Genome-Wide Estimation of Linkage Disequilibrium from Population-Level High-Throughput Sequencing Data

Takahiro Maruki[1] **and Michael Lynch**
Department of Biology, Indiana University, Bloomington, Indiana 47405

ABSTRACT Rapidly improving sequencing technologies provide unprecedented opportunities for analyzing genome-wide patterns of polymorphisms. In particular, they have great potential for linkage-disequilibrium analyses on both global and local genetic scales, which will substantially improve our ability to derive evolutionary inferences. However, there are some difficulties with analyzing high-throughput sequencing data, including high error rates associated with base reads and complications from the random sampling of sequenced chromosomes in diploid organisms. To overcome these difficulties, we developed a maximum-likelihood estimator of linkage disequilibrium for use with error-prone sampling data. Computer simulations indicate that the estimator is nearly unbiased with a sampling variance at high coverage asymptotically approaching the value expected when all relevant information is accurately estimated. The estimator does not require phasing of haplotypes and enables the estimation of linkage disequilibrium even when all individual reads cover just single polymorphic sites.

LINKAGE disequilibrium (LD) refers to the nonrandom association of alleles at different loci. Estimating LD and analyzing its pattern are important for several reasons. First, the genealogies of two physically close sites are identical, unless there are recombination events between the sites (Sved 1971; Hudson 1983). Because the amount of LD between sites declines with the number of historical recombination events between them, the former provides insight into the latter, which cannot be observed directly (Hudson 2001; Stumpf and McVean 2003; McVean *et al.* 2004). Second, studying patterns of LD is important for gene mapping. In association studies, extensive LD facilitates the identification of candidate regions of functional importance (Gabriel *et al.* 2002; International HapMap Consortium 2003), whereas weaker LD enables finer-scale analyses (Zhu *et al.* 2000; Tishkoff and Williams 2002). Third, significant improvements in inference of population demographic parameters can be made from analyses of LD patterns. By comparing the LD decay pattern in the same region across different populations, differences in effective population size or recombination rates among populations can be inferred (Frisse *et al.* 2001; Reich *et al.* 2001; Conrad *et al.* 2006). Furthermore, the historical change in effective population size in a population can be inferred by examining the relationship between map distance and degree of LD between sites (Hill 1981; Hayes *et al.* 2003; Tenesa *et al.* 2007). Finally, LD is one of the most powerful means for detecting signatures of genetic forces such as gene conversion and natural selection (Hudson *et al.* 1994; Langley *et al.* 2000; Frisse *et al.* 2001; Przeworski and Wall 2001; Sabeti *et al.* 2002; Eberle *et al.* 2006; Kim *et al.* 2007).

Population-level high-throughput sequencing (Mardis 2008; Shendure and Ji 2008) provides unprecedented opportunities for analyzing genome-wide and local patterns of LD. The technologies are rapidly improving (Glenn 2011). In particular, read lengths are becoming longer and are expected to be much longer in the near future, an ideal situation for estimating LD, as longer reads allow increasing possibilities for the direct phasing of double heterozygotes. However, high-throughput sequencing also imposes some disadvantages. The error rates are high, ranging from 0.001 to 0.01 per base read, in commonly used platforms. Furthermore, error rates are known to vary among different runs, even with the same platform. In addition, as the two

doi: 10.1534/genetics.114.165514
Manuscript received April 18, 2014; accepted for publication May 20, 2014; published Early Online May 28, 2014.
Supporting information is available online at http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.114.165514/-/DC1.
[1]Corresponding author: Department of Biology, Indiana University, Bloomington, IN 47405. E-mail: tmaruki@indiana.edu

chromosomes are randomly sequenced in a particular genetic region in diploid organisms, confident inference of the genotypes of individuals with low depths of coverage is difficult.

Several statistical methods have recently been developed for estimating population parameters from high-throughput sequencing data (Hellmann *et al.* 2008; Johnson and Slatkin 2008, 2009; Jiang *et al.* 2009; Lynch 2009; Futschik and Schlotterer 2010; Hohenlohe *et al.* 2010; Kim *et al.* 2010; Keightley and Halligan 2011). Of these, Johnson and Slatkin (2009) developed a maximum-likelihood (ML) method for estimating recombination rates from high-throughput sequencing data, considering the difficulties explained above. They focused on estimating recombination rates from LD patterns under an assumed evolutionary model. In this study, we propose an ML method for overcoming the difficulties in the estimation of LD itself.

To study genome-wide patterns of LD using high-throughput sequencing data, researchers can attempt to reconstruct phased haplotypes from unphased genotypes and then estimate LD, using software packages such as fastPHASE (Scheet and Stephens 2006), BEAGLE (Browning and Browning 2007), and MaCH (Li *et al.* 2010). For putative double heterozygotes, these packages infer haplotypes from relative frequencies of putatively unambiguous genotypes, which are themselves called using another software package. Therefore, these software packages do not maximally use the information for LD estimation, in particular on haplotype phase, contained in the sequence-read data (Bansal *et al.* 2008; Long *et al.* 2009). Here we present an ML method for estimating LD in population-level analyses directly from the sequence-read data, without any requirements for phasing haplotypes.

To evaluate the performance of the ML method, computer simulations were carried out to generate sequence-read data and the parameters were estimated and compared to the expectations. The effects of read length and depth of coverage on the precision of the LD estimates were also investigated. The results demonstrate that the proposed method improves the potential of the LD analysis with high-throughput sequencing data and also provide guidelines for designing optimal sequencing strategies with a limited research budget.

Methods

Overall procedure for estimating population parameters

Although there are various measures of LD, the basic one is D (Lewontin and Kojima 1960), which measures the deviation of a haplotype frequency from its expected value under the random association of alleles (Hedrick 1987; Slatkin 2008). Other widely used LD measures such as D' (Lewontin 1964) and r^2 (Hill and Robertson 1968) can also be calculated when D and the allele frequencies at two polymorphic sites of interest are estimated, although their statistical properties are problematical (Lewontin 1988; Eberle *et al.* 2006; Song and Song 2007).

The sequence-read data relevant to our analysis are of two types: those covering just single polymorphic sites and those covering both polymorphic sites. For the former, the numbers of the four different nucleotide reads (*i.e.*, A, C, G, and T) in each individual are recorded as a quartet at the observed sites. For the latter, the numbers of the 16 different types of dinucleotide reads (*i.e.*, AA, AC, AG, AT, . . . , TA, TC, TG, TT) in each individual are recorded.

Our goal is to estimate the allele frequencies at each site, the LD coefficient D between the pair of polymorphic sites, and the error rate per site, by maximizing the likelihood of the observed set of sequence reads in a population sample. The number of alleles segregating at each site is assumed to be at most two, which is empirically known to be true for the majority of single-nucleotide polymorphisms in diploid organisms (Lynch 2007). Random union of gametes (*i.e.*, Hardy–Weinberg equilibrium) is assumed, to infer genotype frequencies from allele frequencies. Because our estimator requires a collection of read data for each individual, each sequence read is assumed to originate from a particular known individual, for example, by tagging reads when sequencing pooled samples. The allele frequencies and error rate at each site are first estimated from the observed set of site-specific sequence reads, using a method analogous to that developed by Lynch (2009). After obtaining this prior information, D is then estimated in a one-dimensional analysis from the entire set of sequence reads (some of which might cover both polymorphic sites of interest). A simple grid search exploring possible values of parameters is used to find the global maximum of the likelihood.

Estimation of allele frequencies at each site

Prior to embarking on a computationally demanding genome-wide survey of LD, it is more efficient to first identify the restricted set of loci at which polymorphisms are likely to exist. In such analyses, the two most abundant nucleotide reads in the population sample are considered to be candidates for alleles at each site. Then, the frequency of the most abundant nucleotide (which is not necessarily the true major allele) and the error rate per site are estimated by maximizing the likelihood of the observed site-specific read data as a function of the allele frequency and the error rate.

In the population sample of site-specific sequence-read data at site α, let A and a denote the first and second most abundant observed nucleotides, respectively. Let us denote the true frequency of the major (the most abundant) allele in the population sample at the site by p and the sequence error rate per site by ε. Denoting the site-specific genotypes AA, Aa, and aa genotypes 1, 2, and 3, respectively, for each individual i, the log-likelihood of the observed set of reads at site α is given by

$$\ln L_i = \ln\left[\sum_{g=1}^{3} \pi_g P_g(n_1, n_2, n_3)\right], \qquad (1)$$

where π_g simply denotes the Hardy–Weinberg expected genotype frequencies [$\pi_1 = p^2$, $\pi_2 = 2p(1-p)$, $\pi_3 = (1-p)^2$

at site α], and n_1, n_2, and n_3 are the observed number of reads of candidate nucleotides A (*e.g.*, C), a (*e.g.*, T), and e (other nucleotides, *e.g.*, in this case A and G), respectively. $P_g(n_1, n_2, n_3)$ is the probability of the specific observed set of nucleotide reads given genotype g, which, given a depth of coverage of the individual $n = n_1 + n_2 + n_3$, is calculated using the formula for the multinomial distribution:

$$P_g(n_1, n_2, n_3) = \frac{n!}{n_1!n_2!n_3!} \prod_{j=1}^{3} p_g(j)^{n_j}. \quad (2)$$

Here, $p_g(j)$ is a probability of observed nucleotide read j with genotype g. $p_g(j)$ is a function of ε and is given by summing conditional probabilities of observed nucleotide read j given the true nucleotide on the sequenced chromosome chosen from the pair (Table 1). For example, when $g = 2$ (*Aa*), the probability of nucleotide read 1 (*A*) is $p_2(1) = (1/2)(1 - \varepsilon) + (1/2)(\varepsilon/3)$.

The ML estimate of the allele frequency is found by maximizing the log-likelihood of the observed site-specific reads in the entire population sample, which is calculated by summing the log-likelihoods (Equation 1) over N individuals:

$$\ln L = \sum_{i=1}^{N} \ln L_i. \quad (3)$$

The estimates of the allele frequency and the error rate at the site, $\hat{p}$ and $\hat{\varepsilon}$, are obtained by maximizing $\ln L$ as a function of p and ε. The significance of the polymorphism at the site can be statistically tested by the likelihood-ratio test (Kendall and Stuart 1979). Specifically, the likelihood of the observed data at the site under the hypothesis of polymorphism can be compared to the corresponding likelihood of the data under the hypothesis of monomorphism to examine whether the former is significantly greater than the latter.

Estimation of the linkage disequilibrium coefficient between sites

The linkage disequilibrium coefficient between the sites, D, is estimated only if there is polymorphism at both sites α and β (*i.e.*, $\hat{p}, \hat{q}$ significantly < 1, where $\hat{q}$ is the estimate of the major allele frequency at site β). The likelihood of the entire set of sequence reads in the sample as a function of D is maximized to obtain the ML estimate $\hat{D}$, using the preestimated values of the allele frequencies and error rate in the likelihood function. D is highly dependent on allele frequencies, and its minimum, $D_{\min}$, and maximum, $D_{\max}$, given $\hat{p}$ and $\hat{q}$, are

$$\begin{aligned} D_{\min} &= -\min[\hat{p}\,\hat{q}, (1-\hat{p})(1-\hat{q})] \\ D_{\max} &= \min[\hat{p}(1-\hat{q}), (1-\hat{p})\hat{q}] \end{aligned} \quad (4)$$

(Lewontin 1988). Thus, the search for $\hat{D}$ need not exceed these bounds.

