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

Yan Wang∗ Lue Ping Zhao† Sandrine Dudoit‡

∗Division of Biostatistics, School of Public Health, University of California, Berkeley, yanw@stat.berkeley.edu
†Quantitative Genetic Epidemiology, Fred Hutchinson Cancer Research Center, Seattle, WA, lzhao@zebra.fhcrc.org
‡Division of Biostatistics, School of Public Health, University of California, Berkeley, sandrine@stat.berkeley.edu

This working paper is hosted by The Berkeley Electronic Press (bepress) and may not be commercially reproduced without the permission of the copyright holder.

http://biostats.bepress.com/ucbbiostat/paper192
Copyright ©2005 by the authors.
A Fine-Scale Linkage Disequilibrium Measure Based on Length of Haplotype Sharing

Yan Wang, Lue Ping Zhao, and Sandrine Dudoit

Abstract

High-throughput genotyping technologies for single nucleotide polymorphisms (SNP) have enabled the recent completion of the International HapMap Project (Phase I), which has stimulated much interest in studying genome-wide linkage disequilibrium (LD) patterns. Conventional LD measures, such as D’ and r-square, are two-point measurements, and their relationship with physical distance is highly noisy. We propose a new LD measure, 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 the new LD measure, which takes into consideration the dependence structure of the observed data, is developed and compared to a naive estimator based on the usual empirical correlation coefficient. Furthermore, we propose methods for inferring LD decay rates based on the new LD measure. The results from coalescent simulation studies and analysis of HapMap SNP data demonstrate that the proposed new LD measure and its estimators are superior to the two most popular conventional LD measures, in terms of their 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 investigating recombination mechanisms based on better-quantified LD.
Introduction

Linkage disequilibrium (LD) refers to the association of alleles at loci on the same chromosome (Lewontin and Kojima 1960). Such allelic associations are mostly due to physical adjacency, 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 references therein). 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 single nucleotide polymorphisms (SNP) have stimulated much interest in studying fine-scale genome-wide patterns of common DNA variations, using > 1 million SNPs from a number of human populations (The International HapMap Consortium 2003; Hinds et al. 2005).

While 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, based only on the allele distributions at these two loci, without exploiting information on the allele distributions of and physical distances from neighboring loci. Despite their popularity, $D'$ and $r^2$ (Box 1) 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 the fraction of common SNPs.
that are in high LD (say, \(r^2 > 0.8\)) (Hinds et al. 2005), in a region of empirically chosen size.

**Box 1. Conventional LD Measures \(D, D', \text{ and } r^2.\)**

Given a pair of biallelic markers \(A\) (with alleles \(A_0\) and \(A_1\)) and \(B\) (with alleles \(B_0\) and \(B_1\)), let \(p_{ij}\) denote the frequency of haplotype \(A_iB_j, i, j \in \{0, 1\}\). Then, \(p_i = \sum_{j=0}^{1} p_{ij}\) and \(p_j = \sum_{i=0}^{1} p_{ij}\) denote the marginal frequencies of allele \(A_i\) at locus \(A\) and allele \(B_j\) at locus \(B\), respectively. The LD measure \(D\) is defined by

\[
D = p_{00} - p_0 \cdot p_0.
\]

The LD measure \(D'\) (Lewontin 1964) is defined by

\[
D' = \frac{p_{00} - p_0 \cdot p_0}{D_{\text{max}}},
\]

where \(D_{\text{max}}\) is the largest value of \(D\), given the marginal frequencies:

\[
D_{\text{max}} = \begin{cases} 
\min\{p_0 \cdot p_1, p_1 \cdot p_0\}, & \text{if } D \geq 0 \\
\min\{p_0 \cdot p_0, p_1 \cdot p_1\}, & \text{if } D < 0.
\end{cases}
\]

The LD measure \(r^2\) (Hill and Robertson 1968; Franklin and Lewontin 1970) is defined by

\[
r^2 = \frac{(p_{00} - p_0 \cdot p_0)^2}{p_0 \cdot p_0 \cdot p_1 \cdot p_1}.
\]

With the aim of better quantifying LD, several new measures have been proposed, based on population genetic models. Morton et al. (2001) propose an association probability (LDU) between a pair of loci, under population genetic assumptions regarding recombination, mutation, migration, etc. Other measures do not directly quantify LD in the usual two-locus manner. Instead, LD is assessed as a function 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 that theoretically is 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 genome-wide LD patterns and the limitations of existing measures, we propose a new LD measure, $\Delta$, which borrows information from multiple neighboring loci and 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 between the ancestor and the current generation. 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 lead to a sensible and stable measure of association between two loci.

