\pi_{21}|/Q\,\pi_{22} = D/[Q\,(1-R-Q+RQ+D)]$. The efficiency of δ relative to ρ decreases as Q increases. The asymptotic standard error for $\ln(1-\delta)$ (11) implies $K_\delta = 0$ if $\pi_{21} = 0$. For determinacy and comparability, we take $D = C\psi$, $\partial D/\partial\psi = C^2/[C - D(\partial C/\partial D)]$, and $K_\psi = K_D(\partial D/\partial\psi)^2$. On the null hypothesis, $\partial D/\partial\psi = C$, and $\chi^2 = \delta^2 K_\delta$. Yule's measure (12) is $y = (\pi_{11}\pi_{22} - \pi_{12}\pi_{21})/(\pi_{11}\pi_{22} + \pi_{12}\pi_{21}) = D/[2Q\,(1-Q)\,R(1-R) + D\,(1-2Q)(1-2R) + 2D^2]$. Yule (12) gave the information in an expression that is quartic in y and goes to zero if any of the π_{ij} is zero. This problem can be solved in the same way as for δ. Table 2 provides $\hat{\psi}$ and K_ψ, distinguishing between the predicted covariance D and its estimate $\hat{D}$.

If there is no doubt about a candidate region, K_ψ should be evaluated under the null hypothesis that $D = 0$, but this is not optimal once a candidate region has been demonstrated and the goal is to obtain the most accurate estimates of the Malecot parameters, perhaps including the location S of a gene affecting

Table 2. Measures of allelic association ψ in random haplotypes

Definition	Symbol	Estimate $\hat{\psi} = \hat{D}/C$
Covariance	D	$\hat{D} = \lvert\pi_{11}\pi_{22} - \pi_{12}\pi_{21}\rvert$
Association	ρ	$\hat{D}/Q(1 - R)$
Correlation	r	$\hat{D}/\sqrt{Q(1 - Q)R(1 - R)}$
Regression	b	$\hat{D}/R(1 - R)$
Frequency difference	f	$\hat{D}/Q(1 - Q)$
Delta	δ	$\hat{D}/Q(1 - R - Q + RQ + D)$
Yule	y	$\hat{D}/[2Q(1 - Q)R(1 - R) + D(1 - 2Q)(1 - 2R) + 2D^2]$

$\hat{D}$ is an estimate with expected value D and information K_D.
$K_D = n/[Q(1 - Q)R(1 - R) + D(1 - 2Q)(1 - 2R) - D^2]$ (ref. 34, equation 3.8).

a quantitative trait or disease. Under H_1, both ψ and K_ψ depend on D. The ALLASS program has an option to estimate ψ and K_ψ simultaneously (http://cedar.genetics.soton.ac.uk/public_html). We find that this refinement has little effect on estimates and standard errors, and does not increase the relative efficiency of alternatives to ρ_0.

Under the Malecot model, as many as four parameters may be estimated: M, L, ε, and S, the location of a disease locus as a function of distance (13). The most general model may be rejected because of a type I error, significant evolutionary variance, nonindependence of samples, variable recombination, map error, or other departure from the model. Then subhypotheses may be tested by using the quadratic form as an error sums of squares with $n - m$ degrees of freedom. The optimal metric consistently minimizes the sums of squares. Because the optimum is chosen on general considerations and not on the sample in hand, confidence intervals are not invalidated by choice of an extremum (14).

Materials and Methods

We applied these methods to several large samples of haplotypes. Two studies observed X chromosomes in males, the other inferred autosomal haplotypes. Ennis *et al.* (15) studied more than 7,000 haplotypes for 8 markers in the *FRAX* region, spanning 790 kb (1.36 cM) on chromosome Xq27–q28. They include two trinucleotide repeats (*FRAXA* and *FRAXE*) that were the focus of the study, 5 dinucleotide repeats (*DXS548*, *FRAXAC1*, *FRAXAC2*, *DXS1691*, and *DXS6687*), and a single-nucleotide polymorphism (SNP), *ATL1*. Nontransmitted haplotypes in mothers of typed males were inferred, assuming no crossover in transmission to the son. This assumption was supported in pedigrees and has minimal error in mother–child pairs. Pre- and full-mutation haplotypes and haplotypes identical by descent from a pedigree founder were excluded to make the gametes representative of the Wessex population. Each common allele or set of alleles of similar size was tested against the other markers to reduce the data to 2×2 tables by an algorithm that works well with major loci (13, 16). Taillon-Miller *et al.* (17) studied 39 SNPs in three populations: an outbred European sample (CEPH) and the more isolated populations of Finland and Sardinia. The SNPs were selectively in two small regions of 1 Mb in Xq25 and 340 kb in Xq28 that suggested strong LD. Eaves *et al.* (18) reported 21 microsatellites in a 6.5-cM interval on chromosome 18q31, with multiple alleles dichotomized around their modes. Families in four populations were studied (U.K., U.S.A., Finland, and Sardinia), yielding 800 unrelated haplotypes for each. Although the families were selected through insulin-dependent diabetes, the putative *IDDM6* locus has a very small relative risk and ascertainment bias was not evident in comparison with affected family-based controls.