In the entire set of sequence reads with respect to the two polymorphic sites of interest, let k, l, and m denote the number of reads covering just site α, just site β, and both sites, respectively. The two-locus genotypes *AB/AB* (which consist of *AB* and *AB* haplotypes), *Ab/Ab*, *aB/aB*, *ab/ab*, *AB/Ab*, *aB/ab*, *AB/aB*, *Ab/ab*, *AB/ab*, and *Ab/aB* are denoted by 1, 2, 3, 4, 5, 6, 7, 8, 9, and 10, respectively. Then, the likelihood of the observed set of reads for the individual is calculated with expressions involving the reads covering sites α, β, or both. Analogous to Equation 1, the likelihood for an individual is calculated by multiplying the probability of each particular genotype and that of the observed set of reads given the genotype and then summing the products over all 10 genotypes,

$$\ln L_i = \ln\left[\sum_{g=1}^{10} \theta_g \{1 - (1 - P_{g1})I_1\}\{1 - (1 - P_{g2})I_2\} \cdot \{1 - (1 - P_{g3})I_3\}\right], \quad (5)$$

where θ_g now denotes the Hardy–Weinberg expected two-locus genotype frequencies, which are each products (for gametic homozygotes) or twice the products (for gametic heterozygotes) of two of the four gamete frequencies $pq + D$, $p(1 - q) - D$, $(1 - p)q - D$, and $(1 - p)(1 - q) + D$. P_{g1}, P_{g2}, and P_{g3} are the probabilities of the observed nucleotide reads covering just site α, just site β, and both sites, respectively. I_1 is an indicator variable equal to one for reads covering just site α and zero otherwise. Thus, $\{1 - (1 - P_{g1})I_1\}$ becomes P_{g1} for reads covering just site α and one otherwise. The same applies to sequence reads covering just site β and both sites. As an example, when there is a mixture of reads covering either site α or site β but not both sites, $\{1 - (1 - P_{g1})I_1\}\{1 - (1 - P_{g2})I_2\}\{1 - (1 - P_{g3})I_3\}$ becomes $P_{g1}P_{g2}$.

The probabilities P_{g1}, P_{g2}, and P_{g3} are calculated as follows. For the observed sequence reads covering just site α,

$$P_{g1} = \frac{k!}{k_1!k_2!k_3!} \prod_{j=1}^{3} p_g(j)^{k_j}, \quad (6)$$

where k_1, k_2, and k_3 are the observed numbers of the most abundant, the second most abundant, and other nucleotides given reads covering just site α, respectively, and k is their sum. $p_g(j)$ is a function of ε and is given by summing conditional probabilities of observed nucleotide read j given the true nucleotide on the sequenced chromosome chosen from

Table 1 Probability $p_g(j)$ of observed nucleotide read j with genotype g as a function of the error rate ε

	Nucleotide read		
Genotype	1 (*A*)	2 (*a*)	3 (*e*)
1 (*AA*)	$1 - \epsilon$	$\frac{\epsilon}{3}$	$\frac{2\epsilon}{3}$
2 (*Aa*)	$\frac{1}{2}\cdot(1-\epsilon)+\frac{1}{2}\cdot\frac{\epsilon}{3}$	$\frac{1}{2}\cdot\frac{\epsilon}{3}+\frac{1}{2}\cdot(1-\epsilon)$	$\frac{2\epsilon}{3}$
3 (*aa*)	$\frac{\epsilon}{3}$	$1 - \epsilon$	$\frac{2\epsilon}{3}$

A and *a* denote candidate alleles (the two most abundant nucleotide reads in the population sample, *e.g.*, C and T) and *e* denotes other nucleotide reads (*e.g.*, in this case A and G).

the pair (Supporting Information, Table S1). Similarly, P_{g2} is calculated for the observed sequence reads covering just site β. For the observed sequence reads covering both sites,

$$P_{g3} = \frac{m!}{m_1!m_2!m_3!m_4!m_5!m_6!m_7!m_8!m_9!}\prod_{j=1}^{9} p_g\,(j)^{m_j}, \quad (7)$$

where m_1, m_2, m_3, m_4, m_5, m_6, m_7, m_8, and m_9 are the observed numbers of dinucleotide reads *AB*, *Ab*, *Ae*, *aB*, *ab*, *ae*, *eB*, *eb*, and *ee*, respectively, and m is their sum. $p_g(j)$ is a function of ε and is given by summing conditional probabilities of observed dinucleotide read j given the true nucleotides on a sequenced chromosome chosen from the pair (Table S2). For example, when $g = 5$ (*AB/Ab*), the probability of dinucleotide read 1 (*AB*) is $p_5(1) = (1/2)(1-\varepsilon)^2 + (1/2)(1-\varepsilon)(\varepsilon/3)$. The log-likelihood for the observed sequence reads in the entire population sample, ln L, is calculated by summing the log-likelihoods (Equation 5) over all N individuals, as in Equation 3. The estimate of the LD coefficient, $\hat{D}$, is obtained by maximizing the likelihood of the observed set of read data as a function of D.

Expectation and sampling variance of the linkage disequilibrium coefficient estimates

When the LD coefficient D is estimated from known genotypes in a sample of N diploid individuals, the expectation of the ML estimate is

$$E(\hat{D}) = \frac{2N-1}{2N}D \quad (8)$$

(Weir 1996), so we need to multiply the ML estimates by $(2N)/(2N-1)$ to remove bias. When the haplotypes are phased, and coupling and repulsion double heterozygotes can be distinguished, and again assuming no uncertainty in genotypes, the sampling variance of the ML estimate is

$$\mathrm{Var}(\hat{D}) = \frac{p(1-p)q(1-q) + (1-2p)(1-2q)D - D^2}{2N} \quad (9)$$

(Hill 1974), where p and q are the allele frequencies at the two polymorphic sites of interest. When coupling and repulsion double heterozygotes cannot be distinguished, the sampling variance of the ML estimate is

$$\mathrm{Var}(\hat{D}) = \frac{p(1-p)q(1-q)}{N-1} + \frac{(1-2p)(1-2q)D}{2N} + \frac{D^2}{N(N-1)} \quad (10)$$

(Cockerham and Weir 1977; Weir 1979). The mean and sampling variance of the ML estimates using the proposed approach are expected to asymptotically correspond to these values when depths of coverage are high.

In high-throughput sequencing data, depths of coverage vary among sites, individuals, and chromosomes within individuals, and Equations 8–10 need to be modified to account for this additional source of variation. To accomplish this, we express the expectation and sampling variance of the ML estimates as functions of the mean depth of coverage, μ, by considering the "effective" number of sampled chromosomes/individuals with respect to the inferences at two polymorphic sites of interest; these are defined as the numbers of sampled chromosomes and individuals for which there are sequence reads enabling haplotype and two-locus genotype estimation, respectively. Because of the variability of the depth of coverage among individuals and random sampling of sequenced chromosomes, these numbers are smaller than the actual numbers.

Let us assume that the depth of coverage at each site per individual (X) is Poisson distributed such that

$$f(X;\mu) = \frac{(\mu)^X e^{-\mu}}{X!}, \quad (11)$$

where f is the probability mass function of X and μ is the mean coverage. The probability of zero coverage in an individual is $e^{-\mu}$.

When the read length is much larger than the distance between polymorphic sites of interest and all reads cover both of the sites, LD can be estimated from haplotypes. In this case, the probability that just one of the two chromosomes is sampled from an individual is

$$\mu e^{-\mu} + \sum_{k=2}^{\infty}\frac{(\mu)^k e^{-\mu}}{k!}\left(\frac{1}{2}\right)^{k-1} = 2\left(e^{\mu/2}-1\right)e^{-\mu}, \quad (12)$$

where the first term on the left side of the equation is the probability that there is only one read for the individual. Both chromosomes are sampled at least once with probability $1 - e^{-\mu} - 2(e^{\mu/2}-1)e^{-\mu} = 1 - (2e^{\mu/2}-1)e^{-\mu}$. Therefore, the effective number of sampled chromosomes with phased haplotypes is defined as

$$N_c = N\left[2\left(e^{\mu/2}-1\right)e^{-\mu} + 2 - 2\left(2e^{\mu/2}-1\right)e^{-\mu}\right] = 2N\left(1-e^{-(\mu/2)}\right). \quad (13)$$

Assuming sequence errors are properly accounted for by the ML method, the expectation and sampling variance of the ML estimates can therefore be anticipated by replacing $2N$ by N_c in Equations 8 and 9 when all reads cover both polymorphic sites.

At the other extreme, when the read length is smaller than the distance between polymorphic sites of interest, all reads cover just single polymorphic sites, and the estimation of LD needs to be made from unphased genotypes. Let ν denote the probability that the total set of read data specific to a polymorphic site is not informative for the inference of the one-locus genotype of an individual, a situation that arises when one or both of the chromosomes are not read at least once. That is,

$$\nu = e^{-\mu} + \mu e^{-\mu} + \sum_{k=2}^{\infty}\frac{(\mu)^k e^{-\mu}}{k!}\left(\frac{1}{2}\right)^{k-1} = \left(2e^{\frac{\mu}{2}}-1\right)e^{-\mu}. \quad (14)$$

The site-specific read data are informative for the genotype inference at either of the sites with probability $2\nu(1-\nu)$ and informative for the genotype inferences at both sites with probability $(1-\nu)^2$. Therefore, assuming that the information at the two loci is additive with respect to the inferences of two-locus genotypes, the effective number of sampled individuals is

$$\begin{aligned} N_i &= N\left[\frac{1}{2}\cdot 2\nu(1-\nu)+(1-\nu)^2\right] \\ &= N(1-\nu) = N\left[1-\left(2e^{\mu/2}-1\right)e^{-\mu}\right] \end{aligned} \tag{15}$$

and again Equations 8 and 10 can be adjusted by substituting N_i for N.