In the Methods section, we first define the new LD measure $\Delta$ 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 $\Delta$, $\hat{\Delta}^U$, which takes into account the dependence structure of the observed lengths of shared haplotypes for pairs of chromosomes. An alternative estimator $\hat{\Delta}$, which 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 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 for estimating LD decay rate is proposed based on the tight relationship between $\hat{\Delta}$ and physical distance $d$. Then, merits of the new LD measure $\Delta$ are demonstrated by analyzing human
X-chromosome SNP data from the HapMap project. We close with a discussion on issues regarding evaluation of the lengths of common ancestral segments.

**Methods**

*A New LD Measure: $\Delta$*

Figure 1 shows the conceptual model which motivates the definition of the LD parameter $\Delta$. For a pair of chromosomes that share a common ancestor around locus $A$, we denote the lengths of the ancestral segments from locus $A$ to their respective breakpoints on one side (right side, say) by random variables $S_1$ and $S_2$. Given a locus $B$, located to the right of $A$ with 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, what one may observe are the lengths of haplotypes shared by the chromosome pair, which approximate the lengths of the shared common ancestral segments, denoted by $X \approx \min(S_1, S_2)$ and $Y \approx \min(T_1, T_2)$. These lengths may be measured either by physical distance, i.e., the number of base-pairs (in units of bp or kb), or by genetic distance (in units of cM). However, the former is more precise and relevant because the most appropriate type of data for the proposed method is that of dense sets of markers (see the Results and Discussion sections regarding marker density).

Two assumptions are involved in approximating the lengths of the 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 decent (IBD). This may appear strong, yet, the Discussion section shows that violation of this assumption does not impact much on the new LD measure. Some other concerns
related to this approximation are further discussed later.

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 identical by decent and by state (under the assumption of no mutation). In this case, A and B are in complete linkage disequilibrium, 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. 1A). On the other hand, when chromosomal segments around loci A and B are inherited from two independent (unrelated) ancestors (blue and red), 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. 1B). Therefore, $\rho_{xy}$ quantifies the magnitude of LD between A and B.

The same situation applies to the lengths of shared haplotypes from the reference loci to the other direction, i.e., $X'$ and $Y'$ in figure 1, but the relationship between $X + X'$ and $Y + Y'$ is more complex. Therefore, we treat the lengths of haplotype sharing to the right and left side of the reference loci separately in our method. The new LD measure $\Delta$ is proposed as the arithmetic mean of $\rho_{xy}$ and $\rho_{x'y'}$.

Statistically there is a challenge in estimating $\Delta$ due to the dependence structure of observed lengths of shared haplotypes between pairs of chromosomes. In the following subsections, an estimator of $\Delta$ is developed based on unbiased U-statistics (Lee 1990). We first consider the simplest scenario, where the sampled haplotypes are distinct by state, which is the case in practice when the population or the number of markers is large enough. Then, we extend the method to the general situation where haplotypes are not necessarily distinct.
**An Estimator Based on U-statistics: \( \hat{\Delta}^U \) for Distinct Haplotypes**

Suppose that \( n \) chromosomes are independently sampled with equal probability from a population, on which \( n \) distinct haplotypes are observed as \( \{h_i : i = 1, \ldots, n\} \). It is assumed that the unobservable ancestral segment lengths \( (S_i, T_i), i = 1, 2, \ldots, n \), are independently and identically distributed (i.i.d.) with cumulative distribution function \( F(S, T), S \geq 0, T \geq 0 \). From \( \{h_i\} \), one observes \( X = \{X_{ij} : i, j = 1, \ldots, n, i < j\} \) and \( Y = \{Y_{ij} : i, j = 1, \ldots, n, i < j\} \) as the pairwise lengths of one-sided shared haplotypes for loci \( A \) and \( B \), respectively, where \( (i, j) \) index the \( \binom{n}{2} \) distinct pairs of haplotypes. Based on 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 \( \{X_{ij}\} \) nor the \( \{Y_{ij}\} \) are sets 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 i.i.d. 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, page 7), the variances and covariance of \( X \) and \( Y \) can be estimated by the average kernels, termed U-statistics due to their unbiasedness. The correlation coefficient of \( X \) and \( Y \) is then estimated by the estimated covariance standardized by the estimated standard deviations of \( X \) and \( Y \).

