A Fine-Scale Linkage-Disequilibrium Measure Based on Length of Haplotype Sharing

Yan Wang,[1,2] Lue Ping Zhao,[2] and Sandrine Dudoit[1]

[1]Division of Biostatistics, University of California, Berkeley; and [2]Quantitative Genetic Epidemiology, Fred Hutchinson Cancer Research Center, Seattle

High-throughput genotyping technologies for SNPs have enabled the recent completion of the International HapMap Project (phase I), which has stimulated much interest in studying genomewide linkage-disequilibrium (LD) patterns. Conventional LD measures, such as D' and r^2, are two-point measurements, and their relationship with physical distance is highly noisy. We propose a new LD measure, Δ, defined in terms of the correlation coefficient for shared haplotype lengths around two loci, thereby borrowing information from multiple loci. A *U*-statistic–based estimator of Δ, which takes into consideration the dependence structure of the observed data, is developed and compared with an estimator based on the usual empirical correlation coefficient. Furthermore, we propose methods for inferring LD-decay rates and recombination hotspots on the basis of Δ. The results from coalescent-simulation studies and analysis of HapMap SNP data demonstrate that the proposed estimators of Δ are superior to the two most popular conventional LD measures, in terms of their close relationship with physical distance and recombination rate, their small variability, and their strong robustness to marker-allele frequencies. These merits may offer new opportunities for mapping complex disease genes and for investigating recombination mechanisms on the basis of better-quantified LD.

Linkage disequilibrium (LD) refers to the association of alleles at different loci on the same chromosome (Lewontin and Kojima 1960). Such allelic associations are mostly due to physical proximity but could be affected by mutation, recombination, gene conversion, selection, genetic drift, or demographic factors such as inbreeding, migration, and population structure (Xiong and Guo [1997] and their references). Investigating LD patterns has profound implications for understanding the architecture of the human genome, for mapping complex disease loci on a fine scale, for studying population genetics, and for elucidating mechanisms of meiotic recombination. High-throughput genotyping technologies for SNPs have stimulated much interest in studying fine-scale genomewide patterns of common DNA variations with the use of >1 million SNPs from a number of human populations (International HapMap Consortium 2003; Hinds et al. 2005).

Although LD is well defined at a conceptual level, existing approaches for quantifying LD suffer from a number of limitations. Conventional LD measures are typically two-point measures—that is, they quantify LD between two loci, A and B, on the basis of only the allele distributions at these two loci, without exploiting information about the allele distributions of and physical distances from neighboring loci. Despite their popularity, D' and r^2 are both sensitive to allele frequencies (Devlin and Risch 1995) and highly variable in their relationship with the physical distance, d, between A and B. The substantial variability of D' and r^2 makes interpretation of individual LD values challenging. Since average values of D' and r^2 are generally monotonically related to physical distance d, LD patterns based on these conventional measures have been summarized by their average values (Dawson et al. 2002) or by the fraction of common SNPs that are in high LD (e.g., $r^2 > 0.8$) (Hinds et al. 2005) in a region of empirically chosen size.

With the aim of better quantifying LD, several new measures based on population genetics models have been proposed. Morton et al. (2001) proposed an association probability between a pair of loci, under population genetics assumptions regarding recombination, mutation, migration, etc. Other measures do not directly quantify LD in the usual two-locus manner. Instead, LD is assessed in terms of one (reference) locus, by an estimate of the expected genetic distance from the reference locus to either edge of an ancestral segment (McPeek and Strahs 1999) or by an estimate of the population rate of crossing over (theoretically, a function of expected r^2) for a given region (Pritchard and Przeworski 2001). For these model-based measures, robustness to any violation of model assumptions is unknown.

Recognizing the increasing interest in assessing genomewide LD patterns and the limitations of existing measures, we propose a new LD measure, Δ, that borrows information from multiple neighboring loci and

Received October 14, 2005; accepted for publication January 23, 2006; electronically published February 13, 2006.

Address for correspondence and reprints: Dr. Yan Wang, Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue North, M2-B500, Seattle, Washington 98109. E-mail: yanw@stat.berkeley.edu

Am. J. Hum. Genet. 2006;78:615–628. © 2006 by The American Society of Human Genetics. All rights reserved. 0002-9297/2006/7804-0009$15.00

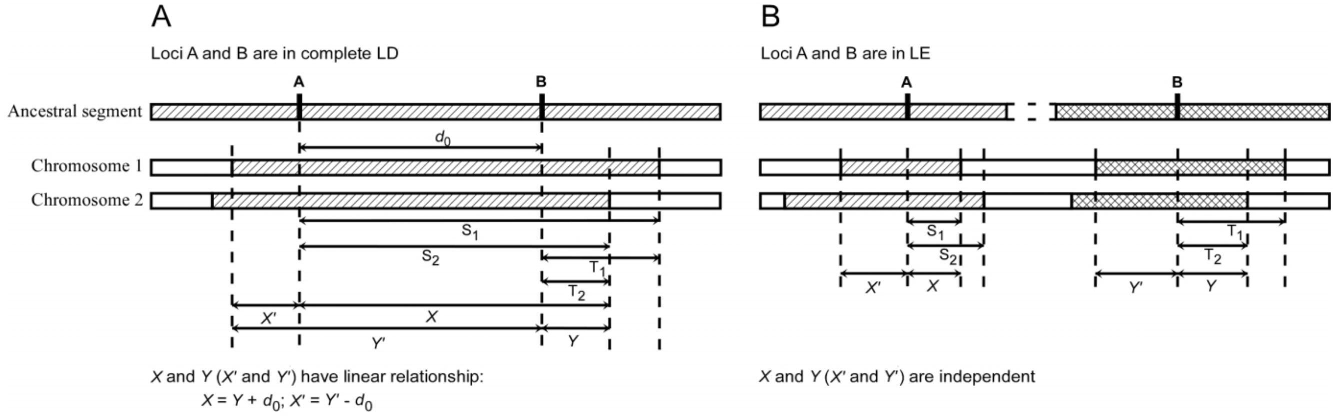
A
Loci A and B are in complete LD
A
B
Ancestral segment
d_0
Chromosome 1
Chromosome 2
S_1
S_2
T_1
T_2
X'
X
Y'
Y
X and Y (X' and Y') have linear relationship:
$X = Y + d_0$; $X' = Y' - d_0$
B
Loci A and B are in LE
A
B
S_1
S_2
T_1
T_2
X'
X
Y'
Y
X and Y (X' and Y') are independent

Figure 1 Conceptual model for motivating the LD measure Δ. When loci A and B are in complete LD, the lengths of haplotype sharing around loci A and B are linearly dependent for all chromosome pairs (*A*). When loci A and B are in LE, the lengths of haplotype sharing around A and B are independent for all chromosome pairs (*B*).

does not require restrictive modeling assumptions. For a reference locus on any chromosome, an ancestral segment refers to the haplotype preserved from an ancestral chromosome. The ancestral segment extends in both directions from the reference locus to breakpoints, which are the closest loci where events such as crossover or gene conversion occurred during meiosis processes intervening between the ancestral and current chromosomes. Given a dense set of markers in a large region, the lengths of common ancestral segments between chromosomes can be well approximated by the lengths of shared haplotypes and can lead to a sensible and stable measure of allelic association between two loci.

In the "Methods" section, we first define Δ as a function of the correlation coefficient between the lengths of common ancestral segments around two loci of interest. Next, we develop a *U*-statistic–based estimator of Δ, $\hat{\Delta}^U$, that takes into account the dependence structure of the observed lengths of shared haplotypes for pairs of chromosomes. An alternative estimator, $\hat{\Delta}$, that naively ignores this dependence structure is also proposed as a simplified and computationally more efficient version. In the "Results" section, simulation studies show that the two estimators, $\hat{\Delta}^U$ and $\hat{\Delta}$, are strikingly similar. Thus, the remaining simulations and applications to HapMap data focus on properties of the simpler and computationally more tractable estimator, $\hat{\Delta}$. A method is proposed for estimating LD-decay rates on the basis of the tight relationship between $\hat{\Delta}$ and physical distance *d*. Then, merits of Δ are demonstrated by analyzing human X-chromosome SNP data from the HapMap Project. We close with a discussion of issues regarding evaluation of the lengths of common ancestral segments.

Methods

A New LD Measure, Δ

Figure 1 shows the conceptual model that motivates the definition of the LD parameter Δ. For a pair of chromosomes that share a common ancestor around locus A, denote the lengths of the ancestral segments from locus A to their respective breakpoints on one side (e.g., right side) by random variables S_1 and S_2. Given a locus B located to the right of A at distance d_0, random variables T_1 and T_2 can be defined in the same way. In practice, neither the ancestral haplotypes nor the breakpoints are observable; thus, neither are S_1, S_2, T_1, and T_2. Given a dense set of markers, one may observe the lengths of haplotypes shared by the chromosome pair that approximate the lengths of the shared common ancestral segments. These shared haplotype lengths, denoted by $X \approx \min(S_1,S_2)$ and $Y \approx \min(T_1,T_2)$, may be measured either by physical distance—that is, the number of base pairs (in bp or kb)—or by genetic distance (in cM). However, the former distance is more precise and more relevant because the most appropriate type of data for our proposed methods is that of dense sets of markers (see the "Results" and "Discussion" sections regarding marker density).