In many cases, there is a mixture of reads covering just single polymorphic sites and those covering both polymorphic sites. To predict the sampling variance of the ML estimates in such cases, let γ denote the probability that there is at least one read covering both polymorphic sites for an individual. Then, the sampling variance of $\hat{D}$ as a function of γ is anticipated, using the above results and the conditional variance formula (Ross 2006), as follows:

$$\begin{aligned} \mathrm{Var}(\hat{D}) &= E\left[\mathrm{Var}\left(\hat{D}\middle|\text{haplotype phase}\right)\right] \\ &\quad + \mathrm{Var}\left(E\left[\hat{D}\middle|\text{haplotype phase}\right]\right) \\ &= \gamma\cdot\frac{p(1-p)q(1-q)+(1-2p)(1-2q)D-D^2}{N_c} \\ &\quad + (1-\gamma)\cdot\left[\frac{p(1-p)q(1-q)}{N_i-1}\right. \\ &\quad \left.+\frac{(1-2p)(1-2q)D}{2N_i}+\frac{D^2}{N_i(N_i-1)}\right] \\ &\quad + \gamma\cdot\left(\frac{N_c-1}{N_c}D\right)^2+(1-\gamma)\cdot\left(\frac{2N_i-1}{2N_i}D\right)^2 \\ &\quad - \left[\gamma\cdot\frac{N_c-1}{N_c}D+(1-\gamma)\cdot\frac{2N_i-1}{2N_i}D\right]^2. \end{aligned} \tag{16}$$

γ is a function of μ and the probability that a read covers both polymorphic sites given it covers one polymorphic site, ϕ. Letting d and L be the distance between the polymorphic sites of interest (in base pairs) and read length (in base pairs),

$$\begin{aligned} \phi &= \frac{L-d}{L+d}, \quad \text{when } d\le L, \\ &= 0, \qquad \text{when } d>L. \end{aligned} \tag{17}$$

Generation of sequence-read data by computer simulation

Stochastic simulations were carried out to generate sequence-read data at two sites of interest in samples of N diploid individuals. Individual genotypes were assigned with probabilities equal to their frequencies under given population parameters, p (major allele frequency at site α), q (major allele frequency at site β), and D (LD coefficient between the sites), assuming random union of gametes. The depth of coverage at each site per individual was assumed to be Poisson distributed with mean μ. Letting ϕ be the probability that a read covers both sites given it covers one site, the numbers of reads covering just site α, just site β, and both sites were assumed to be Poisson distributed with parameters $\mu(1-\phi)$, $\mu(1-\phi)$, and $\mu\phi$, respectively, so that the number of reads at each site was Poisson distributed with parameter μ, satisfying ϕ. Then, the probability that there is at least one read covering both sites for an individual is $\gamma = 1-e^{-\mu\phi}$. Errors were randomly introduced at each site on a read of each individual with probability ε, assuming each error from a true nucleotide has equal probability $\varepsilon/3$.

Comparison of the LD estimation performance by the proposed method with that by an imputation-based method

To compare the performance of the LD estimation by the proposed ML method with that by widely used imputation-based methods, we estimated LD using the imputation-based phasing software package BEAGLE from the sequence reads generated by the computer simulations described above and compared the bias and sampling variance of the LD estimates by the two methods. We mapped the sequence reads to a reference sequence generated by a computer simulation, which consists of 10,000 random nucleotides, using Novoalign (www.novocraft.com). Each of the polymorphic sites was flanked by 22 nucleotides on both sides in a sequence read, which are each uniquely found in the reference sequence, so that every sequence read is mapped to the correct location. We called the genotypes of the individuals at two polymorphic sites of interest, using SAMtools (Li *et al.* 2009). The missing genotypes and haplotype phase of the individuals were imputed by BEAGLE version 4 (Browning and Browning 2013). LD coefficients were calculated from the VCF file of imputed genotype data, using VCFtools (Danecek *et al.* 2011).

Results

The performance of the LD estimator was evaluated using sequence-read data generated by computer simulations according to the methods described above. A relatively high error rate of 0.01 was assumed, to evaluate the performance in the face of the worst effects of sequence errors expected with widely used sequencing devices (Margulies *et al.* 2005; Huse *et al.* 2007). The mean of the LD coefficient estimates, $\hat{D}$, was nearly unbiased (Figure 1). Furthermore, at high depths of coverage, the sampling variance of the estimates approached the asymptotic value (Equation 9), which is expected when all relevant information is obtained without error. Both the accuracy and the precision of the estimates improved when N (number of sampled individuals) was larger.

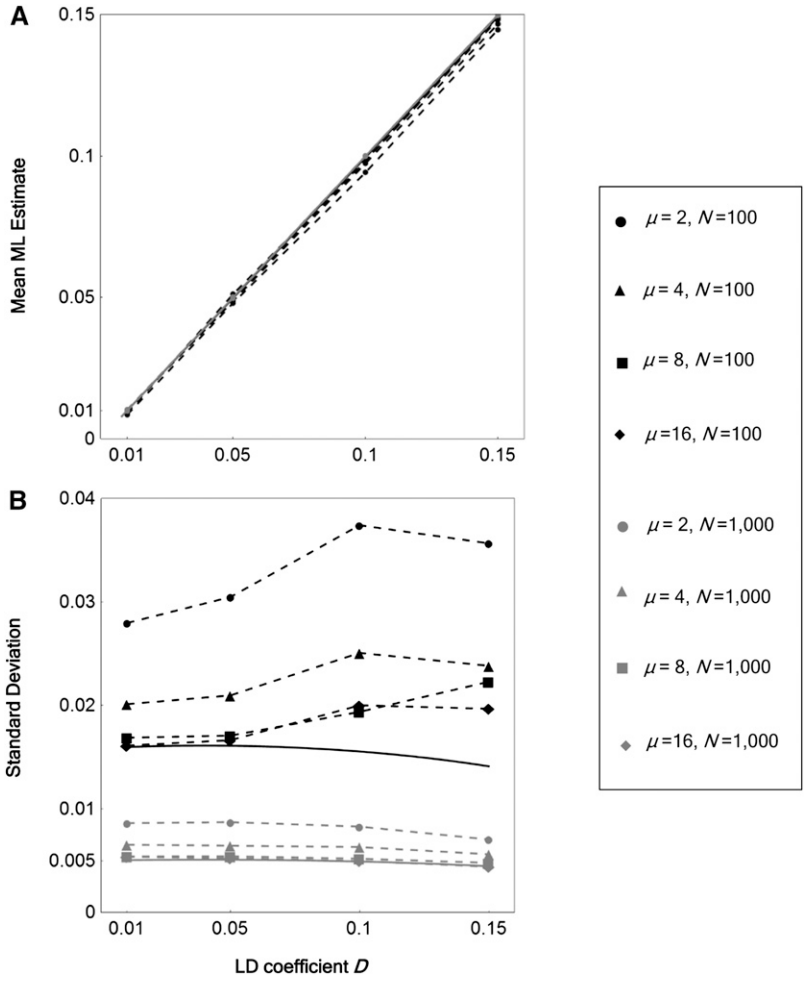

Figure 1 Accuracy and precision of the ML estimates of LD. Sample mean (A) and standard deviation (B) of ML estimates are shown as functions of the LD coefficient D, the number of sampled individuals (sample size) N, and mean depth of coverage μ. Major allele frequency at site α, p = 0.6; major allele frequency at site β, q = 0.7; read length, L = 100; distance between sites of interest, d = 50; and error rate, ε = 0.01. The probability that a read covers both sites given it covers one site, Φ, is 1/3. A total of 1000 simulation replicates were run for each set of parameter values. The asymptotic theoretical sample expectation (given by Equation 8) and standard deviation (given by the square root of Equation 9) of $\hat{D}$ are plotted by the solid lines (A) and curves (B), respectively.

To examine the effect of ϕ (the fraction of informative reads covering both sites of interest) on the performance of the LD estimator, we varied the distance between the sites of interest, with a fixed read length: $L = 100$ and $d = 1, 10, 30, 50, 70, 90, 100$ (Figure 2). The means of the estimates were nearly unbiased regardless of ϕ, although the sampling variance of the estimates declined with increasing ϕ. The precision of the estimates greatly improves with initial increases in the depth of coverage, although this improvement diminishes rapidly beyond $\mu = 4$. When ϕ was lower, the sampling variance was more influenced by the depth of coverage. Even though there are no reads covering both sites when ϕ is zero, in this case, the mean of the estimates is still nearly unbiased, with the sampling variance of the estimates approaching the asymptotic value (Equation 10) when depths of coverage are high.

To examine the sensitivity of the LD estimator to allele frequencies, in particular when all reads covered just single polymorphic sites (*i.e.*, $\phi = 0$), we estimated the LD coefficients with various allele frequencies at the two sites of interest (Table 2). A total of 1000 simulation replicates were run to obtain these results, with sample size $N = 1000$, LD coefficient $D = 0.01$ or 0.05, and error rate $\varepsilon = 0.01$. Because we defined that the sign of D is positive when major or minor alleles at the two sites are on the same chromosome and negative otherwise, when the allele frequencies are 0.5, the sign of D cannot be defined and therefore the estimates of D are not reported; instead, the mean and sampling standard deviation of the squared estimates of D are reported. Overall, the mean and sampling standard deviation of $\hat{D}$ estimated from simulated sequence reads agreed well with the theoretical predictions. Furthermore, they were close to the asymptotic theoretical predictions (Equation 8 and square root of Equation 10) when the mean depth of coverage was 16, as then the effective number of sampled individuals is approximately equal to the actual number of individuals in the sample (Equation 15).

Given that the proposed estimator provides nearly unbiased LD estimates with asymptotically minimal sampling variances, we compared the LD-estimation performance of the proposed estimator to that of a widely used imputation-based method to examine whether the proposed method enables better LD estimates from high-throughput sequencing data than the standard approach of SNP calling followed by genotype imputation. For this purpose, we compared the mean and standard deviation of the LD estimates by the proposed method to those obtained using the widely used imputation-based software package BEAGLE (Browning and

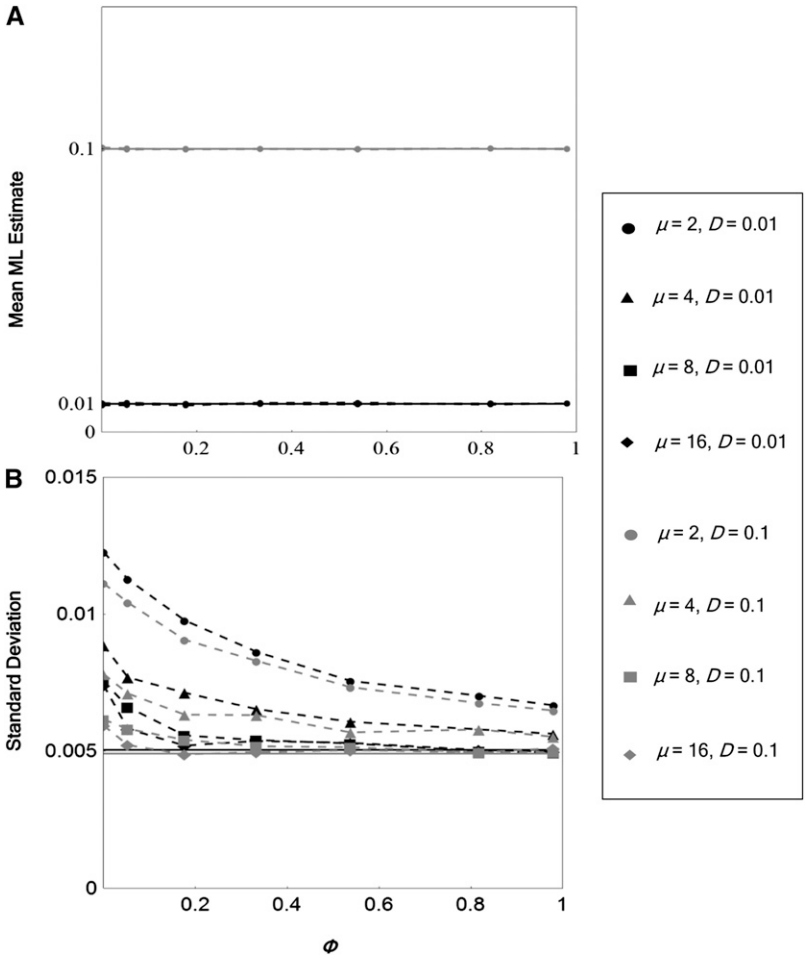

Figure 2 Effect of Φ (the probability that an informative read covers both sites of interest) on the performance of the LD estimator. Sample mean (A) and standard deviation (B) of ML estimates of LD are shown as functions of Φ, LD coefficient D, and mean depth of coverage μ. Sample size N = 1000; major allele frequency at site α, p = 0.6; major allele frequency at site β, q = 0.7; read length, L = 100; and error rate, ε = 0.01. A total of 1000 simulation replicates were run for each set of parameter values. The asymptotic theoretical sample expectation (given by Equation 8) and standard deviation (given by the square root of Equation 9) of $\hat{D}$ are plotted by the solid lines.