**Proposition.** Let \( \sigma_{xy} \) be a statistical functional of degree 4 with kernel function \( \psi \). That is, define \( \sigma_{xy} \) as

\[
\sigma_{xy} = E[\psi((S_1, T_1), \ldots, (S_4, T_4))] = \int_0^\infty \cdots \int_0^\infty \psi((s_1, t_1), \ldots, (s_4, t_4)) \prod_{i=1}^4 dF(s_i, t_i), \quad (1)
\]
where the kernel function is

\[
\psi((S_1, T_1), \ldots, (S_4, T_4)) = \frac{1}{6} \left[ \{\min(S_1, S_2) - \min(S_3, S_4)\}\{\min(T_1, T_2) - \min(T_3, T_4)\} \\
+ \{\min(S_1, S_3) - \min(S_2, S_4)\}\{\min(T_1, T_3) - \min(T_2, T_4)\} \\
+ \{\min(S_1, S_4) - \min(S_2, S_3)\}\{\min(T_1, T_4) - \min(T_2, T_3)\}\right]
\]

and \((S_i, T_i), i = 1, \ldots, n\), are i.i.d. with cumulative distribution function \(F(S, T)\). Then, for \(X \equiv \min(S_1, S_2)\) and \(Y \equiv \min(T_1, T_2)\), \(\sigma_{xy}\) 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 \(\sigma_{xy}\) has the form of a U-statistic,

\[
\hat{\sigma}_{xy}^U = \left(\frac{n}{4}\right)^{-1} \sum_{(n,4)} \psi((S_{i_1}, T_{i_1}), \ldots, (S_{i_4}, T_{i_4})),
\]

where the sum \(\sum_{(n,4)}\) is taken over all distinct four-element subsets \(\{i_1, i_2, i_3, i_4\}\) from \(\{1, \ldots, n\}\). The unobservable random variables \((S_i, T_i)\) in the kernel function \(\psi\) are then approximated by the corresponding observable random variables. For example, \(\min(S_{i_1}, S_{i_2})\) is replaced by \(X_{i_1i_2}\), etc. Hence, the kernel function can be written as

\[
\psi((S_{i_1}, T_{i_1}), \ldots, (S_{i_4}, T_{i_4}))
= \frac{1}{6} \left[ (X_{i_1i_2} - X_{i_3i_4})(Y_{i_1i_2} - Y_{i_3i_4}) + (X_{i_1i_3} - X_{i_2i_4})(Y_{i_1i_3} - Y_{i_2i_4}) + (X_{i_1i_4} - X_{i_2i_3})(Y_{i_1i_4} - Y_{i_2i_3}) \right].
\]

Denote the variances of \(X\) and \(Y\) by \(\sigma_x\) and \(\sigma_y\), respectively. These are also statistical
functionals of degree 4:

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

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 express \( \sigma_y \) and \( \psi_y \) likewise for the variance of \( Y \).

Then, the unique unbiased estimators for \( \sigma_x \) and \( \sigma_y \) are both U-statistics:

$$\hat{\sigma}_x^U = \left(\frac{n}{4}\right)^{-1} \sum_{(n,4)} \psi_x(S_{i_1}, \ldots, S_{i_4}); \quad (5)$$

$$\hat{\sigma}_y^U = \left(\frac{n}{4}\right)^{-1} \sum_{(n,4)} \psi_y(T_{i_1}, \ldots, T_{i_4}), \quad (6)$$

where the kernel functions become

$$\psi_x(S_{i_1}, \ldots, S_{i_4}) = \frac{1}{6} \left\{ (X_{i_1i_2} - X_{i_3i_4})^2 + (X_{i_1i_3} - X_{i_2i_4})^2 + (X_{i_1i_4} - X_{i_2i_3})^2 \right\}, \quad (7)$$

$$\psi_y(T_{i_1}, \ldots, T_{i_4}) = \frac{1}{6} \left\{ (Y_{i_1i_2} - Y_{i_3i_4})^2 + (Y_{i_1i_3} - Y_{i_2i_4})^2 + (Y_{i_1i_4} - Y_{i_2i_3})^2 \right\}. \quad (8)$$

The correlation of \( X \) and \( Y \), \( \rho_{xy} \), is the covariance \( \sigma_{xy} \) standardized by the standard
deviations of \( X \) and \( Y \). A reasonable estimator of \( \rho_{xy} \) is then

\[
\hat{\rho}_{xy}^U = \frac{\hat{\sigma}_{xy}^U}{\sqrt{\hat{\sigma}_{x}^U \hat{\sigma}_{y}^U}}.
\]  

(9)

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 side. An estimator of \( \Delta, \hat{\Delta}^U \), is then the arithmetic mean of \( \hat{\rho}_{xy}^U \) and \( \hat{\rho}_{x'y'}^U \). This measure has reduced variance compared to the two individual correlation coefficients (unpublished results). While theoretically correlation coefficients range from -1 to 1, \( \hat{\Delta}^U \) values are seldom negative in our numerical studies. Negative values may occur due to stochastic variations around the true value of zero. In practice, those negative values can be converted to zero.