For each sample, we fitted the Malecot model with information evaluated under the null hypothesis and then estimated simultaneously. The general model for ε, M, and L gives a residual variance for testing the subhypothesis $L = 0$, which was often tenable over small intervals (<1 Mb). However, over much larger intervals, L was significantly greater than zero, often far too great to be attributed to a small effective population size and almost entirely because of bias in estimating association, which is constrained to be positive, because negative values have no useful interpretation. The bias for ψ is approximately $\sqrt{2/\pi K}$, where K is the mean information per marker pair. Because the bias is relatively large when K_ψ is small, and K_ψ is proportional to sample size, the small samples tolerated by coalescence theory are unsatisfactory for allelic association. Ideally, L would be estimated in each study as the mean association for pairs of unlinked loci. All tests and standard errors are adjusted for the residual variance when it exceeded 1. Efficiency of ψ relative to ρ was estimated as the ratio of the residual variance for ρ to the residual for ψ. Goodness of fit to association parameters was determined by the χ^2 test. We analyzed ρ beginning with *ML* estimates for ψ, giving $-2 \ln lk = A$. Then we estimated M (if $\hat{M} < 1$ for ρ) and L (if $\hat{L} > 0$ for ρ) to give $-2 \ln lk = B$. Then, if the joint *ML* estimates for ρ give $-2 \ln lk = C$ with $k = n - m$ degrees of freedom and $C/k > 1$, we took $(A - B)/(C/k)$ as a χ^2 with 1 or 2 degrees of freedom testing goodness of fit of M, L, and $(B - C)/(C/k)$ as χ^2_1 for goodness of fit of ε. These tests are conservative, because they treat $\hat{\varepsilon}$, $\hat{M}$, $\hat{L}$ for ρ as parameters rather than as estimates for a correlated trait in the same sample.

Results

Estimates of ε for ρ range from 0.0011 to 0.0026, corresponding to swept radii of 893 and 385 kb, respectively (Table 3). M is

Table 3. Estimates of parameters for association ρ under H_0 (distance in kb)

Sample	ε	σ_ε	M	σ_M	L	σ_L	$1/\varepsilon$, kb
Xq, Wessex	0.00238	0.00012	0.604	0.013	0	—	420
Xq, CEPH	0.00255	0.00027	0.793	0.041	0.1276	0.0067	392
Xq, Finland	0.00204	0.00022	0.782	0.039	0.1215	0.0083	490
Xq, Sardinia	0.00112	0.00031	0.675	0.026	0.1300	0.0578	893
18, U.K.	0.00253	0.00014	1	—	0.0342	0.0104	395
18, U.S.A.	0.00260	0.00014	1	—	0.0320	0.0095	385
18, Finland	0.00212	0.00012	1	—	0.0515	0.0118	472
18, Sardinia	0.00234	0.00013	1	—	0.0331	0.0103	427

Table 4. Efficiency relative to association ρ

Source	*D*	r	b	f	δ	y	$\lvert D' \rvert_{max}$
Devlin and Risch (3),* H_0	—	0.417	—	0.389	1.000	0.793	0.822
Collins *et al.* (19), H_0	—	0.576	—	—	—	—	—
This study, H_0	0.262	0.481	0.277	0.773	0.786	0.665	—
This study, except Wessex	0.476	0.610	0.505	0.703	0.652	0.540	—
This study, H_1	0.134	0.231	0.141	0.278	0.449	0.372	—
This study, except Wessex	0.466	0.571	0.490	0.642	0.757	0.623	—

*Case-control sampling.

estimated to be 1 in the chromosome 18 sample of microsatellites, which was dichotomized so as to conserve major modes and is therefore nearly monophyletic. On the contrary, the *FRAX* sample distinguished antimodal alleles that are known to be polyphyletic, and this is reflected in the smallest estimate of *M*, which has no genetic interpretation for metrics other than ρ. Earlier work established that *L* is not significantly different from zero in this 710-kb region (15). Estimates of *L* are greatest in samples with small values of information, K_ρ.