Browning 2013) (Table S3 and Table S4). When the major allele frequencies at two sites of interest were intermediate, LD estimates by the proposed estimator were better than those by the imputation-based method in the majority of the examined cases in terms of both accuracy and precision (Table S3). When the major allele frequencies were high, the root mean square deviations (RMSDs) of the LD estimates by the imputation-based method were smaller than those by the proposed estimator in some of the cases (Table S4). However, in those cases, the means of the LD estimates by the imputation-based methods were underestimated. The degree of the underestimation was especially large when major allele frequencies were high and the depth of coverage was low. On the other hand, the means of the LD estimates by the proposed estimator were close to the true values with their standard deviations essentially equal to the RMSD in all of the cases, indicating our method enables essentially unbiased LD estimation from high-throughput sequencing data.

Finally, to find the optimal sequencing strategy for analyzing linkage disequilibrium with a limited research budget, the sampling standard deviation of the estimates of D was examined, fixing the value of the product of the mean depth of coverage (μ) and sample size (N) (Figure 3). When all sequence reads cover both sites of interest (*i.e.*, $\phi = 1$), the minimum sampling variance with a fixed value of μN is achieved at $\mu = 1$ (Figure 3A). When all sequence reads cover just one of the sites of interest (*i.e.*, $\phi = 0$), the minimum sampling variance is at a slightly higher value of $\mu = 2$ (Figure 3B).

Discussion

High-throughput sequencing technologies provide unprecedented opportunities for analyzing global as well as local patterns of polymorphisms in various organisms (Pool *et al.* 2010). In particular, they harbor great potential for haplotype analyses because sequencing occurs on single DNA molecules, allowing for the unambiguous determination of the phases of haplotypes if sequence errors are correctly accounted for. Genome-wide high-throughput sequencing data at the population level are accumulating in various organisms (Cao *et al.* 2011; Altshuler *et al.* 2012; Mackay *et al.* 2012). One of the major discoveries of these projects is the localization of a number of new polymorphic sites harboring alleles with low frequencies. The resultant denser polymorphisms provide excellent opportunities for finer LD analyses essential for understanding the contributions of

Table 2 Sensitivity of the LD estimator to allele frequencies when all sequence reads cover single polymorphic sites

D	μ	(p, q)	$\hat{D}$ (mean ± SD)	Theoretical SD ($\hat{D}$)	$\hat{D}^2$ (mean ± SD)	Asymptotic theoretical SD ($\hat{D}$)
0.01	2	(0.5, 0.5)	NA	0.013	0.0003 ± 0.0004	0.008
0.01	2	(0.6, 0.6)	0.010 ± 0.013	0.012	0.0003 ± 0.0004	0.008
0.01	2	(0.7, 0.7)	0.010 ± 0.011	0.011	0.0002 ± 0.0003	0.007
0.01	2	(0.8, 0.8)	0.010 ± 0.009	0.008	0.0002 ± 0.0002	0.005
0.01	2	(0.9, 0.9)	0.010 ± 0.006	0.005	0.0001 ± 0.0001	0.003
0.01	2	(0.6, 0.9)	0.010 ± 0.008	0.008	0.0002 ± 0.0002	0.005
0.01	16	(0.5, 0.5)	NA	0.008	0.0002 ± 0.0002	0.008
0.01	16	(0.6, 0.6)	0.010 ± 0.008	0.008	0.0002 ± 0.0002	0.008
0.01	16	(0.7, 0.7)	0.010 ± 0.007	0.007	0.0002 ± 0.0002	0.007
0.01	16	(0.8, 0.8)	0.010 ± 0.005	0.005	0.0001 ± 0.0001	0.005
0.01	16	(0.9, 0.9)	0.010 ± 0.003	0.003	0.0001 ± 0.0001	0.003
0.01	16	(0.6, 0.9)	0.010 ± 0.005	0.005	0.0001 ± 0.0001	0.005
0.05	2	(0.5, 0.5)	NA	0.013	0.0027 ± 0.0014	0.008
0.05	2	(0.6, 0.6)	0.051 ± 0.012	0.012	0.0027 ± 0.0013	0.008
0.05	2	(0.7, 0.7)	0.050 ± 0.011	0.011	0.0027 ± 0.0011	0.007
0.05	2	(0.8, 0.8)	0.050 ± 0.009	0.009	0.0026 ± 0.0009	0.006
0.05	2	(0.9, 0.9)	0.050 ± 0.008	0.008	0.0025 ± 0.0007	0.005
0.05	2	(0.6, 0.9)	0.050 ± 0.007	0.008	0.0026 ± 0.0007	0.005
0.05	16	(0.5, 0.5)	NA	0.008	0.0026 ± 0.0007	0.008
0.05	16	(0.6, 0.6)	0.049 ± 0.007	0.008	0.0025 ± 0.0007	0.008
0.05	16	(0.7, 0.7)	0.050 ± 0.007	0.007	0.0025 ± 0.0007	0.007
0.05	16	(0.8, 0.8)	0.050 ± 0.005	0.006	0.0025 ± 0.0005	0.006
0.05	16	(0.9, 0.9)	0.050 ± 0.005	0.005	0.0025 ± 0.0005	0.005
0.05	16	(0.6, 0.9)	0.050 ± 0.004	0.005	0.0025 ± 0.0004	0.005

Sample means and standard deviations of the LD coefficient $\hat{D}$ and its square $\hat{D}^2$ estimated from simulated sequence data are shown for various major allele frequencies at two polymorphic sites, p and q. The theoretical prediction of the standard deviation (SD) as a function of the mean depth of coverage μ, which is calculated as a square root of the theoretical sampling variance from Equations 10 and 15, is also shown. The theoretical prediction of the asymptotic standard deviation, which is calculated as a square root of the sampling variance given by Equation 10, is the expected level of achievement when individuals are genotyped without errors.

various evolutionary forces, including mutation, recombination, random genetic drift, and natural selection to genome evolution (Lynch 2007).

Despite the promise, the new technologies have some disadvantages. Unlike conventional Sanger sequencing, high-throughput sequencing rapidly produces massive amounts of sequence data, sacrificing quality for quantity. In particular, high sequence error rates and random sampling of sequenced chromosomes in diploid organisms make estimation of allele frequencies and subsequent population-genomic analyses difficult. Because LD measures in general are dependent on allele frequencies (Lewontin 1988), correctly identifying polymorphic sites and estimating allele frequencies at the sites are especially important for LD analyses.

The proposed LD estimator overcomes problems associated with both sequence errors and the process of random sampling of sequenced chromosomes through their direct incorporation into the likelihood function. Because errors associated with DNA-sequence annotation include not only errors arising at the time of sequencing but also those introduced during sample preparation, it is risky to use externally determined "off-the-shelf" error-rate estimates. The proposed estimator avoids this problem by estimating error contributions from the sequence data themselves.

Unless the depth of coverage is high, just one of the two chromosomes may be sequenced at specific sites in a subset of diploid individuals, and to avoid this error in inference, many studies simply employ an arbitrary minimum coverage cutoff. However, such treatment can discard substantial amounts of informative data; and unless they are statistically accounted for, missing data are likely to introduce bias in population-genetic parameter estimates (Pool *et al.* 2010). Our results show that ML estimates of LD are nearly unbiased, with sampling variance at high coverages approaching the expectation when haplotype identities are known with certainty. This indicates that the proposed estimator maximally uses all available information. Thus, previously proposed methods of imputation of missing data, which may introduce biases in subsequent analyses (Lin and Huang 2007; Pool *et al.* 2010; and as shown in our results), are not necessary or often even desirable to obtain unbiased LD estimates. Furthermore, a recent study found that allele frequencies estimated via genotype calling are biased, whereas those estimated directly from mapped sequence reads by ML methods are unbiased, when the depth of coverage is low (Han *et al.* 2014). Given that carefully made reference panels do not exist for most organisms except for humans and flies, the proposed estimator should be especially useful when applied to high-throughput sequencing data for population-genomic analyses of nonmodel organisms. For example, Khatkar *et al.* (2010) substantially improved the quality of the bovine genome assembly by estimating pairwise LD between sites with unknown locations and those with known locations. We expect our estimator will play important roles in these kinds of applications, where unbiased estimation of LD is essential for correct inferences.

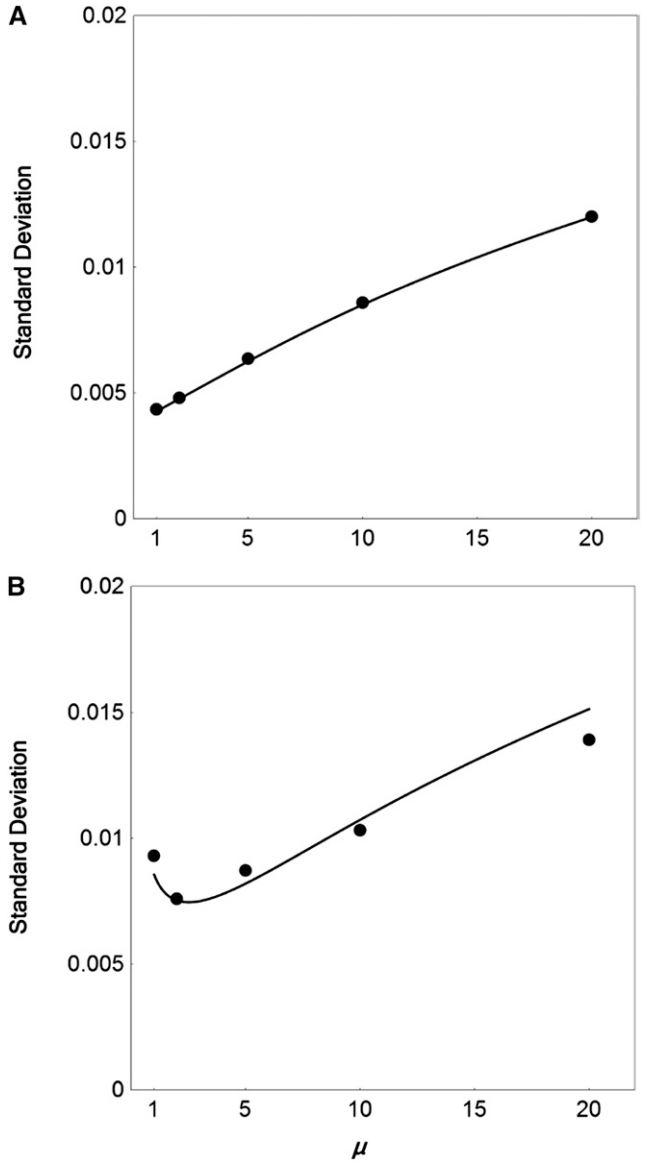

Figure 3 Sampling standard deviation of ML estimates of LD, as a function of the mean depth of coverage μ, estimated from a fixed number of total sequence reads. The product of μ and sample size N was fixed at 1000 such that a twofold increase in μ resulted in a twofold decrease in N. (A) Results when all sequence reads cover both sites of interest; (B) Results when all sequence reads cover just single polymorphic sites. The solid curves show the predicted values of the standard deviation (given by the square root of the theoretical sampling variance) of the estimates of the LD coefficient D from Equations 9 and 13 (A) and Equations 10 and 15 (B). The solid circles show the corresponding values of the standard deviation of $\hat{D}$, estimated from the simulated sequence read data. The major allele frequencies at two sites of interest $p = q = 0.9$, $D = 0.01$, and error rate $\varepsilon = 0.01$. A total of 1000 simulation replicates were run for each set of parameter values.