Two main assumptions are involved in approximating the lengths of shared common ancestral segments by the lengths of shared haplotypes. One is that mutation on the common ancestral segment is ignorable, which is reasonable given the extremely low mutation rate for SNPs. The other is that all alleles identical by state (IBS) are identical by descent (IBD). This second assumption may appear strong, yet the "Discussion" section shows that the new LD measure is robust to violations of the assumption.

Two extreme cases are illustrated in figure 1. When chromosomal segments around loci A and B are co-inherited from the same ancestor, all alleles between loci A and B are IBD (under the assumption of no mutation). In this case, A and B are in complete LD, and the linear relationship $X = Y + d_0$ holds for all chromosome pairs in the population. This perfect linear dependence between X and Y is characterized by a Pearson correlation coefficient of $\rho_{xy} = 1$ (fig. 1*A*). On the other hand, when chromosomal segments around loci A and B are inherited from two independent (unrelated) ancestors, A and B are in complete linkage equilibrium (LE). In this case, the above linear relationship does not hold, and X and Y are independent, corresponding to $\rho_{xy} = 0$ (fig. 1*B*). Therefore, ρ_{xy} quantifies the magnitude of LD between A and B.

The same situation applies to the lengths of shared haplotypes on the other side of the reference loci—that is, X' and Y' in figure 1. However, the relationship between $X + X'$ and $Y + Y'$ is more complex. Therefore, we treat separately the lengths of haplotype sharing to the right and left sides of the reference loci. The new LD measure, Δ, is defined as the arithmetic mean of ρ_{xy} and $\rho_{x'y'}$.

There is a statistical challenge in estimating Δ because of the dependence structure of the observed lengths of shared haplotypes between pairs of chromosomes. In the following subsections, an estimator of Δ is developed on the basis of unbiased *U*-statistics (Lee 1990). We first consider the simplest scenario, in which the sampled haplotypes are distinct by state, which is the case in practice when the population or number of markers is large enough. Then, we extend the method to the general situation where haplotypes are not necessarily distinct.

An Estimator of Δ Based on U*-Statistics: $\hat{\Delta}^U$ for Distinct Haplotypes*

Suppose that one observes n distinct haplotypes, $\{h_i : i = 1,\dots,n\}$, for a random sample of n chromosomes from a population of interest. It is assumed that the unobservable ancestral segment lengths $\{(S_i,T_i) : i = 1,\dots,n\}$ are independently and identically distributed (iid), with joint cumulative distribution function $F(S,T), S \geqslant 0, T \geqslant 0$. From $\{h_i : i = 1,\dots,n\}$, one observes $\mathbf{X} = \{X_{ij} : i,j = 1,\dots,n, i<j\}$ and $\mathbf{Y} = \{Y_{ij} : i,j = 1,\dots,n, i<j\}$, the pairwise lengths of one-sided shared haplotypes for loci A and B, respectively, where (i,j) indexes the $\binom{n}{2}$ distinct pairs of haplotypes. On the basis of the aforementioned assumptions, $X_{ij} \approx \min(S_i,S_j)$ and $Y_{ij} \approx \min(T_i,T_j)$.

As mentioned before, one of the statistical challenges in estimating the correlation of X and Y is that neither the $\mathbf{X}$ nor the $\mathbf{Y}$ is a set of independent random variables. To develop a reasonable estimator for the correlation of X and Y, we use *U*-statistics. As shown in the following proposition, the variances and covariance of X and Y are statistical functionals of degree 4, with kernels that are symmetric functions of four iid random variables. Here, a function is said to be symmetric if it is invariant under permutations of its arguments. As a result, according to Lee (1990, p. 7), the variances and covariance of X and Y can be estimated by the average kernels, termed "*U*-statistics" because of their unbiasedness. The correlation coefficient of X and Y is then estimated by the estimated covariance standardized by the estimated SDs of X and Y.

Proposition.—Let σ_{xy} be a statistical functional of degree 4 with kernel function ψ. That is, define σ_{xy} as

$$\sigma_{xy} = E\{\psi[(S_1,T_1),\dots,(S_4,T_4)]\} = \int_0^\infty \cdots \int_0^\infty \psi[(s_1,t_1),\dots,(s_4,t_4)] \prod_{i=1}^{4} dF(s_i,t_i) \ ,$$

where the kernel function is

$$\begin{aligned}\psi[(S_1,T_1),\dots,(S_4,T_4)] = \frac{1}{6}\{&[\min(S_1,S_2) - \min(S_3,S_4)] \\ &\times[\min(T_1,T_2) - \min(T_3,T_4)] \\ &+[\min(S_1,S_3) - \min(S_2,S_4)] \\ &\times[\min(T_1,T_3) - \min(T_2,T_4)] \\ &+[\min(S_1,S_4) - \min(S_2,S_3)] \\ &\times[\min(T_1,T_4) - \min(T_2,T_3)]\}\end{aligned}$$

and $\{(S_i,T_i) : i = 1,\dots,n\}$ are iid with cumulative distribution function $F(S,T)$. Then, for $X = \min(S_1,S_2)$ and $Y = \min(T_1,T_2)$, σ_{xy} is the covariance of X and Y. The proof is provided in appendix A.

As a result of the proposition, the unique unbiased estimator of the covariance σ_{xy} has the form of a *U*-statistic,

$$\hat{\sigma}^U_{xy} = \binom{n}{4}^{-1} \sum_{(n,4)} \psi[(S_{i_1},T_{i_1}),\dots,(S_{i_4},T_{i_4})] \ ,$$

where the sum $\sum_{(n,4)}$ is taken over all distinct four-element subsets $\{i_1,i_2,i_3,i_4\}$ from $\{1,\dots,n\}$. The unobservable random variables $\min(S_i,S_j)$ and $\min(T_i,T_j)$ in the kernel function ψ are then approximated by the corresponding observable random variables X_{ij} and Y_{ij}. Hence, the kernel function can be approximated as

$$\begin{aligned}\psi[(S_{i_1},T_{i_1}),\dots,(S_{i_4},T_{i_4})] \approx \frac{1}{6}[&(X_{i_1i_2} - X_{i_3i_4})(Y_{i_1i_2} - Y_{i_3i_4}) \\ &+(X_{i_1i_3} - X_{i_2i_4})(Y_{i_1i_3} - Y_{i_2i_4}) \\ &+(X_{i_1i_4} - X_{i_2i_3})(Y_{i_1i_4} - Y_{i_2i_3})] \ .\end{aligned}$$

Denote the variances of X and Y by σ_x and σ_y, respectively. These are also statistical functionals of degree 4:

$$\sigma_x = E[\psi_x(S_1,\dots,S_4)] = \int_0^\infty \cdots \int_0^\infty \psi_x(s_1,\dots,s_4) \prod_{i=1}^{4} dF(s_i) \ ,$$

where

$$\psi_x(S_1,\ldots,S_4) = \frac{1}{6}\{[\min(S_1,S_2) - \min(S_3,S_4)]^2 + [\min(S_1,S_3) - \min(S_2,S_4)]^2 + [\min(S_1,S_4) - \min(S_2,S_3)]^2\} \ .$$

One may likewise express σ_y and ψ_y for the variance of Y.

Then, the unique unbiased estimators for σ_x and σ_y are both U-statistics

$$\hat{\sigma}_x^U = \binom{n}{4}^{-1}\sum_{(n,4)}\psi_x(S_{i_1},\ldots,S_{i_4})$$

and

$$\hat{\sigma}_y^U = \binom{n}{4}^{-1}\sum_{(n,4)}\psi_y(T_{i_1},\ldots,T_{i_4}) \ ,$$

where the kernel functions ψ_x and ψ_y are approximated by

$$\psi_x(S_{i_1},\ldots,S_{i_4}) \approx \frac{1}{6}[(X_{i_1i_2} - X_{i_3i_4})^2 + (X_{i_1i_3} - X_{i_2i_4})^2 + (X_{i_1i_4} - X_{i_2i_3})^2]$$

and

$$\psi_y(T_{i_1},\ldots,T_{i_4}) \approx \frac{1}{6}[(Y_{i_1i_2} - Y_{i_3i_4})^2 + (Y_{i_1i_3} - Y_{i_2i_4})^2 + (Y_{i_1i_4} - Y_{i_2i_3})^2] \ .$$

A reasonable estimator of the correlation ρ_{xy} is then

$$\hat{\rho}_{xy}^U = \frac{\hat{\sigma}_{xy}^U}{\sqrt{\hat{\sigma}_x^U\hat{\sigma}_y^U}} \ .$$

Up to this point, we have considered the length of haplotype sharing to one side of a reference locus. Another correlation coefficient, $\hat{\rho}_{x'y'}^U$, can be computed likewise for the length of haplotype sharing to the other sides of a pair of loci. An estimator of Δ, denoted by $\hat{\Delta}^U$, is then the arithmetic mean of $\hat{\rho}_{xy}^U$ and $\hat{\rho}_{x'y'}^U$. This measure has reduced variance compared with the two individual correlation coefficients (Y. Wang, unpublished results).