An Estimator Based on Weighted \( U \)-statistics: \( \hat{\Delta}^U \) for Non-Distinct Haplotypes

Next, consider the case where the \( n \) observed haplotypes are not necessarily distinct. Suppose there are \( m \) distinct haplotypes \( \{h_i : i = 1, \ldots, m\} \), which follow a multinomial distribution \( (n, \theta) \), where \( \theta = \{\theta_i : i = 1, \ldots, m\} \) 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

\[
w^U(i_1, i_2, i_3, i_4) = \frac{24\theta_{i_1}\theta_{i_2}\theta_{i_3}\theta_{i_4}}{W^U},
\]

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, page 64) implies that unbiased estimators for the variances and
covariance of $X$ and $Y$ may be obtained from U-statistics weighted by $w^U$:

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

(10)

\[
\hat{\sigma}_{x}^U = \left( \frac{m}{4} \right)^{-1} \sum_{(m,4)} w^U(i_1, i_2, i_3, i_4) \psi_x(S_{i_1}, \ldots, S_{i_4}),
\]

(11)

\[
\hat{\sigma}_{y}^U = \left( \frac{m}{4} \right)^{-1} \sum_{(m,4)} w^U(i_1, i_2, i_3, i_4) \psi_y(T_{i_1}, \ldots, T_{i_4}).
\]

(12)

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 $\{\theta_i\}$ are inferred through, for instance, the EM algorithm.

**An Alternative Estimator: $\hat{\Delta}$**

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

In the case of $n$ distinct haplotypes,

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

(13)

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 $X$ and $Y$, respectively.

In the case of non-distinct 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,

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

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

\[
\hat{\sigma}_x = \left(\frac{m}{2}\right)^{-1} \sum_{(m, 2)} w(i, j)(X_{ij} - \bar{X})^2,
\]

\[
\hat{\sigma}_y = \left(\frac{m}{2}\right)^{-1} \sum_{(m, 2)} w(i, j)(Y_{ij} - \bar{Y})^2.
\]

Therefore, we propose a computationally simpler estimator \(\hat{\Delta}\) as the average of \(\hat{\rho}_{xy}\) and \(\hat{\rho}_{x'y'}\). Simulation studies show that \(\hat{\Delta}\) serves as a good approximation for \(\hat{\Delta}^U\) (see Results section). The two estimators are summarized in box 2.

\section*{Results}

In this article, we carried out a series of simulation studies, based on genotype data generated by the “ms” program (Hudson 2002), to investigate properties of the new LD measure \(\Delta\) and its estimators, \(\hat{\Delta}^U\) and \(\hat{\Delta}\). We focused for simplicity on fully phased data, but note that our proposed methods also apply to unphased data.

\subsection*{Comparison of \(\hat{\Delta}^U\) to its Approximation \(\hat{\Delta}\) and Impact of Sample Size}

First we studied how well the two estimators, \(\hat{\Delta}^U\) and \(\hat{\Delta}\), estimate \(\Delta\). As the underlying parameter value \(\Delta\) cannot be explicitly specified in the simulations with “ms”,...
Box 2. Two Estimators of the New LD Measure $\Delta$.

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 $(n, \theta)$, where $\theta = \{\theta_i : i = 1, \ldots, m\}$ are either empirical or inferred haplotype frequencies. Let $X = \{X_{ij} : i, j = 1, \ldots, m, i < j\}$ and $Y = \{Y_{ij} : i, j = 1, \ldots, m, i < j\}$ denote the pairwise lengths of one-sided shared haplotypes for loci $A$ and $B$, respectively. Similarly, let $X'$ and $Y'$ denote the length of shared haplotypes on the other sides of loci $A$ and $B$. The following two estimators of $\Delta$ are both arithmetic means of correlation coefficient estimators $\hat{\rho}_{xy}$ and $\hat{\rho}_{xy'}$, based on two different estimation approaches.

1. For the U-statistic-based estimator $\hat{\Delta}^U$: Define functions

$$\psi_1(i_1, i_2, i_3, i_4) = \frac{1}{6} \left\{ (X_{i_1i_2} - X_{i_3i_4})(Y_{i_1i_2} - Y_{i_3i_4}) ight.\right. + (X_{i_1i_3} - X_{i_2i_4})(Y_{i_1i_3} - Y_{i_2i_4}) + (X_{i_1i_4} - X_{i_2i_3})(Y_{i_1i_4} - Y_{i_2i_3}) \left. \right\},$$

$$\psi_2(i_1, i_2, i_3, i_4) = \frac{1}{6} \left\{ (X_{i_1i_2} - X_{i_3i_4})^2 + (X_{i_1i_3} - X_{i_2i_4})^2 + (X_{i_1i_4} - X_{i_2i_3})^2 \right\},$$

$$\psi_3(i_1, i_2, i_3, i_4) = \frac{1}{6} \left\{ (Y_{i_1i_2} - Y_{i_3i_4})^2 + (Y_{i_1i_3} - Y_{i_2i_4})^2 + (Y_{i_1i_4} - Y_{i_2i_3})^2 \right\}.$$