To exploit the emerging flood of data, an LD estimator specifically designed for the structure of the high-throughput sequencing data is required. Previous LD estimators (Hill 1974; Weir 1979; Feder *et al.* 2012) do not deal with the confounding effects of sequence errors on LD estimation. Although the most recent LD estimator (Feder *et al.* 2012) is designed for high-throughput sequencing data, it is limited to analyses of sequence reads containing both polymorphic sites. Our method not only deals with the confounding effects of sequence errors but also enables LD estimation even when all reads cover just single polymorphic sites. A C++ program for data analysis is available upon request, and we are currently implementing the proposed method as part of a software package for population-genomic analyses of high-throughput sequencing data.

When the present LD estimator is applied to huge numbers of sites across the genome, efficient means of application will be required for the analyses. Because LD is relevant only to pairs of polymorphic sites, it is useful to first apply the ML allele-frequency estimator (Lynch 2009) before pursuing the estimation of LD on the restricted set of relevant data. Then, using the ML estimates of allele frequencies and position-specific error rates, the LD coefficients between pairs of polymorphic sites can be rapidly estimated, as this is reduced to a one-dimensional problem. Of course, the simple grid search for finding the ML estimates taken in the allele-frequency estimator and the LD estimator can be replaced by a more efficient algorithm, *e.g.*, the simplex method by Nelder and Mead (1965). Although error rates were assumed to be the same at two sites of interest in this article, this assumption can also be relaxed by rewriting the likelihood function with different error rates at the sites. Obviously, before applying any method to LD analysis, it is important to exclude sites where paralogous sequences are found elsewhere in the genome, for example, by eliminating sites with abnormally high depths of coverage.

Because random union of gametes is assumed in the proposed estimator, it should be applied to a randomly mating population. To study LD in a nonrandomly mating population, the composite linkage disequilibrium measure (Cockerham and Weir 1977; Weir 1979) needs to be estimated from two-locus genotype frequencies. Such an extension can in principle be made by modifying the present likelihood function (Equation 5), although nine parameters must then be estimated.

The present population-level LD estimator and the single-individual disequilibrium estimator developed by Lynch (2008) complement each other. The latter does not assume Hardy–Weinberg equilibrium and estimates global LD patterns as functions of the distribution of heterozygous sites from high-throughput sequencing data in a single individual. The mean of the LD estimates by the proposed population-level estimator is expected to be consistent with the estimates by the single-individual estimator, provided that the population is randomly mating (Lynch 2008). Specifically, the mean of $\hat{D}^2$ is expected to closely correspond to $\hat{\Delta}\hat{\pi}(1-\hat{\pi})/4$, where $\hat{\Delta}$ and $\hat{\pi}$ are the single-individual genome-wide estimates of LD coefficients and nucleotide diversity, respectively.

Finally, the theoretical results in this investigation can be used to guide the design of optimal strategies for estimating LD from population-level high-throughput sequencing data

in the face of a limited budget. In general, regardless of the read length, a mean 2× depth of coverage appears to be the optimal allocation of resources, assuming the cost is proportional to $N\mu$, the expected total number of sequenced bases per site.

Alternatively, because sequencing costs are decreasing, one may find that the cost for DNA library preparation is more expensive than that for sequencing itself. The optimal strategy in such cases can also be flexibly predicted with the sampling variance formulas. Letting C_l and C_s denote the costs for library preparation per individual and sequencing per genome, respectively, the total cost T is

$$\begin{aligned} T &= C_l N + C_s(\mu N) \\ &= (C_l + \mu C_s)N. \end{aligned} \tag{18}$$

For example, when $C_l = 3C_s$, $(3 + \mu)N = T/C_s$. In this case, assuming $T = 100{,}000$ and $C_s = 20$, the minimum sampling variance is predicted to be achieved when μ is somewhere between 2 and 3 (Figure S1), with N then being 833 (with $\mu = 3$) to 1000 (with $\mu = 2$). In some studies, however, the number of available samples may become the limiting factor in designing the optimal study design. Our concepts of effective number of sampled individuals and chromosomes provide insights for designing the optimal strategy in such cases. Equations 13 and 15 show that the effective numbers quickly approach the actual numbers of chromosomes and individuals with higher depths of coverage.

Another complication in real studies is higher variance in depths of coverage than that by the Poisson distribution (Quail *et al.* 2012). The higher variance in depths of coverage is expected to decrease the effective numbers and therefore increase the optimal depths of coverage. Taken together, if a central goal of a population-genomic survey is the accurate estimation of LD, sequencing as many individuals as possible is the first priority.

Acknowledgments

The authors thank the anonymous reviewers, whose comments improved the manuscript. This work was supported by National Science Foundation grants EF-0827411 and DEB-1257806 and National Institutes of Health (NIH) NIH–National Institute of General Medical Sciences grant 1R01GM101672-01A1.

Literature Cited

Altshuler, D. M., R. M. Durbin, G. R. Abecasis, D. R. Bentley, A. Chakravarti *et al.*, 2012 An integrated map of genetic variation from 1,092 human genomes. Nature 491: 56–65.

Bansal, V., A. L. Halpern, N. Axelrod, and V. Bafna, 2008 An MCMC algorithm for haplotype assembly from whole-genome sequence data. Genome Res. 18: 1336–1346.

Browning, S. R., and B. L. Browning, 2007 Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. Am. J. Hum. Genet. 81: 1084–1097.

Browning, B. L., and S. R. Browning, 2013 Improving the accuracy and efficiency of identity-by-descent detection in population data. Genetics 194: 459–471.

Cao, J., K. Schneeberger, S. Ossowski, T. Gunther, S. Bender *et al.*, 2011 Whole-genome sequencing of multiple *Arabidopsis thaliana* populations. Nat. Genet. 43: 956–963.

Cockerham, C. C., and B. S. Weir, 1977 Digenic descent measures for finite populations. Genet. Res. 30: 121–147.

Conrad, D. F., M. Jakobsson, G. Coop, X. Wen, J. D. Wall *et al.*, 2006 A worldwide survey of haplotype variation and linkage disequilibrium in the human genome. Nat. Genet. 38: 1251–1260.

Danecek, P., A. Auton, G. Abecasis, C. A. Albers, E. Banks *et al.*, 2011 The variant call format and VCFtools. Bioinformatics 27: 2156–2158.

Eberle, M. A., M. J. Rieder, L. Kruglyak, and D. A. Nickerson, 2006 Allele frequency matching between SNPs reveals an excess of linkage disequilibrium in genic regions of the human genome. PLoS Genet. 2: e142.

Feder, A. F., D. A. Petrov, and A. O. Bergland, 2012 LDx: estimation of linkage disequilibrium from high-throughput pooled resequencing data. PLoS ONE 7: e48588.

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

Futschik, A., and C. Schlotterer, 2010 The next generation of molecular markers from massively parallel sequencing of pooled DNA samples. Genetics 186: 207–218.

Gabriel, S. B., S. F. Schaffner, H. Nguyen, J. M. Moore, J. Roy *et al.*, 2002 The structure of haplotype blocks in the human genome. Science 296: 2225–2229.

Glenn, T. C., 2011 Field guide to next-generation DNA sequencers. Mol. Ecol. Resour. 11: 759–769.

Han, E., J. S. Sinsheimer, and J. Novembre, 2014 Characterizing bias in population genetic inferences from low-coverage sequencing data. Mol. Biol. Evol. 31: 723–735.

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

Hedrick, P. W., 1987 Gametic disequilibrium measures: proceed with caution. Genetics 117: 331–341.

Hellmann, I., Y. Mang, Z. Gu, P. Li, F. M. de la Vega *et al.*, 2008 Population genetic analysis of shotgun assemblies of genomic sequences from multiple individuals. Genome Res. 18: 1020–1029.

Hill, W. G., 1974 Estimation of linkage disequilibrium in randomly mating populations. Heredity 33: 229–239.

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

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

Hohenlohe, P. A., S. Bassham, P. D. Etter, N. Stiffler, E. A. Johnson *et al.*, 2010 Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. PLoS Genet. 6: e1000862.

Hudson, R. R., 1983 Properties of a neutral allele model with intragenic recombination. Theor. Popul. Biol. 23: 183–201.

Hudson, R. R., 2001 Two-locus sampling distributions and their application. Genetics 159: 1805–1817.

Hudson, R. R., K. Bailey, D. Skarecky, J. Kwiatowski, and F. J. Ayala, 1994 Evidence for positive selection in the superoxide dismutase (*Sod*) region of *Drosophila melanogaster*. Genetics 136: 1329–1340.

Huse, S. M., J. A. Huber, H. G. Morrison, M. L. Sogin, and D. M. Welch, 2007 Accuracy and quality of massively parallel DNA pyrosequencing. Genome Biol. 8: R143.