Although, theoretically, correlation coefficients range from -1 to 1, $\hat{\Delta}^U$ values are seldom negative in our numerical studies. Negative values may occur because of stochastic variation around the true value of zero. In practice, those negative values can be converted to zero.

An Estimator of Δ Based on Weighted U*-Statistics: $\hat{\Delta}^U$ for Nondistinct Haplotypes*

Next, consider the case in which the n observed haplotypes are not necessarily distinct. Suppose there are m distinct haplotypes $\{h_i : i = 1,\ldots,m\}$ that follow a multinomial distribution with parameters $(n,\boldsymbol{\theta})$, where $\boldsymbol{\theta} = \{\theta_i i = 1,\ldots,m\}$ are haplotype frequencies. The haplotype frequencies are the empirical frequencies for phase-known genotype data or may be inferred in the case of unphased data. Among all the distinct four-element subsets of $\{1,\ldots,m\}$, the probability for a given subset (i_1,i_2,i_3,i_4) is $w^U(i_1,i_2,i_3,i_4) = 4!\theta_{i_1}\theta_{i_2}\theta_{i_3}\theta_{i_4}/W^U$, where the denominator W^U is chosen so that $\sum_{(m,4)} w^U(i_1,i_2,i_3,i_4) = 1$.

Then, Lee (1990, p. 64) implies that unbiased estimators for the variances and covariance of X and Y can be obtained from U-statistics weighted by w^U:

$$\hat{\sigma}_{xy}^U = \binom{m}{4}^{-1}\sum_{(m,4)} w^U(i_1,i_2,i_3,i_4)\psi[(S_{i_1},T_{i_1}),\ldots,(S_{i_4},T_{i_4})] \ ,$$

$$\hat{\sigma}_x^U = \binom{m}{4}^{-1}\sum_{(m,4)} w^U(i_1,i_2,i_3,i_4)\psi_x(S_{i_1},\ldots,S_{i_4}) \ ,$$

and

$$\hat{\sigma}_y^U = \binom{m}{4}^{-1}\sum_{(m,4)} w^U(i_1,i_2,i_3,i_4)\psi_y(T_{i_1},\ldots,T_{i_4}) \ .$$

For n distinct haplotypes, the weighted U-statistics reduce to the unweighted U-statistics. The correlation coefficient based on weighted U-statistics can be readily applied to unphased genotype data, after haplotype frequencies $\{\boldsymbol{\theta}_i\}$ are inferred through, for instance, the expectation-maximization (EM) algorithm (Excoffier and Slatkin 1995; Hawley and Kidd 1995; Long et al. 1995).

An Alternative Naive Estimator of Δ, $\hat{\Delta}$

The computation of $\hat{\Delta}^U$ as defined above involves enumerating all $\binom{m}{4}$ distinct four-element subsets $\{i_1,i_2,i_3,i_4\}$ of $\{1,\ldots,m\}$ and can be burdensome when the number of distinct haplotypes m is large. When the dependence structure within $\mathbf{X}$ and within $\mathbf{Y}$ is ignored, intensive computation can be avoided by using a naive estimator $\hat{\rho}_{xy}$ of ρ_{xy}.

In the case of n distinct haplotypes,

$$\hat{\rho}_{xy} = \binom{n}{2}^{-1}\sum_{(n,2)}\frac{(X_{ij} - \bar{X})(Y_{ij} - \bar{Y})}{\sqrt{\hat{\sigma}_x\hat{\sigma}_y}} \ ,$$

where $(\bar{X},\hat{\sigma}_x)$ and $(\bar{Y},\hat{\sigma}_y)$ are the usual sample means and variances for the $\binom{n}{2}$ elements of $\mathbf{X}$ and $\mathbf{Y}$, respectively.

In the case of nondistinct haplotypes, each term within the summation above can be weighted by the probability of observing the subset (i,j) from $\{1,\ldots,m\}$:

$$w(i,j) = \frac{2\theta_i\theta_j}{1 - \sum_{k=1}^{n}\theta_k^2} \ .$$

Hence,

$$\hat{\rho}_{xy} = \binom{m}{2}^{-1}\sum_{(m,2)}\frac{w(i,j)(X_{ij} - \bar{X})(Y_{ij} - \bar{Y})}{\sqrt{\hat{\sigma}_x\hat{\sigma}_y}} \ ,$$

where $\hat{\sigma}_x$ and $\hat{\sigma}_y$ are also weighted by $w(i,j)$,

$$\hat{\sigma}_x = \binom{m}{2}^{-1} \sum_{(m,2)} w(i,j)(X_{ij} - \bar{X})^2 \ ,$$

and

$$\hat{\sigma}_y = \binom{m}{2}^{-1} \sum_{(m,2)} w(i,j)(Y_{ij} - \bar{Y})^2 \ .$$

Therefore, we propose, as a computationally simpler estimator $\hat{\Delta}$, the average of $\hat{\rho}_{xy}$ and $\hat{\rho}_{x'y'}$. Simulation studies show that $\hat{\Delta}$ serves as a good approximation of $\hat{\Delta}^U$ (see the "Results" section). The two estimators are summarized in appendix B. An R package, *haploshare,* was developed and used to implement computation of the two estimators.

Estimation of Δ for Unphased Data

For estimation of Δ from unphased data, a straightforward two-stage scheme can be adopted. In the first stage, either haplotypes and their frequencies are inferred for the whole data set or phases are inferred for each individual chromosome. We prefer the first approach through the EM algorithm, because it produces unbiased estimates of haplotype frequencies, which can then be used in the second stage to calculate $\hat{\Delta}^U$, or $\hat{\Delta}$ as in the setting of nondistinct haplotypes. The second approach usually results in many ambiguous phases for individual chromosomes, which may compromise the accuracy and stability of $\hat{\Delta}^U$ and $\hat{\Delta}$.

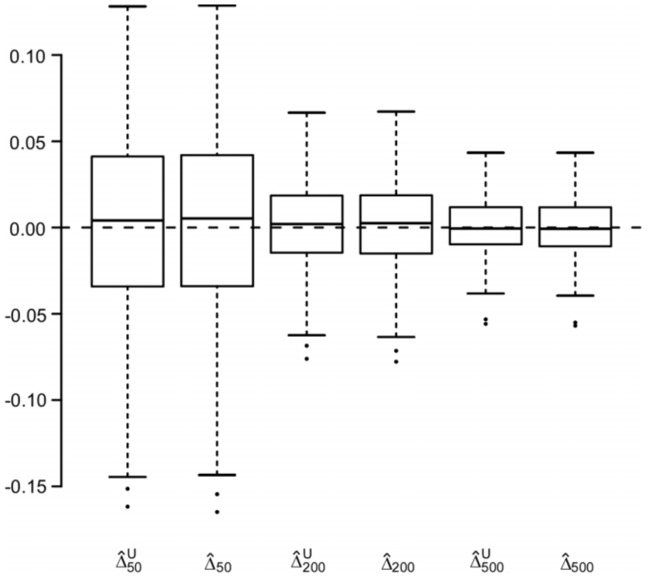

Figure 2 Boxplots of biases for the two estimators $\hat{\Delta}^U$ and $\hat{\Delta}$. From a simulated population of $N = 4{,}000$ haplotypes, 300 random samples of n haplotypes were drawn for $n = 50$, 200, and 500. $\hat{\Delta}^U$ and $\hat{\Delta}$ were calculated for a particular marker pair located 5 kb apart. The six boxplots (*from left to right*) are for $\hat{\Delta}^U_{50}$, $\hat{\Delta}_{50}$, $\hat{\Delta}^U_{200}$, $\hat{\Delta}_{200}$, $\hat{\Delta}^U_{500}$, and $\hat{\Delta}_{500}$, where the numbers in the subscript denote the sample size, n.

Table 1

SDs of $\hat{\Delta}^U$ and $\hat{\Delta}$ in a Simulation Study

Sample Size	SD					
	$d = 5$ kb		$d = 50$ kb		$d = 100$ kb	
(n)	$\hat{\Delta}^U$	$\hat{\Delta}$	$\hat{\Delta}^U$	$\hat{\Delta}$	$\hat{\Delta}^U$	$\hat{\Delta}$
50	.0572	.0579	.0814	.0835	.0532	.0540
200	.0239	.0245	.0387	.0397	.0228	.0231
500	.0154	.0158	.0226	.0232	.0132	.0135

Note.—SDs were calculated for 300 random samples of n haplotypes for three marker pairs arbitrarily chosen with physical distance $d = 5$, 50, and 100 kb.

Results

To investigate properties of Δ and its estimators $\hat{\Delta}^U$ and $\hat{\Delta}$, we performed a series of simulation studies based on genotype data generated by the *ms* program (Hudson 2002) using the finite-site uniform recombination model. Under this model, the recombination rate for a fixed physical distance should be constant in a given sample, leading to our expectation that LD decays at a constant rate. The crossover probability was chosen to be 10^{-8} bp^{-1} for adjacent base pairs, to approximate the average recombination rate for the human genome (i.e., 1 cM per Mb). We then applied our new method to HapMap data, to study human fine-scale genomewide LD patterns.