Then,

$$\hat{\rho}_{xy}^U = \left(\frac{m}{4}\right)^{-1} \sum_{(m,4)} w_{ij}(i_1, i_2, i_3, i_4)\psi_1(i_1, i_2, i_3, i_4),$$

$$\hat{\sigma}_x^U = \left(\frac{m}{4}\right)^{-1} \sum_{(m,4)} w_{ij}(i_1, i_2, i_3, i_4)\psi_2(i_1, i_2, i_3, i_4),$$

$$\hat{\sigma}_y^U = \left(\frac{m}{4}\right)^{-1} \sum_{(m,4)} w_{ij}(i_1, i_2, i_3, i_4)\psi_3(i_1, i_2, i_3, i_4),$$

with the weight function $w_{ij}(i_1, i_2, i_3, i_4)$ proportional to $\theta_i \theta_i \theta_i \theta_i$.

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

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

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

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

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

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

http://biostats.bepress.com/ucrbiostat/paper192
we generated a large population ($N = 4000$) and set, for a given marker pair located 5 kb apart, $\hat{\Delta}^{U}_{4000} = 0.897$ as the true $\Delta$. Then, 300 samples were drawn from the simulated population for each of the following sample sizes: $n = 50, 200, 500$. 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 the sample size $n$ increases, both estimators converge to the true $\Delta$. The same analysis was performed for marker pairs located 50 kb and 100 kb apart, and similar results for biases and variances of $\hat{\Delta}^U$ and $\hat{\Delta}$ were produced. For all the three marker pairs, the standard deviations for $\hat{\Delta}^U$ are slightly smaller than those for $\hat{\Delta}$ (table 1). However, the dependence structures within $\{X_{ij}\}$ and within $\{Y_{ij}\}$ 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.

Based on these results, we used the naive estimator $\hat{\Delta}$ for the rest of the data analysis, due to its simplicity and fast computation.

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

Assuming a finite-site uniform recombination model, in which the crossover probability is $10^{-8}$ bp$^{-1}$ for adjacent base pairs, we simulated a dataset of 620 SNPs in a region of size 270 kb, for 200 chromosomes. The crossover probability was chosen to approximate the average recombination rate for the human genome (i.e., 1 cM per Mb). Among all SNPs, 337 SNPs with minor allele frequencies (MAFs) >5% were the focus of our initial analysis. On this dataset, we computed and displayed the three pairwise LD measures, $D'$ (fig. 3A), $r^2$ (fig. 3B), and $\hat{\Delta}$ (fig. 3C), for each of the $\binom{337}{2} = 56, 616$ SNP pairs. Visualization tools from the Bioconductor hexbin package were applied to produce “2D histograms” that represent the density of data points in a scatter plot. Both $|D'|$ and $r^2$ tend to decrease as physical distance increases, as showed by the locally weighted scatter plot smoothing (lowess) curves. However, $|D'|$ and $r^2$ are highly variable at any given $d$. In
contrast, \( \hat{\Lambda} \) has a nearly deterministic relationship with \( d \).

Moreover, the lowess curve for \( \hat{\Lambda} \) nearly overlaps the exponential function \( e^{\beta d} \) (red line), implying that the relationship between \( \hat{\Lambda} \) 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, \( \hat{\Lambda} \) is nearly 1 for any pair of closely-located SNPs and decreases at a constant rate. By fitting the linear model \( E[\log \hat{\Lambda}] = -\beta d \), the LD decay rate may be estimated as \( \hat{\beta} = 0.010 \), meaning that \( \hat{\Lambda} \) decays exponentially at a rate of 0.010 per kb in the region.

**Impact of Marker Density**

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 \( \hat{\Lambda} \), we randomly selected subsets of SNPs from about 200 SNPs with MAFs >1%, simulated in a 100 kb region. The following percentages of SNPs were selected: 90%, 70%, 50%, 30%, 10%, allowing us to monitor the stability of \( \hat{\Lambda} \), for a fixed physical distance \( d \), as the marker density decreases from 20 to 2 SNPs per 10 kb. Random selection produced similar distributions of MAFs across different subsets, so that the effects of marker density and marker allele frequency (potentially an influential factor on the behavior of an LD measure) are not confounded. Figure 4A shows the distribution of \( \hat{\Lambda} \) for pairs of markers located 15 to 16 kb apart for different marker densities. Although \( \hat{\Lambda} \) tends to decrease slightly as marker density decreases, \( \hat{\Lambda} \) is generally robust to marker density. For other values of \( d \), similar patterns of robustness were observed.