The six measures alternative to ρ provide 48 tests in the eight samples. In no case was the error variance as small as for ρ. Efficiency relative to association is extremely low for *D* and *b* and intermediate for *r*, corresponding to a loss of more than 20% of the information for *f*, *d*, and *y* and more than 70% for *D* and *b* (Table 4). This inefficiency was anticipated by the pilot study of Collins *et al.* (19), which found a relative efficiency of less than 60% for *r*. Relative efficiency of alternatives to ρ is exceptionally low in the *FRAX* sample from Wessex when information is estimated under H_1. Excluding this sample, there is little difference from the estimates under H_0. Convergence under H_1 is slower and the estimates and their standard errors are robust. This lack of improvement under H_1 is not surprising, because the evolutionary variance that exceeds the sampling variance is not modeled. All these metrics are confounded seriously with the gene frequencies *Q* and *R* (2). Devlin and Risch (3) simulated case-control samples from a population with initial frequency $Q_0 = 0.01$. We take δ as a benchmark, because this metric approaches ρ as *Q* approaches 0 and is then a surrogate for ρ. When the average variance among replicates within initial allele frequencies is used, δ has the highest efficiency, followed by $|D'|_{max}$, which equals ρ under random sampling but not in case-control samples (13). Goodness of fit to association parameters under H_0 is extremely poor (Table 5). Clearly, different measures are not equivalent, either for estimates or for error variances.

Discussion

Classical population genetics considered D_{ij} for an arbitrarily specified haplotype, *ij*. If p_i, p_j are the corresponding allele frequencies, D_{ij} lies in the $\pm p_i p_j$ interval. Defined in this way, D_{ij} is as likely to be decreased as increased by genetic drift. This model leads to $D_t = (1 - \theta)^t D_0$, which approaches zero as *t* increases, and does not attempt to explain the initial disequilibrium measured by D_t. The more perceptive geneticists recognized that genetic drift during population contraction was a likely cause, although selection and hybridization between previously isolated populations are not excluded. Hill and Robertson (20) remarked that "any restriction of population size may cause disequilibrium as a result of genetic sampling, and the return to equilibrium will be slow if the loci are tightly linked." Sewall Wright (21) characteristically incorporated this reality into his evolutionary theory:

> This bottleneck effect is greatest in cases in which the total population consists of small demes, each likely to become extinct after a few generations but, if so, always replaced sooner or later by a few stray migrants from populations that have persisted. In this way, every deme at any given time has a history of passage through a great many bottlenecks of small numbers on being traced back from place to place, and since a few momentarily flourishing demes may be the source from which many new colonies are founded, large areas or even the whole species may, in the course of time, trace to a single deme that has passed through many bottlenecks. . . .

The expected value of D_t is $D_0\, e^{-(1/2N+\theta)t}$ in a closed population without mutation (22). This forward equation encounters allele fixation, which led Hill and Robertson (20) to conclude that the expected value of D_t^2 reaches a maximum and then declines to 0. Fixation is obviously irrelevant to a sample polymorphic in generation *t*, and Sved (5) noted that "the derivation of the recurrence relation is a backward calculation, since the value of [LD] is calculated conditional on the observed genotypic distribution in the present generation." Sved was able to show that genetic drift can increase as well as decrease conditional identity by descent, which traces back to the founders and down to the current generation. This seminal result is not directly applicable to allelic association. However, partition of association into a decreasing term (ρ_{rt}) and an increasing term (ρ_{ct}) is made plausible by abandoning D_{ij} in favor of the probability ρ, which in random samples for two diallelic loci (but not for case-control samples) equals $|D'_{ij}|_{max}$, where D_{ij} was defined by Lewontin (23) as a maximum or minimum, without implying a probability. Limitation of ρ to the 0, 1 interval has the same effect as tracing a double path or taking $E(D_{ij}^2)$, introducing a positive value of ρ_{ct}, as in Eq. **3**. As with other evolutionary processes, LD now is modeled as the outcome of stochastic and systematic forces.

Whereas classical theory was preoccupied with change in time, current interest lies in change with distance along the chromosome. For the nominal equivalence of 1 cM to 1 Mb, nearly 700 generations (14,000 years) are required to go halfway to equilibrium at 100 kb (Table 6). For much smaller distances, the halfway time is large relative to duration of our species. Therefore, equilibrium is unlikely for distances less than 100 kb, and doubtful for greater distances.