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

Jiang, R., S. Tavare, and P. Marjoram, 2009 Population genetic inference from resequencing data. Genetics 181: 187–197.
Johnson, P. L., and M. Slatkin, 2008 Accounting for bias from sequencing error in population genetic estimates. Mol. Biol. Evol. 25: 199–206.
Johnson, P. L., and M. Slatkin, 2009 Inference of microbial recombination rates from metagenomic data. PLoS Genet. 5: e1000674.
Keightley, P. D., and D. L. Halligan, 2011 Inference of site frequency spectra from high-throughput sequence data: quantification of selection on nonsynonymous and synonymous sites in humans. Genetics 188: 931–940.
Kendall, M., and A. Stuart, 1979 *The Advanced Theory of Statistics*, Vol. 2, Ed. 4. Charles Griffin & Co..
Khatkar, M. S., M. Hobbs, M. Neuditschko, J. Solkner, F. W. Nicholas *et al.*, 2010 Assignment of chromosomal locations for unassigned SNPs/scaffolds based on pair-wise linkage disequilibrium estimates. BMC Bioinformatics 11: 171.
Kim, S., V. Plagnol, T. T. Hu, C. Toomajian, R. M. Clark *et al.*, 2007 Recombination and linkage disequilibrium in *Arabidopsis thaliana*. Nat. Genet. 39: 1151–1155.
Kim, S. Y., Y. Li, Y. Guo, R. Li, J. Holmkvist *et al.*, 2010 Design of association studies with pooled or un-pooled next-generation sequencing data. Genet. Epidemiol. 34: 479–491.
Langley, C. H., B. P. Lazzaro, W. Phillips, E. Heikkinen, and J. M. Braverman, 2000 Linkage disequilibria and the site frequency spectra in the *su*(s) and *su*(w(a)) regions of the *Drosophila melanogaster* X chromosome. Genetics 156: 1837–1852.
Lewontin, R. C., 1964 The interaction of selection and linkage. I. General considerations; heterotic models. Genetics 49: 49–67.
Lewontin, R. C., 1988 On measures of gametic disequilibrium. Genetics 120: 849–852.
Lewontin, R. C., and K. Kojima, 1960 The evolutionary dynamics of complex polymorphisms. Evolution 14: 458–472.
Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan *et al.*, 2009 The Sequence Alignment/Map format and SAMtools. Bioinformatics 25: 2078–2079.
Li, Y., C. J. Willer, J. Ding, P. Scheet, and G. R. Abecasis, 2010 MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. Genet. Epidemiol. 34: 816–834.
Lin, D. Y., and B. E. Huang, 2007 The use of inferred haplotypes in downstream analyses. Am. J. Hum. Genet. 80: 577–579.
Long, Q., D. MacArthur, Z. M. Ning, and C. Tyler-Smith, 2009 HI: haplotype improver using paired-end short reads. Bioinformatics 25: 2436–2437.
Lynch, M., 2007 *The Origins of Genome Architecture*. Sinauer Associates, Sunderland, MA.
Lynch, M., 2008 Estimation of nucleotide diversity, disequilibrium coefficients, and mutation rates from high-coverage genome-sequencing projects. Mol. Biol. Evol. 25: 2409–2419.
Lynch, M., 2009 Estimation of allele frequencies from high-coverage genome-sequencing projects. Genetics 182: 295–301.
Mackay, T. F. C., S. Richards, E. A. Stone, A. Barbadilla, J. F. Ayroles *et al.*, 2012 The *Drosophila melanogaster* genetic reference panel. Nature 482: 173–178.
Mardis, E. R., 2008 The impact of next-generation sequencing technology on genetics. Trends Genet. 24: 133–141.
Margulies, M., M. Egholm, W. E. Altman, S. Attiya, J. S. Bader *et al.*, 2005 Genome sequencing in microfabricated high-density picolitre reactors. Nature 437: 376–380.
McVean, G. A. T., S. R. Myers, S. Hunt, P. Deloukas, D. R. Bentley *et al.*, 2004 The fine-scale structure of recombination rate variation in the human genome. Science 304: 581–584.
Nelder, J. A., and R. Mead, 1965 A simplex-method for function minimization. Comput. J. 7: 308–313.
Pool, J. E., I. Hellmann, J. D. Jensen, and R. Nielsen, 2010 Population genetic inference from genomic sequence variation. Genome Res. 20: 291–300.
Przeworski, M., and J. D. Wall, 2001 Why is there so little intragenic linkage disequilibrium in humans? Genet. Res. 77: 143–151.
Quail, M. A., M. Smith, P. Coupland, T. D. Otto, S. R. Harris *et al.*, 2012 A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. BMC Genomics 13: 341.
Reich, D. E., M. Cargill, S. Bolk, J. Ireland, P. C. Sabeti *et al.*, 2001 Linkage disequilibrium in the human genome. Nature 411: 199–204.
Ross, S. A., 2006 *A First Course in Probability*. Pearson Education. Upper Saddle River, NJ.
Sabeti, P. C., D. E. Reich, J. M. Higgins, H. Z. Levine, D. J. Richter *et al.*, 2002 Detecting recent positive selection in the human genome from haplotype structure. Nature 419: 832–837.
Scheet, P., and M. Stephens, 2006 A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. Am. J. Hum. Genet. 78: 629–644.
Shendure, J., and H. Ji, 2008 Next-generation DNA sequencing. Nat. Biotechnol. 26: 1135–1145.
Slatkin, M., 2008 Linkage disequilibrium–understanding the evolutionary past and mapping the medical future. Nat. Rev. Genet. 9: 477–485.
Song, Y. S., and J. S. Song, 2007 Analytic computation of the expectation of the linkage disequilibrium coefficient r2. Theor. Popul. Biol. 71: 49–60.
Stumpf, M. P. H., and G. A. T. McVean, 2003 Estimating recombination rates from population-genetic data. Nat. Rev. Genet. 4: 959–968.
Sved, J. A., 1971 Linkage disequilibrium and homozygosity of chromosome segments in finite populations. Theor. Popul. Biol. 2: 125–141.
Tenesa, A., P. Navarro, B. J. Hayes, D. L. Duffy, G. M. Clarke *et al.*, 2007 Recent human effective population size estimated from linkage disequilibrium. Genome Res. 17: 520–526.
Tishkoff, S. A., and S. M. Williams, 2002 Genetic analysis of African populations: human evolution and complex disease. Nat. Rev. Genet. 3: 611–621.
Weir, B. S., 1979 Inferences about linkage disequilibrium. Biometrics 35: 235–254.
Weir, B. S., 1996 *Genetic Data Analysis II*. Sinauer Associates, Sunderland, MA.
Zhu, X., C. A. McKenzie, T. Forrester, D. A. Nickerson, U. Broeckel *et al.*, 2000 Localization of a small genomic region associated with elevated ACE. Am. J. Hum. Genet. 67: 1144–1153.

Communicating editor: J. Wakeley

GENETICS

Supporting Information
http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.114.165514/-/DC1

Genome-Wide Estimation of Linkage Disequilibrium from Population-Level High-Throughput Sequencing Data

Takahiro Maruki and Michael Lynch

DOI: 10.1534/genetics.114.165514

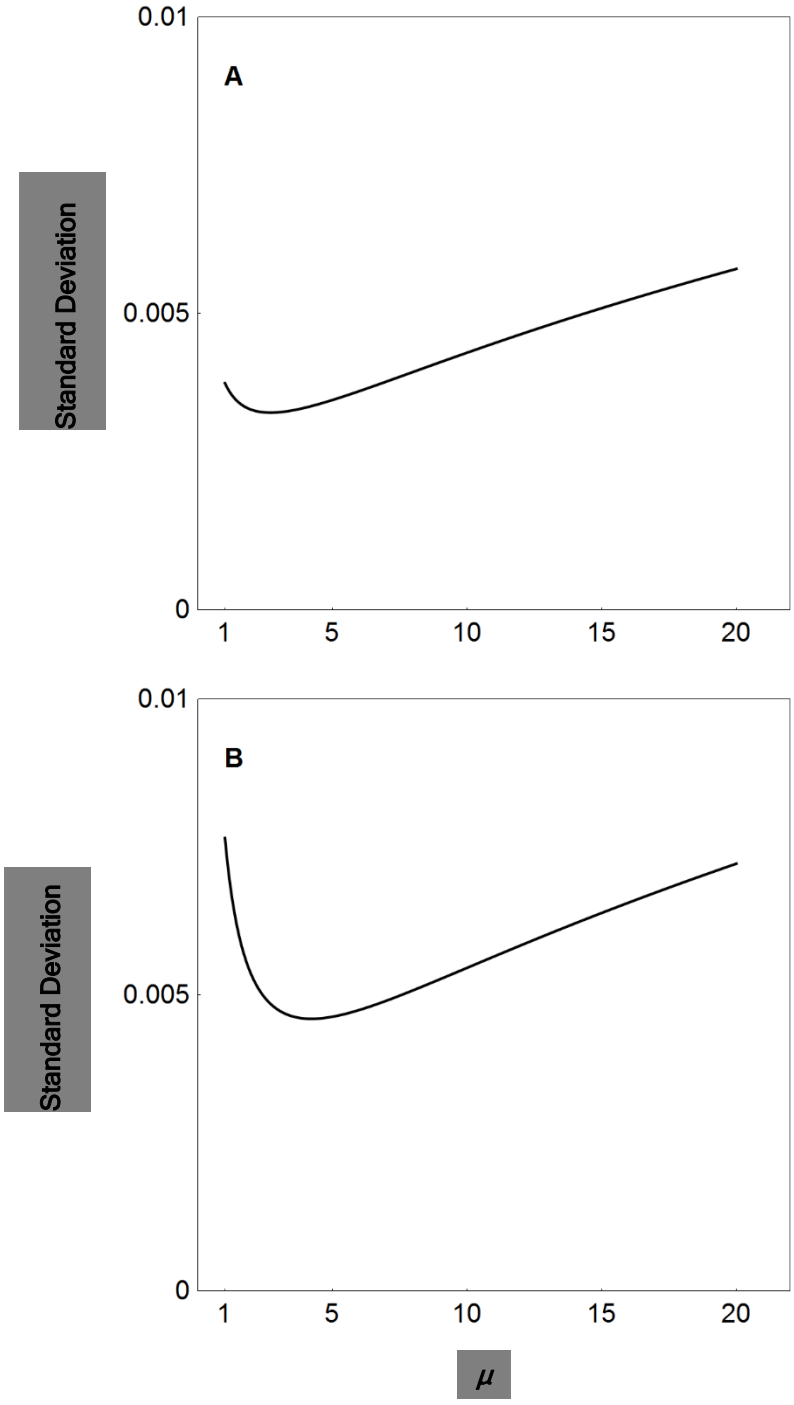

Figure S1: Predicted sampling standard deviation of ML estimates of LD under a fixed total budget, as a function of the mean depth of coverage μ, when library preparation is three times more expensive than sequencing. Figure S1A shows the predicted sampling standard

deviation of the ML estimates of the LD coefficient D when all sequence reads cover both sites of interest (given by the square root of the theoretical sampling variance from Equations 9 and 13). Figure S1B shows that when all sequence reads cover just single polymorphic sites (given by the square root of the theoretical sampling variance from Equations 10 and 15). The major allele frequencies at two sites of interest are $p = q = 0.9$, and $D = 0.01$.

TABLE S1: Probability $p_g(j)$ of observed nucleotide read *j* at site α with two-locus genotype *g* as a function of the error rate ε.