Comparison of $\hat{\Delta}^U$ to Its Approximation $\hat{\Delta}$ and Impact of Sample Size

We first focused, for simplicity, on fully phased data, to assess our two main estimators of the LD parameter Δ. Since the underlying parameter value Δ cannot be explicitly specified in the simulations with *ms,* we generated a large population of haplotypes ($N = 4{,}000$) for 666 markers covering 300 kb and set $\hat{\Delta}^U_{4000} = 0.848$ as the true Δ for a given marker pair located 5 kb apart. Then, 300 samples were drawn from the simulated population for sample sizes $n = 50$, 200, and 500. For each sample, both $\hat{\Delta}^U$ and $\hat{\Delta}$ were calculated. In figure 2, boxplots of the biases $\hat{\Delta}^U - \Delta$ and $\hat{\Delta} - \Delta$ for these 300 samples suggest that, as sample size n increases, both estimators converge to the true Δ. The same analysis was performed for marker pairs located 50 kb and 100 kb apart, and similar results for the biases and variances of $\hat{\Delta}^U$ and $\hat{\Delta}$ were produced. For all three marker pairs, the SDs for $\hat{\Delta}^U$ are slightly smaller than those for $\hat{\Delta}$ (table 1). Overall, however, the dependence structures within **X** and within **Y** do not seem to have much impact on the estimation of the correlation coefficient of X and Y, and, in practice, the two estimators, $\hat{\Delta}^U$ and $\hat{\Delta}$, can be considered equivalent.

On the basis of these results and because of its simplicity and computational efficiency, we used the naive estimator $\hat{\Delta}$ for the remainder of the simulation studies and for the HapMap data analysis.

Relationship of $\hat{\Delta}$ with Physical Distance

A data set of 245 SNPs, with minor-allele frequencies (MAFs) >5%, was simulated in a region of 500 kb for 200 chromosomes. We focused on 81 SNPs spanning the 200-kb region of interest. Other markers on the flanking regions of the 200-kb region provide information about the length of haplotype sharing for a reference locus at the edge of the region of interest. We computed and displayed the three pairwise LD measures D' (fig. 3*A*), r^2 (fig. 3*B*), and $\hat{\Delta}$ (fig. 3*C*) for each of the $\binom{81}{2} = 3{,}240$ SNP pairs in the region of interest. Visualization tools from the Bioconductor R *hexbin* package were applied to produce "2D histograms," which represent the density of data points in a scatterplot by using hexagonal bins of varying areas. Both $|D'|$ and r^2 tend to decrease as physical distance d increases, as shown by the locally weighted scatterplot-smoothing (lowess) curves. However, $|D'|$ and r^2 are highly variable at any given d. In contrast, $\hat{\Delta}$ has a nearly deterministic relationship with d.

Moreover, the lowess curve for $\hat{\Delta}$ nearly overlaps the exponential function $exp(-\hat{\beta}d)$ (fig. 3*C*, dotted line), implying that the relationship between $\hat{\Delta}$ and d fits the expectation that LD decays exponentially with increasing genetic distance, which is equivalent to physical distance d on a fine scale under the uniform-recombination model. Specifically, $\hat{\Delta}$ is nearly 1 for any pair of closely located SNPs and decreases at a constant rate. By fitting the linear model $E[\log\hat{\Delta}] = -\beta d$, the LD-decay rate may be estimated as $\hat{\beta} = 0.014$, meaning that $\hat{\Delta}$ decays exponentially at a rate of 0.014 per kb in the region of interest.

Impact of Phase-Information Loss on $\hat{\Delta}$

Estimation of Δ, given phase-unknown genotype data, relies on the inference of haplotypes and their frequencies. For a large number of markers, haplotype inference can be very computationally challenging, and many existing programs adopt partition-ligation techniques in which the sequence of markers is partitioned into blocks. However, this strategy may compromise the accuracy of inferred haplotypes for the entire marker sequence. In the following simulation study, we studied how well $\hat{\Delta}$ performs for unphased data with different numbers of markers used for haplotype estimation. It was expected that the more accurate the inferred haplotypes, the more robust the estimated Δ. Here, haplotypes were estimated using the software package HPlus, which applies estimating equation theory to implement efficient maxi-

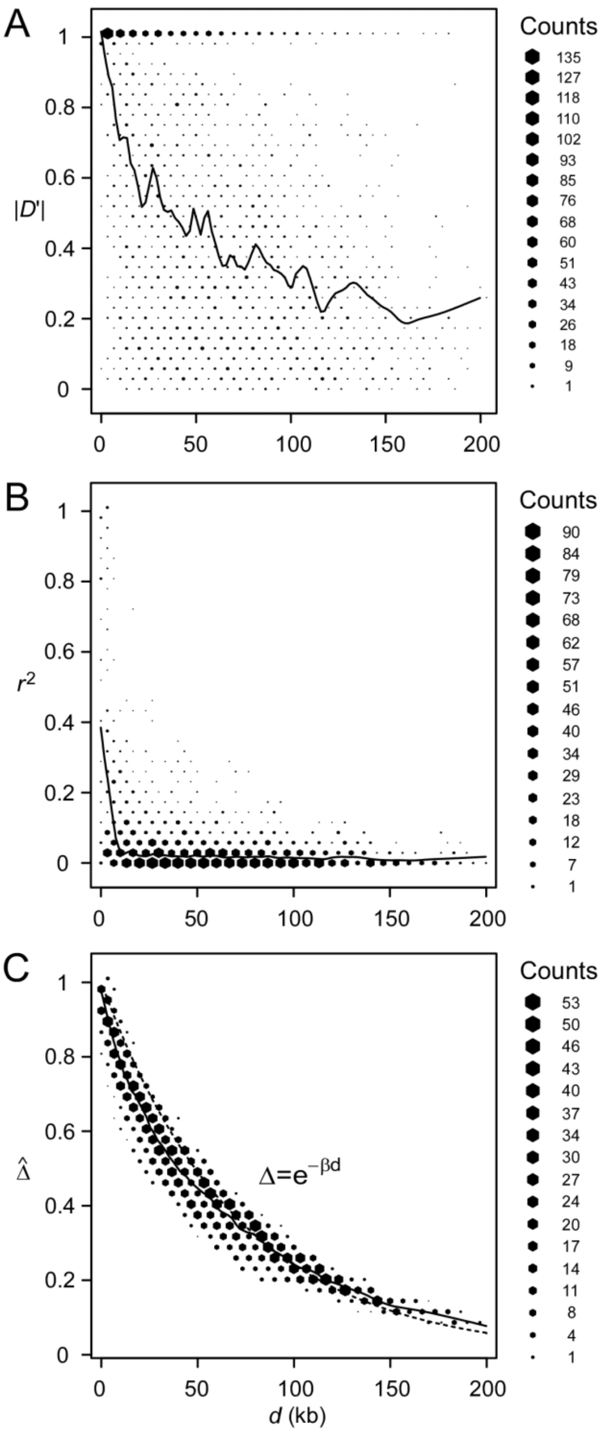

Figure 3 Pairwise LD measured as a function of physical distance *d*. *A*, $|D'|$; *B*, r^2; *C*, $\hat{\Delta}$. Hexagonal bins of different areas are used to represent counts (Bioconductor R package *hexbin*). Locally weighted scatterplot-smoothing (lowess) curves are plotted in black, with the smooth (the *Y*-axis value) at each value influenced by 5% data points. For SNP data simulated under the uniform recombination model, LD decays exponentially at a constant rate, which can be estimated on the basis of the linear regression model $E[\log\hat{\Delta}] = -\beta d$. This relationship is plotted with the dotted line.

mum-likelihood estimation of haplotype frequencies and their variances (Li et al. 2003).

We randomly paired the above 200 haplotypes to create 100 diploid individuals with unphased genotypes. The region of interest was still the 200 kb containing the 81 SNPs. First, haplotypes were estimated for the entire set of 245 SNPs. On the basis of inferred haplotypes with estimated frequencies >0.02%, the naive estimator $\hat{\Delta}$ was then computed for the 3,240 SNP pairs and was plotted against physical distances d (fig. 4*A*). Next, we reduced the number of markers to be haplotyped to 141 and 101, corresponding to 30 and 10, respectively, SNPs in each flanking region, in addition to the 81 SNPs in the region of interest. Figure 4*B* and 4*C* show that smaller variations of $\hat{\Delta}$, at any given distance d, are observed for the smaller numbers of markers. These results suggest that $\hat{\Delta}$ performs well for unphased data and has a similar relationship with physical distance d (fig. 3*C*), especially when markers are chosen in such a way that haplotypes are inferred reliably.

However, it is not true that the fewer the markers the better. Comparing figures 4*B* and 4*C*, we observe that Δ is underestimated when only 10 SNPs instead of 30 are used in the flanking regions to evaluate the length of haplotype sharing. As a result, the LD-decay rates in the two analyses are different, with $\hat{\beta} = 0.016$ in figure 4*B* and $\hat{\beta} = 0.024$ in figure 4*C*, the former being much closer to that estimated with phased data ($\hat{\beta} = 0.014$) using all 245 SNPs.