The impact of marker density on the estimated rate of LD decay was also investigated. The process for random selection of the original markers was repeated 200 times for each marker density, and the LD decay rate \( \beta \) was estimated each time (fig. 4B).
In general, $\hat{\beta}$ appears to be fairly stable for marker densities of 6 or more SNPs per 10 kb, but not for the low density of 2 SNPs per 10 kb, due to 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**

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 dataset 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 about 23, 19, 12, 10, and 6 SNPs per 10 kb. In this range of marker densities, based on the above results, $\Delta$ and its decay rate $\beta$ can be robustly and reliably estimated. Clearly, the exponential relationship between $\hat{\Delta}$ and $d$, and the low variability of $\hat{\Delta}$ at a given $d$, are both maintained across the five SNP subsets. For pairs of SNPs located a certain distance away from each other, say, 10 to 11 kb (fig. 5), $\hat{\Delta}$ is fairly stable in terms of its median and interquartile range across subsets of SNPs. 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 $\beta$ is also very robust to SNP allele frequency, ranging from 0.009 to 0.011 across subsets of SNPs.

**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, $d = 15$ to 16 kb, $\hat{\Delta}$ decreases as the recombination rate increases (fig. 6A). Furthermore, LD decay rates were estimated for 200 independently simulated datasets in
each setting, and were strongly related to recombination rates (fig. 6B).

**Analysis of HapMap SNP Data for the X-chromosome**

We applied the new LD measure $\Delta$ to HapMap data (phase I) for the X-chromosomes of 30 mothers in the CEPH population (Utah residents with ancestry from Northern and Western Europe). Genotypes at 56,001 SNPs were fully phased, among which there were 19,127 monomorphic SNPs. LD decay rates $\beta$ were estimated at every polymorphic SNP locus, by applying the method of least squares to the exponential decay model for $\hat{\Delta}$.

Specifically, pairwise $\hat{\Delta}$ were computed from neighboring polymorphic SNPs within 100 kb windows, as long as there were $\geq 7$ polymorphic SNPs so that the number of marker pairs used to estimate $\beta$ was at least 21 (fig. 7A). 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 estimates of $\Delta$ and its decay rate $\beta$. The results show that LD on the X-chromosome decays exponentially at an average rate of 0.0073 per kb within 100 kb windows, while at certain loci the rate can reach 0.054 per kb. Figure 7B provides a higher resolution display for 100 polymorphic SNPs in the 12.89 Mb to 13.17 Mb region. Pseudo-color images of pairwise $|D'|$, $r^2$ and $\hat{\Delta}$ matrices in this region are displayed in figures 7D-F.

Actual genomic data are different from simulated data in one important aspect. The recombination rate can be fixed in the simulated data, while 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, it is expected that LD does not 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 based on region of
only 100 kb, with the hope that the LD decay rate does not change dramatically within this relatively small region. However, this assumption could still be violated due to recombination hotspots. A recombination hotspot is a site prone to recombination and experimentally identified as a region as narrow as 1-2 kb, where recombination rates are higher than in neighboring regions (Jefferey 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 is needed for developing other indices for the investigation of fine-scale LD. We anticipate that the new LD measure $\Lambda$ may make meaningful contributions to this endeavor.

As an example of usefulness of our new LD measure, heuristic analysis of data for the above 275 kb region in figure 7B suggests that recombination hotspots may be identified based on $\hat{\Lambda}$. We focus on all adjacent marker pairs in the region of interest, as long as $\hat{\Lambda}$ can be calculated based on markers with density higher than 2 SNP per 10 kb. The tight relationship between $\hat{\Lambda}$ 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 the adjacent marker pairs, the regression model $E[\log \hat{\Lambda}] = -\beta d$ was fit. Outliers among the observed marker pairs that have 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, due to the dependence of $\hat{\Lambda}$ between adjacent marker pairs (as shown by the residual plot in figure 7C). We do not intend to address the issue of outlier detection in depth in this article. Instead, we graphically show that the marker pairs with the extreme negative residuals (plotted by the red and blue dots in figure 7C) correspond to potential recombination hotspots (indicated by the red and blue lines, respectively, in figures 7D-F).
Discussion

The proposed LD measure $\Delta$ 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 $\Delta$. Here, we examine the following factors one by one: distinction between IBD and IBS status, marker density, and censoring.

Firstly, the length of a common ancestral segment is best measured based on alleles IBD for a chromosome pair. However, in practice, it is often impossible to distinguish between IBD and IBS given genotype data from unrelated individuals. 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.

Secondly, higher marker densities lead to better approximation of the length of a common ancestral segment by the length of shared haplotype. Based on simulation studies, the impact of marker density on estimation of $\hat{\Delta}$ is very limited once it is above a certain threshold, i.e., 2 SNPs per 10 kb, which is a feasible density given the imminent availability of ultra-high-volume genotyping platforms. Note that efforts are not needed to
identify tagSNPs when markers are used for 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 due to reduced marker density.