	Nucleotide read		
Genotype	1 (*A*)	2 (*a*)	3 (*e*)
1 (*AB/AB*)	$1-\epsilon$	$\frac{\epsilon}{3}$	$\frac{2\epsilon}{3}$
2 (*Ab/Ab*)	$1-\epsilon$	$\frac{\epsilon}{3}$	$\frac{2\epsilon}{3}$
3 (*aB/aB*)	$\frac{\epsilon}{3}$	$1-\epsilon$	$\frac{2\epsilon}{3}$
4 (*ab/ab*)	$\frac{\epsilon}{3}$	$1-\epsilon$	$\frac{2\epsilon}{3}$
5 (*AB/Ab*)	$1-\epsilon$	$\frac{\epsilon}{3}$	$\frac{2\epsilon}{3}$
6 (*aB/ab*)	$\frac{\epsilon}{3}$	$1-\epsilon$	$\frac{2\epsilon}{3}$
7 (*AB/aB*)	$\frac{1}{2}\cdot(1-\epsilon)+\frac{1}{2}\cdot\frac{\epsilon}{3}$	$\frac{1}{2}\cdot\frac{\epsilon}{3}+\frac{1}{2}\cdot(1-\epsilon)$	$\frac{2\epsilon}{3}$
8 (*Ab/ab*)	$\frac{1}{2}\cdot(1-\epsilon)+\frac{1}{2}\cdot\frac{\epsilon}{3}$	$\frac{1}{2}\cdot\frac{\epsilon}{3}+\frac{1}{2}\cdot(1-\epsilon)$	$\frac{2\epsilon}{3}$
9 (*AB/ab*)	$\frac{1}{2}\cdot(1-\epsilon)+\frac{1}{2}\cdot\frac{\epsilon}{3}$	$\frac{1}{2}\cdot\frac{\epsilon}{3}+\frac{1}{2}\cdot(1-\epsilon)$	$\frac{2\epsilon}{3}$
10 (*Ab/aB*)	$\frac{1}{2}\cdot(1-\epsilon)+\frac{1}{2}\cdot\frac{\epsilon}{3}$	$\frac{1}{2}\cdot\frac{\epsilon}{3}+\frac{1}{2}\cdot(1-\epsilon)$	$\frac{2\epsilon}{3}$

A and *a* denote candidate alleles (the two most abundant nucleotide reads in the population, e.g., C and T) and *e* denotes other nucleotide reads (e.g., in this case A and G) at site α. *B* and *b* denote candidate alleles at site β. In the two-locus genotype notation, the slash separates haplotypes.

TABLE S2: Probability $p_g(j)$ of observed dinucleotide read *j* at the two sites of interest with two-locus genotype *g* as a function of the error rate ε.

	Dinucleotide read								
Genotype	1 (*AB*)	2 (*Ab*)	3 (*Ae*)	4 (*aB*)	5 (*ab*)	6 (*ae*)	7 (*eB*)	8 (*eb*)	9 (*ee*)
1 (*AB/AB*)	$(1-\epsilon)^2$	$(1-\epsilon)\cdot\frac{\epsilon}{3}$	$(1-\epsilon)\cdot\frac{2\epsilon}{3}$	$\frac{\epsilon}{3}\cdot(1-\epsilon)$	$\left(\frac{\epsilon}{3}\right)^2$	$\frac{\epsilon}{3}\cdot\frac{2\epsilon}{3}$	$\frac{2\epsilon}{3}\cdot(1-\epsilon)$	$\frac{2\epsilon}{3}\cdot\frac{\epsilon}{3}$	$\left(\frac{2\epsilon}{3}\right)^2$
2 (*Ab/Ab*)	$(1-\epsilon)\cdot\frac{\epsilon}{3}$	$(1-\epsilon)^2$	$(1-\epsilon)\cdot\frac{2\epsilon}{3}$	$\left(\frac{\epsilon}{3}\right)^2$	$\frac{\epsilon}{3}\cdot(1-\epsilon)$	$\frac{\epsilon}{3}\cdot\frac{2\epsilon}{3}$	$\frac{2\epsilon}{3}\cdot\frac{\epsilon}{3}$	$\frac{2\epsilon}{3}\cdot(1-\epsilon)$	$\left(\frac{2\epsilon}{3}\right)^2$
3 (*aB/aB*)	$\frac{\epsilon}{3}\cdot(1-\epsilon)$	$\left(\frac{\epsilon}{3}\right)^2$	$\frac{\epsilon}{3}\cdot\frac{2\epsilon}{3}$	$(1-\epsilon)^2$	$(1-\epsilon)\cdot\frac{\epsilon}{3}$	$(1-\epsilon)\cdot\frac{2\epsilon}{3}$	$\frac{2\epsilon}{3}\cdot(1-\epsilon)$	$\frac{2\epsilon}{3}\cdot\frac{\epsilon}{3}$	$\left(\frac{2\epsilon}{3}\right)^2$
4 (*ab/ab*)	$\left(\frac{\epsilon}{3}\right)^2$	$\frac{\epsilon}{3}\cdot(1-\epsilon)$	$\frac{\epsilon}{3}\cdot\frac{2\epsilon}{3}$	$(1-\epsilon)\cdot\frac{\epsilon}{3}$	$(1-\epsilon)^2$	$(1-\epsilon)\cdot\frac{2\epsilon}{3}$	$\frac{2\epsilon}{3}\cdot\frac{\epsilon}{3}$	$\frac{2\epsilon}{3}\cdot(1-\epsilon)$	$\left(\frac{2\epsilon}{3}\right)^2$
5 (*AB/Ab*)	$\frac{1}{2}(1-\epsilon)^2+\frac{1}{2}\cdot(1-\epsilon)\cdot\frac{\epsilon}{3}$	$\frac{1}{2}\cdot(1-\epsilon)\cdot\frac{\epsilon}{3}+\frac{1}{2}(1-\epsilon)^2$	$(1-\epsilon)\cdot\frac{2\epsilon}{3}$	$\frac{1}{2}\cdot\frac{\epsilon}{3}\cdot(1-\epsilon)+\frac{1}{2}\cdot\left(\frac{\epsilon}{3}\right)^2$	$\frac{1}{2}\cdot\left(\frac{\epsilon}{3}\right)^2+\frac{1}{2}\cdot\frac{\epsilon}{3}\cdot(1-\epsilon)$	$\frac{\epsilon}{3}\cdot\frac{2\epsilon}{3}$	$\frac{1}{2}\cdot\frac{2\epsilon}{3}\cdot(1-\epsilon)+\frac{1}{2}\cdot\frac{2\epsilon}{3}\cdot\frac{\epsilon}{3}$	$\frac{1}{2}\cdot\frac{2\epsilon}{3}\cdot\frac{\epsilon}{3}+\frac{1}{2}\cdot\frac{2\epsilon}{3}\cdot(1-\epsilon)$	$\left(\frac{2\epsilon}{3}\right)^2$
6 (*aB/ab*)	$\frac{1}{2}\cdot\frac{\epsilon}{3}\cdot(1-\epsilon)+\frac{1}{2}\cdot\left(\frac{\epsilon}{3}\right)^2$	$\frac{1}{2}\cdot\left(\frac{\epsilon}{3}\right)^2+\frac{1}{2}\cdot\frac{\epsilon}{3}\cdot(1-\epsilon)$	$\frac{\epsilon}{3}\cdot\frac{2\epsilon}{3}$	$\frac{1}{2}(1-\epsilon)^2+\frac{1}{2}\cdot(1-\epsilon)\cdot\frac{\epsilon}{3}$	$\frac{1}{2}\cdot(1-\epsilon)\cdot\frac{\epsilon}{3}+\frac{1}{2}(1-\epsilon)^2$	$(1-\epsilon)\cdot\frac{2\epsilon}{3}$	$\frac{1}{2}\cdot\frac{2\epsilon}{3}\cdot(1-\epsilon)+\frac{1}{2}\cdot\frac{2\epsilon}{3}\cdot\frac{\epsilon}{3}$	$\frac{1}{2}\cdot\frac{2\epsilon}{3}\cdot\frac{\epsilon}{3}+\frac{1}{2}\cdot\frac{2\epsilon}{3}\cdot(1-\epsilon)$	$\left(\frac{2\epsilon}{3}\right)^2$

7 (*AB/aB*)	$\frac{1}{2}(1-\epsilon)^2+\frac{1}{2}\cdot\frac{\epsilon}{3}\cdot(1-\epsilon)$	$\frac{1}{2}\cdot(1-\epsilon)\cdot\frac{\epsilon}{3}+\frac{1}{2}\cdot\left(\frac{\epsilon}{3}\right)^2$	$\frac{1}{2}\cdot(1-\epsilon)\cdot\frac{2\epsilon}{3}+\frac{1}{2}\cdot\frac{\epsilon}{3}\cdot\frac{2\epsilon}{3}$	$\frac{1}{2}\cdot\frac{\epsilon}{3}\cdot(1-\epsilon)+\frac{1}{2}(1-\epsilon)^2$	$\frac{1}{2}\cdot\left(\frac{\epsilon}{3}\right)^2+\frac{1}{2}\cdot(1-\epsilon)\cdot\frac{\epsilon}{3}$	$\frac{1}{2}\cdot\frac{\epsilon}{3}\cdot\frac{2\epsilon}{3}+\frac{1}{2}\cdot(1-\epsilon)\cdot\frac{2\epsilon}{3}$	$\frac{2\epsilon}{3}\cdot(1-\epsilon)$	$\frac{2\epsilon}{3}\cdot\frac{\epsilon}{3}$	$\left(\frac{2\epsilon}{3}\right)^2$
8 (*Ab/ab*)	$\frac{1}{2}\cdot(1-\epsilon)\cdot\frac{\epsilon}{3}+\frac{1}{2}\cdot\left(\frac{\epsilon}{3}\right)^2$	$\frac{1}{2}(1-\epsilon)^2+\frac{1}{2}\cdot\frac{\epsilon}{3}\cdot(1-\epsilon)$	$\frac{1}{2}\cdot(1-\epsilon)\cdot\frac{2\epsilon}{3}+\frac{1}{2}\cdot\frac{\epsilon}{3}\cdot\frac{2\epsilon}{3}$	$\frac{1}{2}\cdot\left(\frac{\epsilon}{3}\right)^2+\frac{1}{2}\cdot(1-\epsilon)\cdot\frac{\epsilon}{3}$	$\frac{1}{2}\cdot\frac{\epsilon}{3}\cdot(1-\epsilon)+\frac{1}{2}(1-\epsilon)^2$	$\frac{1}{2}\cdot\frac{\epsilon}{3}\cdot\frac{2\epsilon}{3}+\frac{1}{2}\cdot(1-\epsilon)\cdot\frac{2\epsilon}{3}$	$\frac{2\epsilon}{3}\cdot\frac{\epsilon}{3}$	$\frac{2\epsilon}{3}\cdot(1-\epsilon)$	$\left(\frac{2\epsilon}{3}\right)^2$
9 (*AB/ab*)	$\frac{1}{2}(1-\epsilon)^2+\frac{1}{2}\cdot\left(\frac{\epsilon}{3}\right)^2$	$(1-\epsilon)\cdot\frac{\epsilon}{3}$	$\frac{1}{2}\cdot(1-\epsilon)\cdot\frac{2\epsilon}{3}+\frac{1}{2}\cdot\frac{\epsilon}{3}\cdot\frac{2\epsilon}{3}$	$\frac{\epsilon}{3}\cdot(1-\epsilon)$	$\frac{1}{2}\cdot\left(\frac{\epsilon}{3}\right)^2+\frac{1}{2}(1-\epsilon)^2$	$\frac{1}{2}\cdot\frac{\epsilon}{3}\cdot\frac{2\epsilon}{3}+\frac{1}{2}\cdot(1-\epsilon)\cdot\frac{2\epsilon}{3}$	$\frac{1}{2}\cdot\frac{2\epsilon}{3}\cdot(1-\epsilon)+\frac{1}{2}\cdot\frac{2\epsilon}{3}\cdot\frac{\epsilon}{3}$	$\frac{1}{2}\cdot\frac{2\epsilon}{3}\cdot\frac{\epsilon}{3}+\frac{1}{2}\cdot\frac{2\epsilon}{3}\cdot(1-\epsilon)$	$\left(\frac{2\epsilon}{3}\right)^2$
10 (*Ab/aB*)	$(1-\epsilon)\cdot\frac{\epsilon}{3}$	$\frac{1}{2}(1-\epsilon)^2+\frac{1}{2}\cdot\left(\frac{\epsilon}{3}\right)^2$	$\frac{1}{2}\cdot(1-\epsilon)\cdot\frac{2\epsilon}{3}+\frac{1}{2}\cdot\frac{\epsilon}{3}\cdot\frac{2\epsilon}{3}$	$\frac{1}{2}\cdot\left(\frac{\epsilon}{3}\right)^2+\frac{1}{2}(1-\epsilon)^2$	$\frac{\epsilon}{3}\cdot(1-\epsilon)$	$\frac{1}{2}\cdot\frac{\epsilon}{3}\cdot\frac{2\epsilon}{3}+\frac{1}{2}\cdot(1-\epsilon)\cdot\frac{2\epsilon}{3}$	$\frac{1}{2}\cdot\frac{2\epsilon}{3}\cdot\frac{\epsilon}{3}+\frac{1}{2}\cdot\frac{2\epsilon}{3}\cdot(1-\epsilon)$	$\frac{1}{2}\cdot\frac{2\epsilon}{3}\cdot(1-\epsilon)+\frac{1}{2}\cdot\frac{2\epsilon}{3}\cdot\frac{\epsilon}{3}$	$\left(\frac{2\epsilon}{3}\right)^2$