Therefore, there appears to be the following trade-off. With reduction of the number of markers to be haplotyped and used to evaluate the lengths of haplotype sharing, haplotype inference is more reliable, leading to more robust estimation of Δ. However, the lengths of haplotype sharing might be more censored (see the "Discussion" section) or evaluated on the basis of sparser sets of SNPs (see the "Impact of Marker Density on $\hat{\Delta}$" subsection), which would lead to biased estimation of Δ. Note that, even though estimation of Δ is biased (as in fig. 4*C*), the strong relationship of $\hat{\Delta}$ with physical distance is still present, which implies the potential usefulness of $\hat{\Delta}$ in this situation.

Impact of Marker Density on $\hat{\Delta}$

Marker density can have a significant impact on how well the length of haplotype sharing approximates the length of the common ancestral segment. Generally speaking, the denser the marker map, the better the approximation. To study the effect of marker density on $\hat{\Delta}$, we simulated a data set of 200 haplotypes in a region of size 180 kb. From the 196 SNPs with MAFs >1% in the middle 100-kb region of interest, we randomly selected subsets of SNPs according to the following percentages: 90%, 70%, 50%, 30%, and 10%. These per-

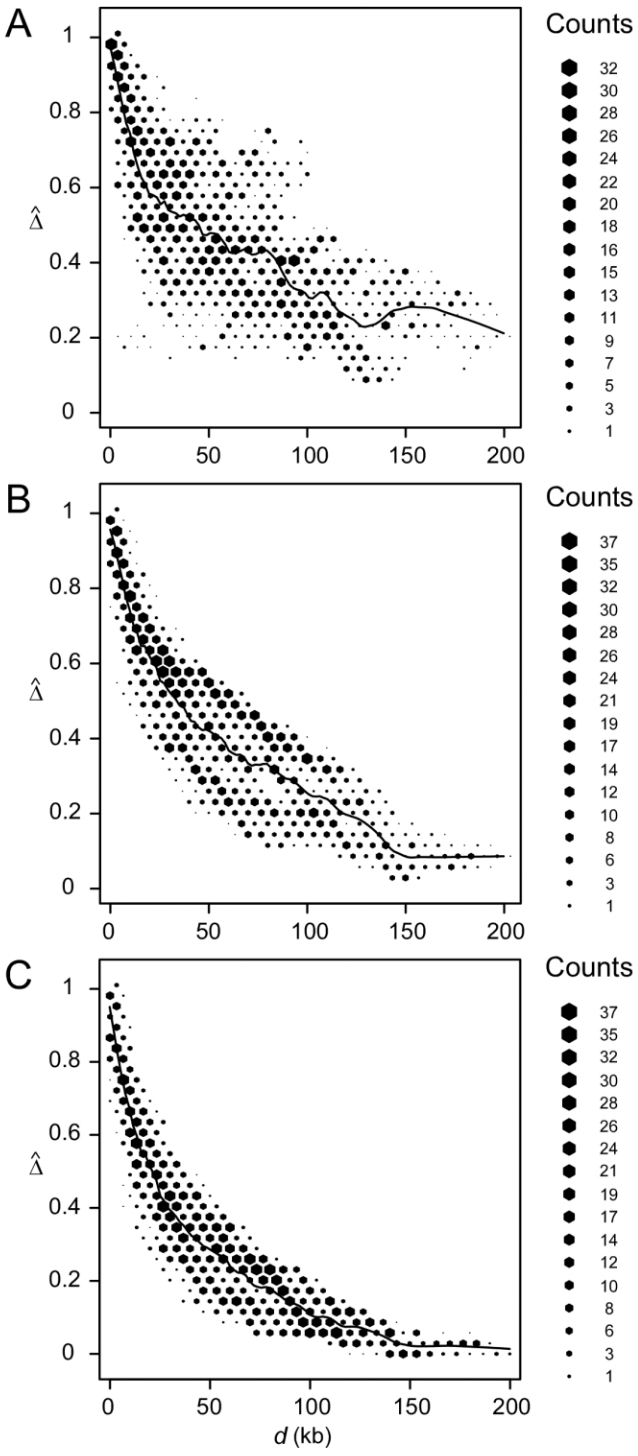

Figure 4 Application of $\hat{\Delta}$ to unphased genotype data. All 245 SNPs (*A*), 30 SNPs in each of the two flanking regions in addition to the 81 SNPs in the region of interest (*B*), and 10 SNPs in each of the two flanking regions, in addition to the 81 SNPs in the region of interest (*C*), are used for haplotype estimation before pairwise $\hat{\Delta}$ is calculated for the 81 SNPs and plotted against physical distance in each situation.

centages allowed us to monitor the stability of $\hat{\Delta}$ for a fixed physical distance *d,* since the marker density decreases from 20 to 2 SNPs per 10 kb. This random selection of markers produced similar distributions of MAFs across different subsets, so that the effects of marker density and marker-allele frequency (a potentially influential factor in the behavior of an LD measure) are not confounded. Figure 5*A* shows the distributions of $\hat{\Delta}$ for pairs of markers located 15–16 kb apart for different marker densities. Although $\hat{\Delta}$ tends to decrease slightly as marker density decreases, $\hat{\Delta}$ is generally robust to marker density. Similar patterns of robustness were observed for other values of the physical distance *d.*

The impact of marker density on the estimated rate of LD decay was also investigated. The process for selecting random subsets of the original markers was repeated 200 times for each marker density, and the LD-decay rate β was estimated each time (fig. 5*B*). In general, $\hat{\beta}$ appears to be fairly stable for marker densities of $\geqslant 6$ SNPs per 10 kb but not for the low density of 2 SNPs per 10 kb, because of the loss of precision for measuring the length of haplotype sharing. Note that, as the marker density decreases, the number of marker pairs decreases, so that $\hat{\beta}$ is estimated with larger variance.

Impact of Marker-Allele Frequency on $\hat{\Delta}$

Conventional two-point LD measures are very sensitive to marker-allele frequency. To investigate the sensitivity of $\hat{\Delta}$ to marker-allele frequency, we used subsets of SNPs with different MAFs from one simulated data set to calculate $\hat{\Delta}$ for pairs of markers in each SNP subset. The minimum MAFs in each subset were 0%, 1%, 5%, 10%, and 20%. The corresponding marker densities for each subset were approximately 23, 19, 12, 10, and 6 SNPs per 10 kb. In this range of marker densities, on the basis of the above results, Δ and its decay rate β can be robustly and reliably estimated. Clearly, the exponential relationship between $\hat{\Delta}$ and *d* and the low variability of $\hat{\Delta}$ at any given *d* are both maintained across the five SNP subsets. For pairs of SNPs located a certain distance away from each other—say, 10–11 kb—$\hat{\Delta}$ is fairly stable in terms of its median and interquartile range across subsets of SNPs with different MAFs (fig. 6). In contrast, $|D'|$ is more likely to be 1 and r^2 is more likely to be close to 0 when SNPs with lower MAFs are included in the analysis. Thus, both $|D'|$ and r^2 are highly sensitive to allele frequencies. Furthermore, the estimated rate of LD decay $\hat{\beta}$ is also very robust to SNP allele frequency, ranging from 0.009 to 0.011 across subsets of SNPs.

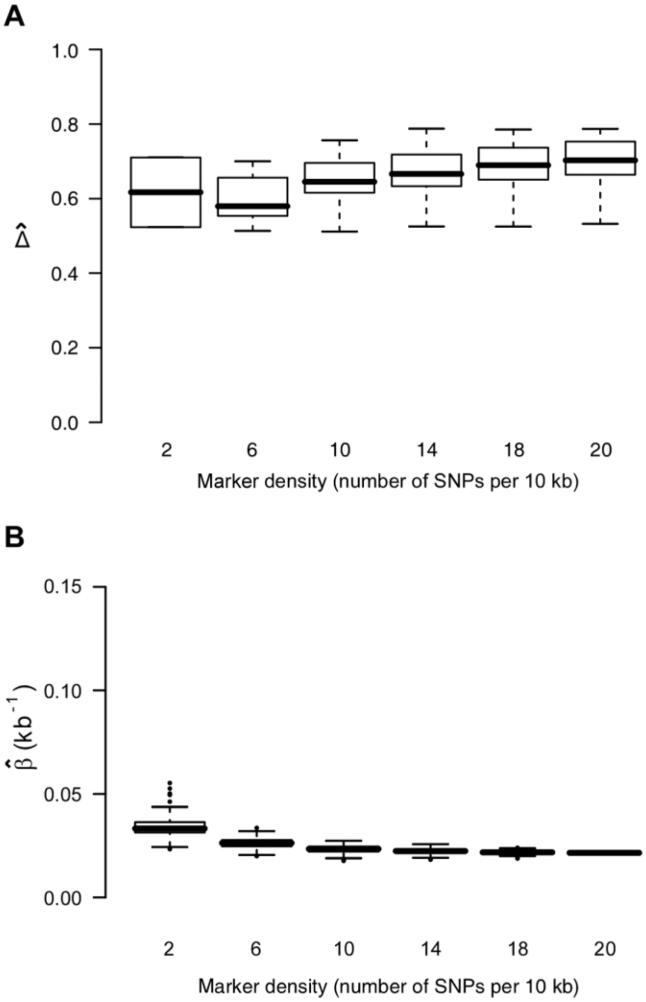

Figure 5 Robustness of $\hat{\Delta}$ and its estimated decay rate $\hat{\beta}$ to marker density. *A,* Distributions of $\hat{\Delta}$ for pairs of SNPs that are located 15–16 kb apart for different marker densities. *B,* Distributions of estimated LD-decay rates $\hat{\beta}$ for 200 SNP subsets randomly selected for each marker density. The scale of the vertical axis was chosen to match that in figure 7.