Thirdly, censoring at the edge of the genotyped region is an important 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. 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 the flanking regions. For instance, when LD decays fast, smaller flanking regions are considered as adequate. The guideline we applied was to use the size of the region across which LD decays to a value near 0. For example, in our simulation study, when the crossover probability was set to $10^{-8}$, a flanking region of $>200$ kb could stabilize $\hat{\Delta}$, whereas for a crossover probability of $10^{-7}$, a flanking region of $100$ kb was good enough. For HapMap X-chromosome data, the lengths of haplotype sharing were calculated using $500$ kb flanking regions on both sides of the reference locus. For the edge of large genotyped regions, such as telomeres in chromosome-wide genotype data, a possible way to avoid censoring is that, for markers at the left (or right, respectively) edge, only the lengths of haplotype sharing to the right (or left, respectively) should be involved in calculating $\hat{\Delta}^U$ or $\hat{\Delta}$. 
To estimate $\Delta$ for unphased data, a straightforward two-stage scheme can be adopted. In the first stage, haplotypes and their frequencies are inferred for the whole dataset 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 $\hat{\Lambda}^U$ or $\hat{\Lambda}$ as in the setting of non-distinct haplotypes. The second approach usually results in many ambiguous phases for individual chromosomes, which may compromise the accuracy and stability of $\hat{\Lambda}^U$ or $\hat{\Lambda}$.

Finally, we would like to address the connections and differences between recombination hotspots and boundaries for haplotype blocks, as the latter have become accepted as a general model for LD patterns throughout the genome. The two terms both 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 block boundaries were simulated as recombination hotspots. 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, where LD decays faster than in other regions, LD decay rate is an important aspect and physical distance plays an essential role. In the HapMap data analysis, hotspots were identified based on residuals for a fitted exponential decay model for $\hat{\Lambda}$ and $d$, instead of based on $\hat{\Lambda}$ only. In contrast, block boundaries are usually chosen to achieve low haplotype diversity within each block, based on significantly low LD value, without taking physical distance into consideration.

In conclusion, simulation studies and analysis of HapMap data demonstrate that $\Delta$ and its estimators 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 genetic models, ∆ 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 have been proposed. It was shown that the new measure may greatly facilitate the identification of recombination hotspots. ∆ can be a helpful tool in studying population genetics, using data from different populations, and in mapping complex disease genes, using samples with different phenotypes. The method can also be readily applied to data for more polymorphic DNA markers (e.g. microsatellites) or amino acid sequence data without further extension.

**Acknowledgements**

This work was partially supported by NIH Training Grant 5 T32 HG00047, NIH grant RO1 LM07609, and 5RO1 CA106320-05. We thank Ying Qing Chen, Sue Li, and Najma Khalid for helpful discussion.

**Appendix A**

**Proof of the Proposition**

Let

\[ \psi((S_1, T_1), \ldots, (S_4, T_4)) \]

\[ = \frac{1}{6} [\phi_1((S_1, T_1), \ldots, (S_4, T_4)) + \phi_2((S_1, T_1), \ldots, (S_4, T_4)) + \phi_3((S_1, T_1), \ldots, (S_4, T_4))] , \]
where

\[ \phi_1((S_1, T_1), \ldots, (S_4, T_4)) = \min(S_1, S_2) - \min(S_3, S_4) \mid \min(T_1, T_2) - \min(T_3, T_4), \]

\[ \phi_2((S_1, T_1), \ldots, (S_4, T_4)) = \min(S_1, S_3) - \min(S_2, S_4) \mid \min(T_1, T_3) - \min(T_2, T_4), \]

\[ \phi_3((S_1, T_1), \ldots, (S_4, T_4)) = \min(S_1, S_4) - \min(S_2, S_3) \mid \min(T_1, T_4) - \min(T_2, T_3). \]

Further 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 i.i.d., 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,

\[
E[\phi_1((S_1, T_1), \ldots, (S_4, T_4))] = E[(X_{12} - X_{34})(Y_{12} - Y_{34})] \\
= E[X_{12}Y_{12} - X_{12}Y_{34} - X_{34}Y_{12} + X_{34}Y_{34}] \\
= 2E[X_{12}Y_{12}] - 2E[X_{12}]E[Y_{12}].
\]

Similarly,

\[
E[\phi_2((S_1, T_1), \ldots, (S_4, T_4))] = E[\phi_3((S_1, T_1), \ldots, (S_4, T_4))] = 2E[X_{12}Y_{12}] - 2E[X_{12}]E[Y_{12}].
\]

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 \( Y \).