A and *a* denote candidate alleles (the two most abundant nucleotide reads in the population, e.g., C and T) and *e* denotes other nucleotide reads (e.g., in this case A and G) at site α. *B* and *b* denote candidate alleles at site β. In the two-locus genotype notation, the slash separates haplotypes.

TABLE S3: Comparison of LD estimates by the proposed ML estimator to those by an imputation-based method when the major allele frequencies at two sites of interest, p and q, are intermediate.

D	μ	(p,q)	Φ	Method	$\widehat{D}$ (mean±SD)	$\widehat{D}^2$ (mean±SD)	Asymptotic theoretical SD($\widehat{D}$)	Theoretical $E[\widehat{D}^2]$	RMSD($\widehat{D}$)	RMSD($\widehat{D}^2$)
0.01	2	(0.6,0.7)	0	ML	0.007±0.039	0.0016±0.0024	0.023	0.0006	0.039	0.0028
0.01	2	(0.6,0.7)	0	Imputation	0.002±0.018	0.0003±0.0007	0.023	0.0006	0.020	0.0007
0.01	2	(0.6,0.7)	1/3	ML	0.008±0.025	0.0007±0.0008	0.016	0.0004	0.025	0.0010
0.01	2	(0.6,0.7)	1/3	Imputation	0.002±0.016	0.0003±0.0006	0.016	0.0004	0.018	0.0006
0.01	2	(0.6,0.7)	1	ML	0.009±0.020	0.0005±0.0008	0.016	0.0004	0.020	0.0008
0.01	2	(0.6,0.7)	1	Imputation	0.009±0.025	0.0007±0.0019	0.016	0.0004	0.025	0.0020
0.01	10	(0.6,0.7)	0	ML	0.007±0.025	0.0006±0.0008	0.023	0.0006	0.025	0.0010
0.01	10	(0.6,0.7)	0	Imputation	0.006±0.026	0.0007±0.0014	0.023	0.0006	0.026	0.0016
0.01	10	(0.6,0.7)	1/3	ML	0.009±0.016	0.0003±0.0005	0.016	0.0004	0.016	0.0005
0.01	10	(0.6,0.7)	1/3	Imputation	0.007±0.017	0.0003±0.0004	0.016	0.0004	0.017	0.0005
0.01	10	(0.6,0.7)	1	ML	0.009±0.019	0.0004±0.0007	0.016	0.0004	0.018	0.0007
0.01	10	(0.6,0.7)	1	Imputation	0.008±0.026	0.0007±0.0017	0.016	0.0004	0.026	0.0018
0.1	2	(0.6,0.7)	0	ML	0.092±0.042	0.0103±0.0073	0.023	0.0105	0.043	0.0073
0.1	2	(0.6,0.6)	0	Imputation	0.036±0.029	0.0022±0.0026	0.023	0.0105	0.070	0.0083
0.1	2	(0.6,0.7)	1/3	ML	0.095±0.031	0.0101±0.0049	0.016	0.0101	0.031	0.0049
0.1	2	(0.6,0.7)	1/3	Imputation	0.043±0.031	0.0028±0.0028	0.016	0.0101	0.065	0.0077
0.1	2	(0.6,0.7)	1	ML	0.099±0.030	0.0107±0.0042	0.016	0.0101	0.030	0.0042
0.1	2	(0.6,0.7)	1	Imputation	0.082±0.034	0.0078±0.0062	0.016	0.0101	0.039	0.0066
0.1	10	(0.6,0.7)	0	ML	0.098±0.030	0.0104±0.0039	0.023	0.0105	0.030	0.0039
0.1	10	(0.6,0.7)	0	Imputation	0.107±0.026	0.0121±0.0045	0.023	0.0105	0.027	0.0049
0.1	10	(0.6,0.7)	1/3	ML	0.100±0.016	0.0103±0.0031	0.016	0.0101	0.016	0.0031
0.1	10	(0.6,0.7)	1/3	Imputation	0.110±0.021	0.0125±0.0040	0.016	0.0101	0.023	0.0047

0.1	10	(0.6,0.7)	1	ML	0.099±0.028	0.0106±0.0037	0.016	0.0101	0.027	0.0037
0.1	10	(0.6,0.7)	1	Imputation	0.108±0.026	0.0123±0.0047	0.016	0.0101	0.027	0.0052

Sample means and standard deviations of the LD coefficient $\widehat{D}$ and its square $\widehat{D}^2$ estimated from simulated data by the ML estimator and the imputation-based method are compared with different values of the parameters. Root mean square deviations (RMSD) of the LD estimates are also compared. The comparisons are made when the mean depth of coverage, μ is low (2) or moderately high (10). Furthermore, comparisons with different values of the probability that an informative read covers both polymorphic sites, Φ, are made. The theoretical prediction of the asymptotic standard deviation, which is calculated as a square root of the sampling variance given by Equation 9 (when $\Phi > 0$) or10 (when $\Phi = 0$), is the expected level of achievement when individual genotypes are known without errors. Sample size $N = 100$, error rate $\varepsilon = 0.01$. A total of 100 simulation replicates were run for each set of parameter values.

TABLE S4: Comparison of LD estimates by the proposed ML estimator to those by an imputation-based method when major allele frequencies at two sites of interest, p and q, are high.

D	μ	(p,q)	Φ	Method	$\widehat{D}$ (mean±SD)	$\widehat{D}^2$ (mean±SD)	Asymptotic theoretical SD($\widehat{D}$)	Theoretical $E[\widehat{D}^2]$	RMSD($\widehat{D}$)	RMSD($\widehat{D}^2$)
0.01	2	(0.9,0.9)	0	ML	0.012±0.014	0.00034±0.00049	0.011	0.00021	0.014	0.00054
0.01	2	(0.9,0.9)	0	Imputation	0.003±0.005	0.00003±0.00005	0.011	0.00021	0.009	0.00008
0.01	2	(0.9,0.9)	1/3	ML	0.011±0.014	0.00030±0.00043	0.008	0.00017	0.014	0.00047
0.01	2	(0.9,0.9)	1/3	Imputation	0.003±0.006	0.00005±0.00015	0.008	0.00017	0.009	0.00016
0.01	2	(0.9,0.9)	1	ML	0.011±0.011	0.00022±0.00031	0.008	0.00017	0.011	0.00033
0.01	2	(0.9,0.9)	1	Imputation	0.004±0.006	0.00005±0.00009	0.008	0.00017	0.009	0.00010
0.01	10	(0.9,0.9)	0	ML	0.009±0.010	0.00017±0.00024	0.011	0.00021	0.010	0.00025
0.01	10	(0.9,0.9)	0	Imputation	0.005±0.008	0.00008±0.00013	0.011	0.00021	0.010	0.00013
0.01	10	(0.9,0.9)	1/3	ML	0.010±0.009	0.00018±0.00023	0.008	0.00017	0.009	0.00024
0.01	10	(0.9,0.9)	1/3	Imputation	0.006±0.008	0.00009±0.00013	0.008	0.00017	0.009	0.00013
0.01	10	(0.9,0.9)	1	ML	0.010±0.008	0.00016±0.00018	0.008	0.00017	0.008	0.00019
0.01	10	(0.9,0.9)	1	Imputation	0.006±0.008	0.00011±0.00025	0.008	0.00017	0.009	0.00025
0.05	2	(0.9,0.9)	0	ML	0.048±0.021	0.00280±0.00245	0.016	0.00271	0.021	0.00245
0.05	2	(0.9,0.9)	0	Imputation	0.011±0.010	0.00022±0.00044	0.016	0.00271	0.041	0.00232
0.05	2	(0.9,0.9)	1/3	ML	0.048±0.020	0.00270±0.00204	0.014	0.00266	0.020	0.00204
0.05	2	(0.9,0.9)	1/3	Imputation	0.015±0.013	0.00039±0.00054	0.014	0.00266	0.037	0.00218
0.05	2	(0.9,0.9)	1	ML	0.049±0.018	0.00274±0.00204	0.014	0.00266	0.018	0.00204
0.05	2	(0.9,0.9)	1	Imputation	0.028±0.017	0.00106±0.00104	0.014	0.00266	0.027	0.00177
0.05	10	(0.9,0.9)	0	ML	0.051±0.015	0.00284±0.00153	0.016	0.00271	0.015	0.00156
0.05	10	(0.9,0.9)	0	Imputation	0.046±0.022	0.00257±0.00189	0.016	0.00271	0.023	0.00188
0.05	10	(0.9,0.9)	1/3	ML	0.051±0.015	0.00282±0.00158	0.014	0.00266	0.015	0.00161
0.05	10	(0.9,0.9)	1/3	Imputation	0.044±0.022	0.00242±0.00189	0.014	0.00266	0.023	0.00188

0.05	10	(0.9,0.9)	1	ML	0.052±0.014	0.00284±0.00152	0.014	0.00266	0.014	0.00156
0.05	10	(0.9,0.9)	1	Imputation	0.045±0.021	0.00248±0.00186	0.014	0.00266	0.021	0.00185

Sample means and standard deviations of the LD coefficient $\widehat{D}$ and its square $\widehat{D}^2$ estimated from simulated data by the ML estimator and the imputation-based method are compared with different values of the parameters. Root mean square deviations (RMSD) of the LD estimates are also compared. The comparisons are made when the mean depth of coverage, μ is low (2) or moderately high (10). Furthermore, comparisons with different values of the probability that an informative read covers both polymorphic sites, Φ, are made. The theoretical prediction of the asymptotic standard deviation, which is calculated as a square root of the sampling variance given by Equation 9 (when $\Phi > 0$) or10 (when $\Phi = 0$), is the expected level of achievement when individual genotypes are known without errors. Sample size $N = 100$, error rate $\varepsilon = 0.01$. A total of 100 simulation replicates were run for each set of parameter values.