Relationship of $\hat{\Delta}$ with Recombination Rate

In the *ms* program, recombination rates for the simulated data can be controlled by crossover probabilities for adjacent base pairs. The following four values for the crossover probability were considered: 10^{-9}, 10^{-8}, 10^{-7}, and 10^{-6} per kb. For a fixed physical distance—say, d = 15–16 kb—$\hat{\Delta}$ decreases as the recombination rate increases (fig. 7*A*). Furthermore, LD-decay rates, estimated for 200 independently simulated data sets in each setting, were strongly related to recombination rates (fig. 7*B*).

Analysis of HapMap SNP Data for the X Chromosome

We applied Δ to HapMap data (phase I) for the X chromosomes of 30 mothers in the CEPH population

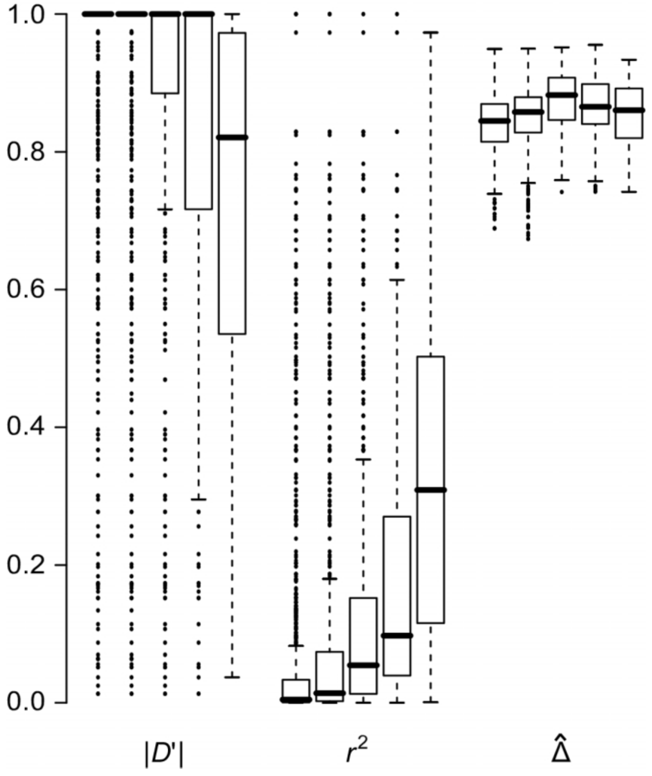

Figure 6 Robustness of LD measures to marker-allele frequency. Distributions of $|D'|$, r^2, and $\hat{\Delta}$ for SNP pairs that are located 10–11 kb apart, with use of subsets of SNPs with MAFs (*from left to right*) of at least 0%, 1%, 5%, 10%, and 20%, for each measure.

(Utah residents with ancestry from northern and western Europe). Genotypes were fully phased at 56,001 SNPs, among which there were 19,127 monomorphic SNPs. LD-decay rates β were estimated at every polymorphic SNP locus by applying the method of least squares to the exponential-decay model for $\hat{\Delta}$. Specifically, pairwise $\hat{\Delta}$ was computed from neighboring polymorphic SNPs within 100-kb windows, as long as there were $\geqslant 7$ polymorphic SNPs, so that the number of marker pairs used to estimate β was at least 21 (fig. 8*A*). The lengths of haplotype sharing were evaluated in 1.1-Mb regions surrounding every polymorphic SNP. The marker density was adequately high in 99% of these regions (>2 per 10 kb) to support reliable estimation of Δ and its decay rate β. The results show that LD on the X chromosome decays exponentially at an average rate of 0.0073 per kb within 100-kb windows, whereas, at certain loci, the rate can reach 0.054 per kb. Figure 8*B* provides a higher-resolution display for 100 polymorphic SNPs in the 12.89–13.17-Mb region. Pseudocolor images of pairwise $|D'|$, r^2, and $\hat{\Delta}$ matrices in this region are displayed in figure 8*D*–8*F*. Note the much "smoother" appearance of the pseudocolor image for $\hat{\Delta}$, which clearly suggests blocks of markers with high LD.

Actual genomic data are different from simulated data in one important aspect. The recombination rate can be fixed in the simulated data, whereas it varies greatly in the real data. Since recombination causes LD decay, the recombination rate is directly related to the LD-decay rate, as shown in the simulation studies above. Thus, we do not expect LD to decay at a homogeneous rate in the human genome. However, in the linear-regression model used to estimate the LD-decay rate, LD is assumed to decay at a constant rate within the region of interest. The result therefore reflects the rate at which LD decays, on average, over the region. We have chosen to estimate LD-decay rates on the basis of regions of only 100 kb, with the hope that the LD-decay rate does not change

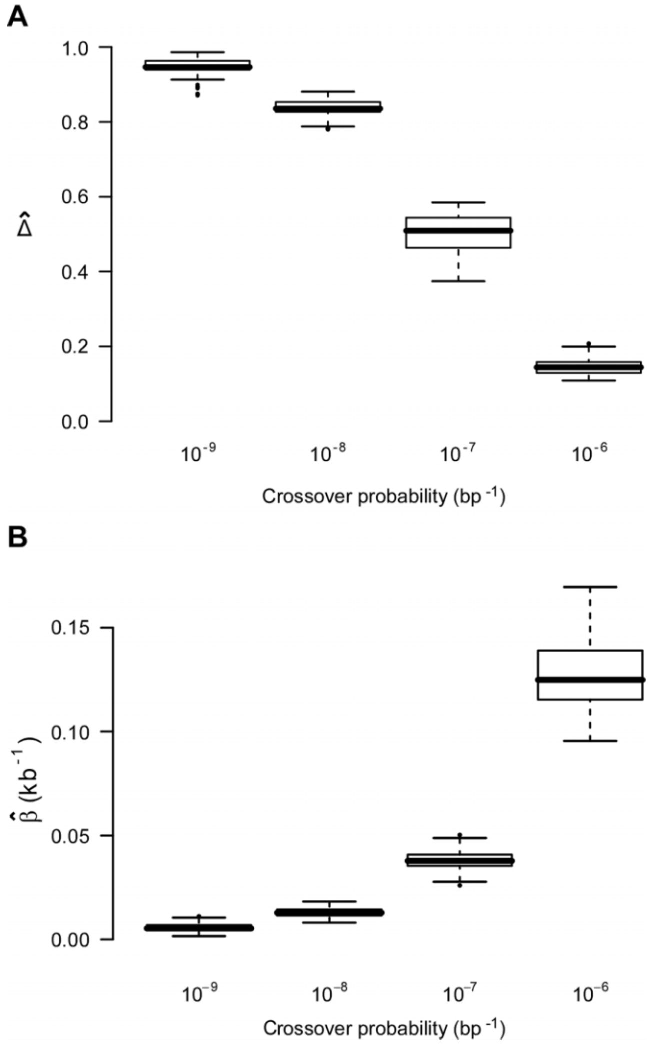

Figure 7 $\hat{\Delta}$ and its estimated decay rate $\hat{\beta}$ as a function of recombination rate. *A,* Distributions of $\hat{\Delta}$ for pairs of SNPs that are located 15–16 kb apart, for different crossover probabilities for adjacent base pairs. *B,* Distributions of estimated LD-decay rates $\hat{\beta}$, for 200 simulations at different crossover probabilities.