Table 5. χ^2 goodness of fit to association parameters under H_0

Source	df	*D*	*r*	*b*	*f*	δ	*y*
M, *L*	11	8961	2024	3841	2656	609	3025
ε	8	48	32	87	32	170	316

Table 6. Time to go halfway to equilibrium if $\theta \gg v + 1/2N$

Recombination, θ	cM	Nominal kb	Generations, $t = (\ln 2)/\theta$	Years, $20t$
10^{-6}	0.0001	0.1	693,147	13,862,940
10^{-5}	0.001	1	69,315	1,386,294
10^{-4}	0.01	10	6,931	138,629
10^{-3}	0.1	100	693	13,863
10^{-2}	1	1,000	69	1,386
10^{-1}	10	10,000	7	139

During the last 2 years, less parsimonious models have been introduced, selecting the estimate that is closest to known location (14, 24, 25). The overparametrized Bayesian variant (26) uses a parameter termed "penetrance" that has no relation to that genetic concept or other biological property. The proportion $1 - M$ of disease alleles not derived from the major founder is miscalled penetrance, with the claim that "previous approaches fail to explicitly allow for this in their association models." The best estimates of location by these methods agree with the Malecot model, but estimation with hindsight is not feasible for an unknown location, and haplotype-based methods are not applicable to genes with effects too small to be reliably assigned to a haplotype.

Estimation of time back to founders is central to coalescence theory but peripheral to allelic association. However, when time in number of generation can be inferred correctly from other evidence, it may be used in the Malecot model to estimate ε as t/z, where z is the number of distance units per morgan. If $M = 1$ and $L = 0$, a single marker with association ρ to a disease locus provides an estimate of distance as $(-\ln \hat{\rho})/\varepsilon$. For example, the *DSTST* locus for diastrophic dysplasia was cloned positionally from this information (27). Because the disease is distributed evenly in Finland with a frequency of $Q = 0.008$, it was assumed that the mutation was present among the first Finnish settlers ($t = 100$). The integrated map available at that time, and currently, is consistent with $z = 10^5$ kb/morgan and therefore with $\varepsilon = 0.001$. Association for restriction fragment length polymorphisms within *CSF1R* increases toward the centromere (28). The most proximal marker is *Eco*RI with $\hat{\rho} = 0.94$, about 64 kb away from the *DSTST* mutation with an error of about 8 kb. This precision was lucky, because the information about ρ under H_0 is $K_\rho = \chi^2/\hat{\rho}^2 = 268$, and so the standard error is about $\sqrt{1/K_\rho}/\varepsilon\hat{\rho} = 65$ kb. The Malecot model under these simplifying assumptions is equivalent to the Luria–Delbruck model but is more general. For example, the three points, *BT1*, *CSF1R*/*Eco*RI, and *CSF1R*/*TAGA*, are consistent with $\varepsilon = 0.001$ and localize *DSTST* 88 kb proximal to the *Eco*RI marker with a standard error of 44 kb. No method has been demonstrated to have higher power than the association probability ρ under the Malecot model.

Turning now to marker × marker evidence, it is clear that most genomic regions have swept radii much greater than the 3 kb suggested by a recent simulation (29). Results in Table 3 are consistent with other reports of significant disequilibrium extending to several hundred kb (15, 19, 30–32). However, some small regions have much less or greater LD, which may be caused by a recombination hotspot, selection, or chance. Whatever the cause, an LD map can make positional cloning more efficient by adjusting the density of SNPs to be proportional to ε. Each SNP has its own value of ε, estimated by fitting the Malecot equation to ρ for all pairs with syntenic markers. Then, if two adjacent SNPs at distance d_{12} have estimates ε_1, ε_2, their midpoint contributes $[(\varepsilon_1 + \varepsilon_2)](d_{12}/2)$ to the LD map, the midpoint of the ε_2, ε_3 pair at distance d_{23} contributes $[(\varepsilon_1 + \varepsilon_2)](d_{12}/2) + [(\varepsilon_2 + \varepsilon_3)](d_{23}/2)$, and so on. Collins *et al.* (33) have shown how ε estimated for each SNP delineates the same transition from low to high LD in two populations, confirming the reliability of this approach. Our evidence on relative efficiency of association measures suggests that heterozygosity and other weakly correlated surrogates will give less reliable estimates than ρ in this application as in others.

We are grateful to I. A. Eaves and J. A. Todd for the X chromosome data used in ref. 18. This work was supported by grants from the Medical Research Council.