References

- Dawson E, Abecasis GR, Bumpstead S, Chen Y, Hunt S, Beare DM, Pabial J,


• Hill WG, Robertson A (1968) Linkage disequilibrium in finite populations. Theor Appl Genet 38:226-231


• 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


Table 1 Standard deviations of $\hat{\Delta}U$ and $\hat{\Lambda}$ for 300 random samples of $n$ haplotypes from a simulated population of $N = 4000$ haplotypes. The estimates $\hat{\Delta}U$ and $\hat{\Lambda}$ are computed for 3 pairs of markers located $d = 5, 50,$ and $100$ kb apart.

<table>
<thead>
<tr>
<th>sample size, $n$</th>
<th>$d = 5$ kb</th>
<th>$d = 50$ kb</th>
<th>$d = 100$ kb</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\hat{\Delta}U$</td>
<td>$\hat{\Lambda}$</td>
<td>$\hat{\Delta}U$</td>
</tr>
<tr>
<td>50</td>
<td>0.0416</td>
<td>0.0418</td>
<td>0.0740</td>
</tr>
<tr>
<td>200</td>
<td>0.0177</td>
<td>0.0179</td>
<td>0.0361</td>
</tr>
<tr>
<td>500</td>
<td>0.0110</td>
<td>0.0111</td>
<td>0.0202</td>
</tr>
</tbody>
</table>
Figure 1 Conceptual model for motivating the LD measure $\Delta$. (A) When loci $A$ and $B$ are in complete LD, the lengths of haplotype sharing around $A$ and $B$ are linearly dependent for all chromosome pairs. (B) When loci $A$ and $B$ are in complete linkage equilibrium (LE), the lengths of haplotype sharing around $A$ and $B$ are independent for all chromosome pairs.
**A**

Loci A and B are in complete linkage disequilibrium (LD):

Ancestral segment

Chromosome 1

Chromosome 2

X and Y (X' and Y') have linear relationship:

\[ X = Y + d_0 \]; \[ X' = Y' - d_0 \].

**B**

Loci A and B are in complete linkage equilibrium (LE):

Ancestral segment

Chromosome 1

Chromosome 2

X and Y (X' and Y') are independent.
**Figure 2** Boxplots for biases of $\hat{\Delta}^U$ and $\hat{\Delta}$ for 300 random samples of $n$ haplotypes from a simulated population of $N = 4000$ haplotypes for a particular marker pair located 5 kb apart. The 6 boxplots (from left to right) are for $\hat{\Delta}^U_{50}$, $\hat{\Delta}_{50}$, $\hat{\Delta}^U_{200}$, $\hat{\Delta}_{200}$, $\hat{\Delta}^U_{500}$, $\hat{\Delta}_{500}$, where the numbers in the subscript denote the sample size, $n$. 
**Figure 3** Pairwise LD measures 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 scatter plot smoothing (lowess) curves are plotted in black, with the smooth at each value influenced by 1% data points. For SNP data simulated under the uniform recombination model, LD decays exponentially at a constant rate, which can be estimated based on the linear regression model: $E[\log \hat{\Delta}] = -\beta d$. This relationship is plotted in red.
A

B

C

\[ \hat{\Delta} = e^{-\hat{\beta}d} \]
**Figure 4** Robustness of $\hat{\Delta}$ and its decay rate to marker density. (A) Distributions of $\hat{\Delta}$ for pairs of SNPs that are located 15 to 16 kb apart for different marker densities. (B) Distributions of $\hat{\beta}$ estimated from 200 SNP subsets randomly selected for each marker density. The scale of the vertical axis was chosen to match that in figure 6.
Figure 5 Robustness of $|D'|$, $r^2$, and $\hat{\Delta}$ to marker allele frequency. Distribution of $|D'|$, $r^2$, and $\hat{\Delta}$ for SNP pairs that are located 10 to 11 kb apart, using subsets of SNPs with different MAFs (from left to right, at least 0, 1%, 5%, 10%, and 20%, for each measure).
Figure 6 $\hat{\Delta}$ and its decay rate as a function of recombination rate. (A) Distributions of $\hat{\Delta}$ for pairs of SNPs that are located 15 to 16 kb apart, for different crossover probabilities for adjacent base-pair. (B) Distributions of estimated rates of LD decay, $\hat{\beta}$, for 200 simulations at different crossover probabilities.
Figure 7 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. (B) A display with higher resolution for an arbitrarily selected region of 275 kb. (C) Residual plot from fitting the linear regression model $E[\log \hat{\Lambda}] = -\beta d$ using data from the 275 kb region. Marker pairs with large negative residuals are heuristically considered as recombination hotspots, and are plotted using colored dots, with red representing even more extreme residuals than blue. (D)-(F) Pseudo-color images of pairwise $|D'|$, $r^2$, and $\hat{\Lambda}$ matrices. The red and blue dashed lines correspond to the marker pairs plotted using red and blue dots in figure 7C.