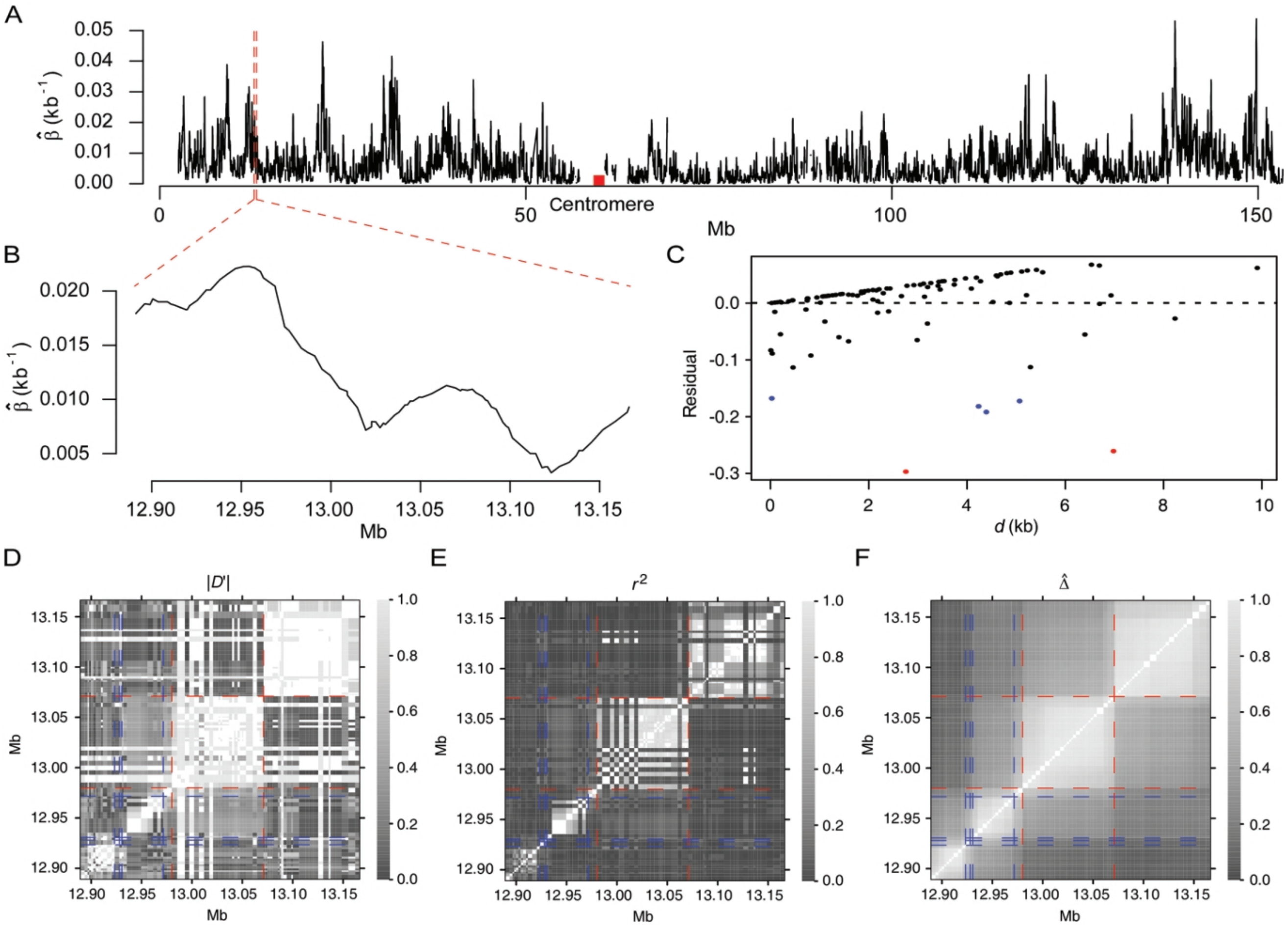

Figure 8 LD measures for HapMap data for 56,001 SNPs on the X chromosomes of 30 women in the CEPH population. *A,* LD-decay rates at every SNP locus, estimated from polymorphic SNPs in the neighboring 100-kb region. *B,* Display with higher resolution for an arbitrarily selected region of 275 kb. *C,* Residual plot from fitting the linear-regression model $E[\log\hat{\Delta}] = -\beta d$ with data from the 275-kb region. Adjacent marker pairs with large negative residuals are heuristically considered recombination hotspots and are plotted using colored dots, with red representing even more extreme residuals than blue. *D–F,* Pseudocolor images of pairwise $|D'|$, r^2, and $\hat{\Delta}$ matrices. The red and blue dashed lines correspond to the marker pairs plotted using red and blue dots in panel C.

dramatically within such relatively small regions. However, this assumption could still be violated because of recombination hotspots. A recombination hotspot is a site prone to recombination and is experimentally identified as a region as narrow as 1–2 kb, where recombination rates are higher than in neighboring regions (Jeffreys et al. 2001). Therefore, LD decays faster across such a hotspot. If a smaller window size is used, the reduced number of markers may be insufficient for stable parameter estimation in the regression model. More methodological work concerning the development of indices for the investigation of fine-scale LD is needed. We anticipate that the new LD measure Δ will make valuable contributions to this endeavor.

As an example of the usefulness of our new LD measure, a heuristic analysis of data for the 275-kb region in figure 8*B* suggests that recombination hotspots may be identified as follows on the basis of $\hat{\Delta}$. We focus on all adjacent marker pairs in the region of interest, as long as $\hat{\Delta}$ can be calculated on the basis of markers with density higher than 2 SNPs per 10 kb. The tight relationship between $\hat{\Delta}$ and physical distance d is expected to be maintained for these marker pairs. Under the assumption that LD decays at the same rate across all adjacent marker pairs, the regression model $E[\log\hat{\Delta}] = -\beta d$ was fit. Outlier adjacent marker pairs, with unexpectedly small residuals (i.e., large negative values), can be considered as recombination hotspots and identified through model diagnostic techniques. However, usual model diagnostic techniques are not applicable here because of the dependence of $\hat{\Delta}$ between adjacent marker pairs (as shown by the residual plot in fig. 8*C*). In this article, we do not intend to address in depth the issue of outlier detection. Instead, we graphically illustrate that adjacent marker pairs with extreme negative residuals (plotted by the red and blue dots in fig. 8*C*) correspond to potential recombination hotspots (indicated by the red and blue lines, respectively, in fig. 8*D*–8*F*).

Discussion

The proposed LD measure Δ is based on the unobservable lengths of common ancestral segments, which are approximated by shared haplotype lengths. The degree of precision for this approximation, influenced by several factors, directly affects estimation of Δ. Here, we examine the following factors one by one: distinction between IBD and IBS status, marker density, and censoring of shared haplotype lengths.

First, the length of a common ancestral segment is best measured on the basis of alleles IBD for a chromosome pair. However, in practice, it is often impossible to distinguish between IBD and IBS. Here, we argue that $\hat{\Delta}^U$ and $\hat{\Delta}$ remain robust to discrepancies between IBD and IBS. In the presence of alleles IBS for a long sequence of contiguous loci, the probability of IBD at each locus is greatly elevated and so is the probability that these loci belong to a common ancestral segment. The larger the length of haplotype sharing by state, the higher the probability of IBD. On the other hand, for chromosome pairs that do not share common ancestral segments, the probability of sharing alleles IBS at a long sequence of contiguous loci is very small. We do not expect the background level of haplotype sharing due to IBS to have a significant effect on $\hat{\Delta}^U$ and $\hat{\Delta}$, because these estimators are mostly determined by large shared haplotype lengths at both loci, which are more likely to be due to IBD. Therefore, $\hat{\Delta}^U$ and $\hat{\Delta}$ should be robust to the approximation of IBD by IBS.

Second, higher marker densities lead to better approximation of the lengths of common ancestral segments by the lengths of shared haplotypes. On the basis of our simulation studies, the impact of marker density on estimation of Δ is very limited once this density is above a certain threshold—namely, 2 SNPs per 10 kb—which is feasible given the imminent availability of ultrahigh-volume genotyping platforms. Note that one need not identify tagging SNPs when markers are used for the purpose of tracking the length of haplotype sharing. In fact, subsetting SNPs does not enhance but impairs accurate evaluation of the length of haplotype sharing because of reduced marker density.

Third, censoring at the edge of the genotyped region is an important practical issue to be considered. For a region of relatively small size, the length of haplotype sharing may not be observed to its full extent for some chromosome pairs that share extensively long common haplotypes. For genome-scan data, the same problem is present when evaluating the length of haplotype sharing for a reference locus close to a telomere or when dealing with phase-unknown data with a moderate number of markers used for haplotype inference. This phenomenon is very similar to censoring for survival time and may bias $\hat{\Delta}^U$ and $\hat{\Delta}$. Further research is needed to adjust these estimators if censoring is involved at one or both markers. For the time being, we recommend that caution be taken for small genotyped regions and that $\hat{\Delta}^U$ or $\hat{\Delta}$ be calculated only if flanking regions of decent sizes are also genotyped. Just as there exists a threshold for marker density above which $\hat{\Delta}$ stabilizes, there is such a threshold for the size of flanking regions. Adequate flanking region sizes are usually determined by how fast LD decays in these regions. For instance, when LD decays fast, smaller flanking regions are considered adequate. For HapMap X-chromosome data, the lengths of haplotype sharing were calculated using 500-kb flanking regions on both sides of the reference locus. In addition, to avoid censoring around telomeres, we used only the correlation coefficient for the right-sided or left-sided lengths of haplotype sharing for markers around the left or right, respectively, telomere.

Finally, we address the connections and differences between recombination hotspots and boundaries for haplotype blocks, since the latter have become accepted as a general model for LD patterns throughout the genome. Both terms describe patterns of LD and were often used interchangeably in the past. For instance, Anderson and Novembre (2003) evaluated their method for identifying block boundaries by simulation studies in which recombination hotspots were generated as block boundaries. From the example in the "Results" section, the identified hotspots seemingly are good candidates for block boundaries. However, the two terms refer to different phenomena, and different methods may be required in practice to detect them. For recombination hotspots at which LD decays faster than in other regions, LD decay rate is an important aspect because physical distance plays an essential role. In the HapMap data analysis, hotspots were identified on the basis of residuals for a fitted exponential-decay model for $\hat{\Delta}$ and d instead of on the basis of $\hat{\Delta}$ only. In contrast, block boundaries are traditionally chosen to achieve low haplotype diversity within each block, on the basis of significantly low LD values, without taking physical distance into consideration.