1. Arunachalam, V. & Owen, A. R. G. (1971) *Polymorphisms with Linked Loci* (Chapman & Hall, London).
2. Hedrick, P. W. (1987) *Genetics* **117,** 331–341.
3. Devlin, B. & Risch, N. (1995) *Genomics* **29,** 311–322.
4. Crow, J. F. & Kimura, M. (1970) *An Introduction to Population Genetics Theory* (Harper & Row, New York).
5. Sved, J. A. (1971) *Theor. Popul. Biol.* **2,** 125–141.
6. Malecot, G. (1969) *The Mathematics of Heredity* (Freeman, San Francisco).
7. Malecot, G. (1973) in *Genetic Structure of Populations*, ed. Morton, N. E. (Univ. of Hawaii Press, Honolulu), pp. 72–75.
8. Morton, N. E., Klein, D., Hussels, I. E., Dodinval, P., Todorov, A., Lew, R. & Yu, S. (1973) *Am. J. Hum. Genet.* **25,** 347–361.
9. Morton, N. E. & Collins, A. (1998) *Proc. Natl. Acad. Sci. USA* **95,** 11389–11393.
10. Risch, N. & Merikangas, K. (1996) *Science* **273,** 1516–1517.
11. Walter, S. D. (1975) *Biometrika* **62,** 371–374.
12. Yule, G. U. (1900) *Philos. Trans. R. Soc. London A* **184,** 257–319.
13. Collins, A. & Morton, N. E. (1998) *Proc. Natl. Acad. Sci. USA* **95,** 1741–1745.
14. Cordell, H. J. & Elston, R. C. (1999) *Genet. Epidemiol.* **17,** 237–252.
15. Ennis, S., Collins, A., Murray, A., Macpherson, J. N. & Morton, N. E. (2000) *Ann. Hum. Genet.* **64,** 513–518.
16. Lonjou, C., Collins, A., Ajioka, R. S., Jorde, L. B., Kushner, J. P. & Morton, N. E. (1998) *Proc. Natl. Acad. Sci. USA* **95,** 11366–11370.
17. Taillon-Miller, P., Bauer-Sardina, I., Saccone, N. L., Putzel, J., Laitinen, T., Cao, A., Kere, J., Pilia, G., Rice, J. P. & Kwok, P.-Y. (2000) *Nat. Genet.* **25,** 324–328.
18. Eaves, I. A., Merriman, T. R., Barber, R. A., Nutland, S., Tuomelehto-Wolf, E., Tuomelehto, J., Cucca, F. & Todd, J. A. (2000) *Nat. Genet.* **25,** 320–323.
19. Collins, A., Lonjou, C. & Morton, N. E. (1999) *Proc. Natl. Acad. Sci. USA* **96,** 15173–15177.
20. Hill, W. G. & Robertson, A. (1968) *Theor. Appl. Genet.* **38,** 226–231.
21. Wright, S. (1969) *The Theory of Gene Frequencies*, Evolution and the Genetics of Populations (Univ. of Chicago Press, Chicago), Vol. 2, p. 215.
22. Hill, W. G. & Robertson, A. (1966) *Genet. Res.* **8,** 269–294.
23. Lewontin, R. C. (1964) *Genetics* **49,** 49–67.
24. Xiong, M. & Guo, S.-W. (1997) *Am. J. Hum. Genet.* **57,** 487–498.
25. McPeek, M. S. & Strahs, A. (1999) *Am. J. Hum. Genet.* **65,** 858–875.
26. Morris, A. P., Whittaker, J. B. & Balding, D. J. (2000) *Am. J. Hum. Genet.* **67,** 155–169.
27. Hastbacka, J., de la Chapelle, A., Kaitila, I., Sistonen, P., Weaver, A. & Lander, E. S. (1992) *Nat. Genet.* **2,** 204–211.
28. Hastbacka, J., de la Chapelle, A., Mahtani, M. M., Clines, G., Reeve-Daly, M. P., Daly, M., Hamilton, B. A., Kusumi, K., Trivedi, B., Weaver, A., *et al.* (1994) *Cell* **78,** 1073–1087.
29. Krugylak, L. (1999) *Nat. Genet.* **22,** 139–144.
30. Huttley, G. A., Smith, M. W., Carrington, M. & O'Brian, S. J. (1999) *Genetics* **152,** 1711–1722.
31. Morton, N. E. & Wu, D. (1988) *Am. J. Hum. Genet.* **42,** 173–177.
32. Jorde, L., Watkins, W. S., Carlson, M., Groden, J., Albertsen, H., Thliveris, A. & Leppert, M. (1994) *Am. J. Hum. Genet.* **54,** 884–898.
33. Collins, A., Ennis, S., Taillon-Miller, P., Kwok, P.-Y. & Morton, N. E. (2001) *Hum. Mutat.*, in press.
34. Weir, B. S. (1990) *Genetic Data Analysis* (Sinauer, Sunderland, MA).