In conclusion, simulation studies and analysis of HapMap data demonstrate that our proposed LD measure Δ and its estimators $\hat{\Delta}^U$ and $\hat{\Delta}$ are superior to two of the most popular two-point LD measures, in terms of their relationship with physical distance, their small variability at any given distance, and their robustness to

SNP allele frequencies. In contrast to alternative LD measures that are based on population genetics models, Δ is a robust empirical measure and should be applicable regardless of population structure. A definition of LD-decay rate and a regression-based method for estimating such rates were proposed, and simulation studies demonstrated that the LD-decay rate was a function of the recombination rate. The new LD measure Δ is a promising tool for studying population genetics and for mapping complex disease genes. Our proposed methods can also be readily applied to data for more polymorphic DNA markers (e.g., microsatellites) or amino acid sequence data without further extension.

Acknowledgments

This work was partially supported by National Institutes of Health (NIH) training grant 5 T32 HG00047, NIH grant RO1 LM07609, and NIH grant 5RO1 CA106320-05. We thank Ying Qing Chen, Sue Li, and Najma Khalid, for helpful discussions.

Appendix A

Proof of the Proposition

Define $X_{ij} = \min(S_i,S_j)$ and $Y_{ij} = \min(T_i,T_j)$. Since $(S_1,T_1),\ldots,(S_4,T_4)$ are iid, we have $E(X_{12}Y_{12}) = E(X_{34}Y_{34})$ and $E(X_{12}Y_{34}) = E(X_{34}Y_{12}) = E(X_{12})E(Y_{12})$. Then,

$$\begin{aligned} E\{[\min(S_1,S_2) - \min(S_3,S_4)][\min(T_1,T_2) - \min(T_3,T_4)]\} &= E[(X_{12} - X_{34})(Y_{12} - Y_{34})] \\ &= E(X_{12}Y_{12} - X_{12}Y_{34} - X_{34}Y_{12} + X_{34}Y_{34}) \\ &= 2E(X_{12}Y_{12}) - 2E(X_{12})E(Y_{12})\ . \end{aligned}$$

Similarly,

$$\begin{aligned} E\{[\min(S_1,S_3) - \min(S_2,S_4)][\min(T_1,T_3) - \min(T_2,T_4)]\} &= E\{[\min(S_1,S_4) - \min(S_2,S_3)] \\ &\quad \times[\min(T_1,T_4) - \min(T_2,T_4)]\} \\ &= 2E(X_{12}Y_{12}) - 2E(X_{12})E(Y_{12})\ . \end{aligned}$$

Therefore, $E\{\psi[(S_1,T_1),\ldots,(S_4,T_4)]\} = E(X_{12}Y_{12}) - E(X_{12})E(Y_{12})$ is the covariance of X and Y.

Appendix B

Two Estimators of Δ

Suppose that, among a random sample of n chromosomes, there are m distinct haplotypes $\{h_i{:}i = 1,\ldots,m\}$ for a region that covers two loci of interest, A and B. The haplotypes $\{h_i\}$ follow a multinomial distribution with parameters (n,θ), where $\theta = \{\theta_i{:}i = 1,\ldots,m\}$ are either empirical or inferred haplotype frequencies. Let $\mathbf{X} = \{X_{ij}{:}i,j = 1,\ldots,m,i<j\}$ and $\mathbf{Y} = \{Y_{ij}{:}i,j = 1,\ldots,m,i<j\}$ denote the pairwise lengths of one-sided shared haplotypes for loci A and B, respectively. Similarly, let $\mathbf{X}'$ and $\mathbf{Y}'$ denote the lengths of shared haplotypes on the other sides of loci A and B. The following two estimators of Δ are both arithmetic means of correlation-coefficient estimators $\hat{\rho}_{xy}$ and $\hat{\rho}_{x'y'}$, based on two different estimation approaches.

1. For the U-statistic–based estimator $\hat{\Delta}^U$, define functions

$$\hat{\psi}_{xy}(i_1,i_2,i_3,i_4) = \frac{1}{6}[(X_{i_1i_2} - X_{i_3i_4})(Y_{i_1i_2} - Y_{i_3i_4}) + (X_{i_1i_3} - X_{i_2i_4})(Y_{i_1i_3} - Y_{i_2i_4}) + (X_{i_1i_4} - X_{i_2i_3})(Y_{i_1i_4} - Y_{i_2i_3})]\ ,$$

$$\hat{\psi}_x(i_1,i_2,i_3,i_4) = \frac{1}{6}[(X_{i_1i_2} - X_{i_3i_4})^2 + (X_{i_1i_3} - X_{i_2i_4})^2 + (X_{i_1i_4} - X_{i_2i_3})^2] \ ,$$

and

$$\hat{\psi}_y(i_1,i_2,i_3,i_4) = \frac{1}{6}[(Y_{i_1i_2} - Y_{i_3i_4})^2 + (Y_{i_1i_3} - Y_{i_2i_4})^2 + (Y_{i_1i_4} - Y_{i_2i_3})^2] \ .$$

Then,

$$\hat{\rho}^U_{xy} = \binom{m}{4}^{-1} \sum_{(m,4)} \frac{w^U(i_1,i_2,i_3,i_4)\hat{\psi}_{xy}(i_1,i_2,i_3,i_4)}{\sqrt{\hat{\sigma}^U_x\hat{\sigma}^U_y}} \ ,$$

where

$$\hat{\sigma}^U_x = \binom{m}{4}^{-1} \sum_{(m,4)} w^U(i_1,i_2,i_3,i_4)\hat{\psi}_x(i_1,i_2,i_3,i_4)$$

and

$$\hat{\sigma}^U_y = \binom{m}{4}^{-1} \sum_{(m,4)} w^U(i_1,i_2,i_3,i_4)\hat{\psi}_y(i_1,i_2,i_3,i_4) \ ,$$

with the weight function $w^U(i_1,i_2,i_3,i_4)$ proportional to $\theta_{i_1}\theta_{i_2}\theta_{i_3}\theta_{i_4}$.

2. For the naive estimator $\hat{\Delta}$,

$$\hat{\rho}_{xy} = \binom{m}{2}^{-1} \sum_{(m,2)} \frac{w(i,j)(X_{ij} - \bar{X})(Y_{ij} - \bar{Y})}{\sqrt{\hat{\sigma}_x\hat{\sigma}_y}} \ ,$$

where $\bar{X}$ and $\bar{Y}$ denote the sample means for **X** and **Y**, and

$$\hat{\sigma}_x = \binom{m}{2}^{-1} \sum_{(m,2)} w(i,j)(X_{ij} - \bar{X})^2$$

and

$$\hat{\sigma}_y = \binom{m}{2}^{-1} \sum_{(m,2)} w(i,j)(Y_{ij} - \bar{Y})^2 \ ,$$

with the weight function $w(i,j)$ proportional to $\theta_i\theta_j$.

References

Anderson EC, Novembre J (2003) Finding haplotype block boundaries by using the minimum-description-length principle. Am J Hum Genet 73:336–354

Dawson E, Abecasis GR, Bumpstead S, Chen Y, Hunt S, Beare DM, Pabial J, et al (2002) A first-generation linkage disequilibrium map of human chromosome 22. Nature 418:544–548

Devlin B, Risch N (1995) A comparison of linkage disequilibrium measures for fine-scale mapping. Genomics 29:311–322

Excoffier L, Slatkin M (1995) Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. Mol Biol Evol 12:921–927

Hawley ME, Kidd KK (1995) HAPLO: a program using the EM algorithm to estimate the frequencies of multi-site haplotypes. J Hered 86:409–411

Hinds DA, Stuve LL, Nilsen GB, Halperin E, Eskin E, Ballinger DG, Frazer KA, Cox DR (2005) Whole-genome patterns of common DNA variation in three human populations. Science 307:1072–1079

Hudson RR (2002) Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics 18:337–338

International HapMap Consortium (2003) The International HapMap Project. Nature 426:789–796

Jeffreys AJ, Kauppi L, Neumann R (2001) Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. Nat Genet 29:217–222

Lee AJ (1990) U-statistics: theory and practice. Marcel Dekker, New York

Lewontin RC, Kojima K (1960) The evolutionary dynamics of complex polymorphisms. Evolution 14:458–472

Li SS, Khalid N, Carlson C, Zhao LP (2003) Estimating haplotype frequencies and standard errors for multiple single nucleotide polymorphisms. Biostatistics 4:513–522

Long JC, Williams RC, Urbanek M (1995) An EM algorithm and testing strategy for multiple-locus haplotypes. Am J Hum Genet 56:799–810

McPeek MS, Strahs A (1999) Assessment of linkage disequilibrium by the decay of haplotype sharing, with application to fine-scale genetic mapping. Am J Hum Genet 65:858–875

Morton NE, Zhang W, Taillon-Miller P, Ennis S, Kwok PY, Collins A (2001) The optimal measure of allelic association. Proc Natl Acad Sci USA 98:5217–5221

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

Xiong M, Guo SW (1997) Fine-scale mapping based on linkage disequilibrium: theory and applications. Am J Hum Genet 60:1513–1531